STRUCTURED OUTPUT PREDICTION AND MODELING SOIL FUNCTIONS

Stevanche Nikoloski

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Supervisor: Prof. Sašo Džeroski, IPS and Jožef Stefan Institute, Ljubljana, Slovenia
Co-Supervisor: Dr. David P. Wall, TEAGASC: Agriculture and Food Development Authority, Johnstown Castle, Ireland

Evaluation Board:
Associate Prof. Marko Debeljak, Chair, IPS and Jožef Stefan Institute, Ljubljana, Slovenia
Dr. Dragi Kocev, Member, IPS and Jožef Stefan Institute, Ljubljana, Slovenia
Associate Prof. Michelangelo Ceci, Member, University of Bari Aldo Moro, Bari, Italy
Prof. Milena Horvat, Member, IPS and Jožef Stefan Institute, Ljubljana, Slovenia

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Stevanche Nikoloski

STRUCTURED OUTPUT PREDICTION AND MODELING SOIL FUNCTIONS

Doctoral Dissertation

NAPOVEDOVANJE STRUKTURIRANIH VREDNOSTI IN MODELIRANJE FUNKCIJ TAL

Doktorska disertacija

Supervisor: Prof. Sašo Džeroski

Co-Supervisor: Dr. David P. Wall

Ljubljana, Slovenia, December 2020

Посветено на моите најдраги...

Acknowledgments

My dreams to enroll in and successfully complete PhD studies have become reality thanks to my supervisor, Prof. Sašo Džeroski. I would like to thank him for choosing me as his student, for sharing his extraordinary knowledge in the field of Computer Science and for supporting and guiding me throughout the PhD studies at the Jožef Stefan International Postgraduate School. Special thanks go to Dragi Kocev, who was always available for my requests and for providing the right advice for improving the quality of my work. Dragi, I would not have succeeded in completing my PhD project without you. Thank you for that! At this point, I would like to acknowledge all my colleagues who I shared wonderful moments with inside and outside the Jožef Stefan Institute. In particular, I want to cordially thank Vladimir Kuzmanovski for his help with comments and suggestions for the dissertation part about modeling soil functions. Additionally, I would like to thank the other members of the evaluation board: Prof. Michelangelo Ceci, Prof. Marko Debeljak and Prof. Milena Horvat, for the careful reading and the constructive feedback on this work.

I would not be able to complete this dissertation without the funds, relevant data and domain expertise that mainly came from TEAGASC, Ireland, through the Walsh Fellowship Programme, especially from my TEAGASC co-supervisor Dr. David P. Wall. I would like to extend my special thanks to my dear colleagues Lilian O'Sullivan, Noeleen McDonald, Philip Murphy, Matthias Bacher and many others, for their unselfish support, assistance and unforgettable moments while staying in Ireland. Moreover, I would like to acknowledge the support of the European Commission through the project LANDMARK - Land management, assessment, research, knowledge base (H2020 Grant number 635201) and Environment Soils and Land-use Department at TEAGASC for providing the data and for constructive comments and suggestions. I am very grateful that I had Francesca Bampa, the LANDMARK project administrator, as my great support, friend and the main bridge between me and the members of the LANDMARK Consortium.

I acknowledge the research funding received through H2020 RIA HBP SGA 2 - Human Brain Project (H2020 Grant number 785907) and INTERREG V-A Italy and Slovenia cooperation program (TRAIN), which I took part in.

Finally, I would like to thank the most important people in my life, my sister Mirjana, my parents, relatives, friends and colleagues, for their understanding and unconditional support and motivation throughout the whole period of my PhD project. But most importantly, I would like to thank my love and motivation, Elena, for understanding and being with me and for sharing all the good and bad moments.

Abstract

The proposed dissertation belongs primarily to the field of machine learning on the one hand, but also to the field of soil science on the other hand. In terms of machine learning, it is concerned with the improvement of existing machine learning algorithms for predicting structured outputs, more specifically for multi-target prediction. In terms of soil science, it addresses two case studies of applying machine learning methods for multi-target prediction to two practical problems of modeling two different soil functions from data in the context of Irish agriculture.

The majority of approaches for multi-target prediction (MTP) do not explicitly take into account the dependencies among the multiple targets. In order to address this drawback, in the proposed dissertation, we propose approaches that find dependencies in the target space by explicitly structuring, in a hierarchical manner, the different targets. Using different representations of the target's attributes (based on the feature importance scores of the input attributes for predicting each target), we use hierarchical clustering of the targets. Having discovered a hierarchy on the target space, we obtain a reformulation of the original task of multi-target prediction into a task of hierarchical multi-target prediction. We then employ approaches for hierarchical multi-target prediction on the transformed task, expecting improved predictive performance.

We address two tasks of MTP, namely multi-label classification (MLC) and multi-target regression (MTR). In both cases, we use feature importance estimation based on tree-ensembles, for classification and regression, respectively, based on the GENIE3 approach. We use different clustering approaches for structuring the target space, including balanced k-means, agglomerative clustering, and predictive clustering trees (PCTs): Of these, balanced k-means gives the best results. On the hierarchical versions of the problems, we use PCT ensembles for hierarchical MLC (HMLC) and hierarchical MTR (HMTR), respectively.

We conduct an extensive experimental evaluation on various benchmark datasets for MTP (MLC and MTR) tasks, showing the advantage of using our proposed method for structuring the output space. Using ensembles of PCTs for HMLC and HMTR on the structured output spaces performs clearly better than using PCT ensembles for MLC and MTR on the original spaces. The differences in performance are largest for large output spaces (with more than 100 targets).

We also address two case studies of applying machine learning methods for multi-target prediction to two practical problems of modeling two different soil functions from data in the context of Irish agriculture. The data were provided by TEAGASC, Environment Soils and Land-use Department, from Ireland. TEAGASC was also the source of expertise about the tasks.

First, we apply PCTs for MTR, as well as ensembles (random forests) thereof to the task of estimating the total herbage production and nutrient uptake, i.e., the task of modeling the soil function of primary productivity, on Irish dairy farms. We then apply PCTs (and ensembles) for semi-supervised MTR to model a combination of another two soil func-

tions, i.e., water regulation and purification, and provision and cycling of nutrients. More specifically, we learn models for assessing the chemical quality (nitrogen and phosphorus loss from soils through runoff and leaching) and the biological quality of water in Irish agricultural lands. In the latter case, we used incompletely (partially) labeled data, which has missing values for the target variables we want to predict. This is an innovative use of semi-supervised PCTs for MTR, as only fully labeled (all target values present) or fully unlabeled (no target values) examples had been used so far, whereas the real-world data from this study has partially labeled examples (with some but not all target values).

In both case studies, models are learnt in the form of PCTs and PCT ensembles. They are both accurate (especially ensembles) and understandable (individual PCTs). They reveal knowledge about the studied domains, which is both consistent with existing knowledge of domain experts and provides new insights, important for practical use in the context of achieving better soil function outcomes for given fields/agricultural lands.

Povzetek

Predlagana disertacija sodi na področje strojnega učenja, po eni strani, ter na področje pedologije, po drugi strani. Z vidika strojnega učenja je v disertaciji predstavljena metodologija za izboljšanje obstoječih algoritmov strojnega učenja za napovedovanje strukturiranih vrednosti, oz. za napovedovanje več ciljnih spremenljivk. Z vidika pedologije, disertacija obravnava dve praktični študiji primerov, in sicer uporabo metod strojnega učenja za modeliranje različnih funkcij tal na osnovi realnih kmetijskih podatkov iz Republike Irske.

Večina pristopov za večciljno napovedovanje ne upošteva neposredno odvisnosti med ciljnimi spremenljivkami. Za odpravo te pomanjkljivosti disertacija predlaga pristope, ki najdejo odvisnosti med ciljnimi spremenljivkami in jih nato organizirajo v hierarhično strukturo. S pomočjo obstoječih metod hierarhično razvrščamo ciljne spremenljivke na podlagi ocen pomembnosti značilk za napovedovanje vsake posamezne ciljne spremenljivke. Na ta način transformiramo problem večciljnega napovedovanja v problem hierarhičnega večciljnega napovedovanja. Nato uporabimo pristope za hierarhično večciljno napovedovanje na preoblikovanem problemu in dobimo boljše točnosti napovedi.

Obravnavamo dve nalogi večciljnega napovedovanja, in sicer večciljno klasifikacijo in večciljno regresijo. V obeh primerih uporabljamo ocene pomembnosti značilk dobljene iz ansamblov dreves za klasifikacijo in regresijo. Za strukturiranje prostora ciljnih spremenljivk uporabljamo različne pristope razvrščanja, vključno z uravnotežena metoda k-means, hierarhičnim aglomerativnim razvrščanjem ter drevesi za napovedno razvrščanje, pri čemer da uravnoteženo metodo k-means najboljše rezultate. Pri hierarhičnih različicah problemov napovednega modeliranja uporabljamo ansamble dreves za napovedno razvrščanje za hierarhično večciljno klasifikacijo oz. hierarhično večciljno regresijo.

Prednosti uporabe predlagane metode pokažemo z obsežno eksperimentalno evalvacijo na različnih naborih podatkov za probleme večciljnega napovedovanja (večciljne klasifikacije in regresije). Uporaba ansamblov dreves za hierarhično večciljno klasifikacijo in hierarhično večciljno regresijo na hierarhičnih različicah problemov daje signifikantno boljše rezultate kot uporaba ansamblov dreves za večciljno klasifikacijo in večciljno regresijo na izvirnih problemih. Razlike v uspešnosti so največje pri velikih izhodnih prostorih (z več kot 100 ciljnimi spremenljivkami).

Obravnavali smo tudi dve študiji primerov uporabe metod strojnega učenja za večciljno napovedovanje na dveh praktičnih problemih modeliranja različnih funkcij tal na realnih kmetijskih podatkih. Podatke je zagotovil TEAGASC, Oddelek za okolje in rabo zemljišč Republike Irske. Strokovnjaki TEAGASC so bili tudi vir domenskega znanja pri interpretaciji dobljenih rezultatov tj. naučenih modelov in njihovih napovedi.

Najprej smo uporabili drevesa za napovedno razvrščanje za večciljno regresijo kot tudi njihove ansamble (naučene z metodo naključnih gozdov) za nalogo ocenjevanja pridelave travinje in vnosa hranil. Gre za nalogo modeliranjene funkcije tal, tj. primarne produktivnosti na irskih mlečnih kmetijah. Nato smo drevesa za napovedno razvrščanje (in ansamblem teh dreves) uporabili za polnadzorovano večciljno regresijo pri modeliranju druge funkcije tal, in sicer regulacije in čiščenja vode. Učili smo se modelov za ocenje-

vanje kemijske kakovosti (izgube dušika in fosforja) in biološke kakovosti voda na irskih kmetijskih zemljiščih. V slednjem primeru smo uporabili nepopolno (delno) označene podatke, v katerih so manjkale vrednosti za ciljne spremenljivke. Gre za inovativno uporabo polnadzorovanih dreves za napovedno razvrščanje za večciljno regresijo, saj so bili doslej v objavljenih znanstvenih publikacijah uporabljeni popolnoma označeni podatki (z vsemi ciljnimi vrednostmi) in popolnoma neoznačeni podatki (brez ciljnih vrednosti), ne pa tudi delno označeni podatki.

V obeh študijah primerov smo se učili modelov v obliki dreves za napovedno razvrščanje in ansamblov dreves za napovedno razvrščanje. Oba pristopa sta natančna (zlasti ansambli) in razumljiva (posamezna drevesa za napovedno razvrščanje). Iz podatkov sta odkrila znanje o preučevanih domenah, ki je v skladu z obstoječim domenskim znanjem ter hkrati ponuja nove vpoglede, pomembne za praktično uporabo v smislu doseganja boljših funkcij tal na danih kmetijskih zemljiščih.

Contents

List of Figures **xv**

List of Algorithms **xvii**

Abbreviations **xix**

1 Introduction **1**
1.1 Motivation 2
1.2 Goals, Hypotheses and Methodology 4
1.3 Contributions 6
1.4 Organization of the Thesis 8

2 Background **11**
2.1 Structured Output Prediction (SOP) 11
2.2 Multi-Target Prediction (MTP) 12
2.3 Hierarchical Multi-Target Prediction (HMTP) 13
2.4 Semi-Supervised Learning for SOP 15
2.4.1 Semi-supervised learning (SSL) with labeled and unlabeled data . . 16
2.4.2 Semi-supervised learning (SSL) with partially-labeled data 18
2.5 Modeling Soil Functions 19

3 Related Work **23**
3.1 Methods for Multi-Target Prediction 23
3.1.1 State-of-the-art methods for multi-label classification 23
3.1.2 State-of-the-art methods for multi-target regression 25
3.1.3 State-of-the-art methods for semi-supervised MTR 28
3.2 Methods for HMTP 29
3.2.1 State-of-the-art methods for HMLC 29
3.2.2 State-of-the-art methods for HMTR 30
3.3 Methods for Structuring the Output Space in MTP 31

4 State-of-the-art Methods for Modeling Soil Functions **33**
4.1 Primary Productivity 34
4.1.1 DSS for modeling primary productivity 34
4.1.2 ML for modeling primary productivity 35
4.2 Water Regulation and Purification 37
4.2.1 DSS for modeling water regulation and purification 37
4.2.2 ML for modeling water regulation and purification 38
4.3 Carbon Storage and Climate Regulation 39
4.3.1 DSS for modeling carbon storage and climate regulation 39
4.3.2 ML for modeling carbon storage and climate regulation 40
4.4 Provision of Functional and Intrinsic Biodiversity 41

4.4.1 DSS for modeling provision of functional and intrinsic biodiversity . 41
4.4.2 ML for modeling provision of functional and intrinsic biodiversity . . 42
4.5 Provision and Cycling of Nutrients 43
4.5.1 DSS for modeling provision and cycling of nutrients 43
4.5.2 ML for modeling provision and cycling of nutrients 44

5 Structuring the Output Space in Multi-Target Prediction 47
5.1 Structuring the Label Space in Multi-Label Classification Using a Feature Ranking Representation of the Labels 48
5.2 Structuring the Output Space in Multi-Target Regression Using Different Representations of the Targets 66

6 Multi-Target Prediction for Modeling Soil Functions 89
6.1 Estimation of Herbage Production and Nutrient Uptake on Irish Dairy Farms 90
6.2 Exploiting Partially-Labeled Data for Learning Water Quality Models in Irish Agricultural Catchments 110

7 Conclusions 159
7.1 Summary of Contributions 160
7.2 Further Work 161

References 163

Bibliography 179

Biography 181

List of Figures

Figure 2.1: An illustrative example of a multi-target regression (MTR) dataset. The table contains a set of images described with visual features extracted from the images, such as the number of lines of different type (curved, horizontal, diagonal, etc.) and target attributes given as numerical values (numbers of objects of the given types). 12

Figure 2.2: An illustrative example of a multi-label classification (MLC) dataset. The table contains a set of images described with visual features extracted from the images, such as the number of lines of different type (curved, horizontal, diagonal, etc.) and target attributes denoting the presence ($\top$) or absence ($\bot$) of objects of different types. 13

Figure 2.3: An illustrative example of a hierarchical multi-label classification (HMLC) dataset. The table contains a set of images described with visual features extracted from the images, such as the number of lines of different type (curved, horizontal, diagonal, etc.) and target values given as a hierarchical representation of labels, i.e., strings where child labels from different branches of the hierarchy are separated by the symbol '@'. . . 15

Figure 2.4: An illustration of the semi-supervised learning in structured output prediction with respect to different data labeling. 16

Figure 2.5: A simple example to demonstrate semi-supervised learning with labeled and unlabeled data. 17

Figure 2.6: Illustrative representation of the five soil functions (Schulte et al., 2014). 20

Figure 5.1: An illustration of the proposed framework for structuring the output space in MTP. We consider two target representations for clustering: the representation that uses original output space, i.e., the values of each target/label for each example and the feature ranking representation. The importance scores for each feature with respect to a given target/label are used as a representation of the targets. We thus transform the original MTP tasks into a HMTP task, for which we build a HMTP model. 48

List of Algorithms

Algorithm 5.1: The algorithm for structuring the output space in multi-target prediction (MTP). 48

Abbreviations

AI	...	Artificial intelligence
XAI	...	Explainable artificial intelligence
ML	...	Machine learning
DSS	...	Decision-support system
DEX	...	Decision expert
SOP	...	Structured output prediction
MTP	...	Multi-target prediction
MTR	...	Multi-target regression
MLC	...	Multi-label classification
HMTP	...	Hierarchical multi-target prediction
HMTR	...	Hierarchical multi-target regression
HMLC	...	Hierarchical multi-label classification
PCT	...	Predictive clustering tree
RF	...	Random forests ensemble of PCTs
RRMSE	...	Relative root mean squared error
LANDMARK	...	LAND, Management, Assessment, Research, Knowledge base
SF	...	Soil function

Chapter 1

Introduction

We live in the age of artificial intelligence. The term *artificial intelligence* may sound threatening, yet it has been used for a considerable length of time and its applications are more mundane than human fantasy may envision. Artificial intelligence (AI) (Russell & Norvig, 2009) aids each aspect of our lives, regardless of whether we are attempting to read our emails, get driving directions, start a new business, or even trying to book an accommodation in a hotel, chatting directly with "chatbots".

Artificial intelligence, as a sub-field of computer science referring to the intelligent behavior of machines (i.e., computers), has been gaining momentum in almost every domain where the large amounts of data and knowledge are available. AI helps to better understand the data, produce novel knowledge and facilitate the decision-making process. Machine learning (ML) is a branch of AI that is concerned with designing methods for data analysis that can automatically induce predictive models from data, rather than following some pre-programmed rules (Mitchell, 1997). Data mining is a process that uses different approaches (including machine learning) to extract potential knowledge and/or interesting patterns from data (Witten & Frank, 2005).

The proposed dissertation addresses the topic of machine learning, currently one of the most prominent topics in the field of information and communication technologies. The data used for predictive modeling, the most common task of machine learning, are composed of inputs (attributes) and outputs (targets). The input space consists of vectors of values of the descriptive attributes, while the output space can be represented differently. The values taken by the output space may be simple primitive data types (discrete, Boolean, continuous values, etc.) or complex data structures, such as tuples of values, sequences (including time series) and hierarchies (Panov et al., 2016). The specific task concerned with learning structured outputs is called structured output prediction (Bakır et al., 2007; Džeroski, 2006; Panov et al., 2014).

The main goal in structured output prediction is to learn a model predicting the target value(s) of previously unseen examples. The model is learned from a set of examples with known values of the target variable(s). If the target space consists of a single target, the learning task at hand is a single-target prediction task. If the target space consists of more than one target, the task at hand is called multi-target prediction (MTP).

A multi-target prediction task can be addressed by learning local or global models. In the first case, a separate model for each target attribute is learned. In the second case, a single, global model is learned, predicting all the target variables simultaneously.

If the values of the target attributes are numerical, the learning task is called a multi-target regression task. If the target attributes are discrete/nominal, the learning task is called multi-target classification task. Specifically, if each example can be associated with multiple labels, i.e. target attributes with binary values (0 or 1), the learning task is called multi-label classification task (Tsoumakas & Katakis, 2007).

There are many real-life examples, where there is a practical need for MTP models. Prominent examples can be found in ecology, for instance, predicting the abundance of different species occupying the same habitat (Demšar et al., 2006), estimating different vegetation quality indices for the same site (Kocev et al., 2009), predicting the weed cover profile from crop-related input variables (Debeljak et al., 2011), predicting the composition of a community of organisms (Levatić, Kocev, Debeljak, et al., 2015), predicting soil bulk density based on visual parameters (Bondi et al., 2018) and mapping vegetation-impervious-soil fractions across multiple cities by using multi-target regression models (Okujeni et al., 2018).

1.1 Motivation

Both local and global models have been extensively studied in the field of MTP. It has been shown that transformations of the output space can yield improved predictive performance as compared to standard local and global models using the original targets. An example of this are the ensembles of local and global models, which consist of base models learned on parts of the output space (Breskvar et al., 2018).

Next, several studies have investigated the influence of introducing a structure in the output space on predictive performance (Levatić, Kocev, & Džeroski, 2015; Madjarov et al., 2019). Furthermore, Madjarov et al. (2016) propose a methodology for data-driven structuring of the label space in multi-label classification, by constructing a hierarchical representation on top of the labels. Then, the hierarchy of labels is used in hierarchical multi-label classification, a hierarchical variant of the MLC task.

Further, Szymanski et al. (2016) show that data-driven hierarchies obtained by structuring the label space in MLC, are superior to random generated graphs. Both Madjarov et al. (2016) and Szymanski et al. (2016) exploit hierarchies obtained by clustering the labels space. In this dissertation, we propose a novel approach, for constructing a label hierarchy in MLC by considering feature importance scores for each label (Nikoloski et al., 2018).

Unlike the case of MLC, in the case of MTR, there is no research investigating and exploring the dependencies among the target variables. In this dissertation, like for the case of MLC, we identify relations among targets and structure the output space in MTR. We do this by clustering the values of the target attributes, or the values of the feature importance scores for the target attributes (Nikoloski, Kocev, & Džeroski, 2019). We investigate whether data-driven structuring of the target space is superior to considering the original MTR task. We transform the original MTR task to a task of hierarchical multi-target regression (HMTR) (Mileski et al., 2017) from the learned target hierarchy. The proposed approach is also of interest for eliciting domain knowledge, since the domain experts are typically interested in the relations among targets, i.e., response variables. We thus provide a methodological pipeline to upgrade the existing and potentially discover new knowledge about the relations among the targets.

We also investigate whether the concept of structuring large target spaces is applicable to environmental data. We apply our approach of structuring the output space to two datasets with large output spaces, one with 111 target attributes indicating the relative abundance of diatom species in a label and another one with 492 target attributes indicating the abundance of water bioindicator species in Slovenian rivers (Nikoloski, Kocev, & Džeroski, 2019).

The main focus of the dissertation is on the machine learning task of structured output prediction, i.e., predictive modeling with multiple response variables, known as multi-target prediction. The structured output can be either ‘complete’, i.e., all the values of all target attributes are known (supervised learning) or ‘incomplete’, i.e., some of the values of some of the target attributes in some examples are not known (semi-supervised learning). This happens in practice in domains with a very complex process of monitoring or very expensive acquisition of the values of the target attributes. In the dissertation, we investigate learning scenarios where the data can have complete or incomplete structured output values.

In many domains, such as computer vision (Navaratnam et al., 2007; Rosenberg et al., 2005), computer linguistics (Yarowsky, 1995), etc., it is not always possible to obtain complete data, but there are vast numbers of ’incomplete’ examples. In that case, a novel methodology for exploiting those ’incomplete’ data is needed. Semi-supervised learning (SSL) is able to solve this problem to some extent. Namely, recent methods for semi-supervised learning, e.g. for multi-target regression, consider only unlabeled data (i.e., examples with completely unknown values for all target attributes) in addition to labeled data (i.e., examples with completely known values for all target attributes) (Levatić et al., 2018). However, in practice, the collected samples (neither all nor none) may have known values for some of the target attributes. In that case, the partially-labeled examples are discarded from the learning set. In this dissertation, we propose the use of a re-designed semi-supervised method that can handle partially-labeled examples applied on soil-related data.

In collaborative interdisciplinary research, there is often a gap in communicating the expertise between computer scientists and domain experts, hence it is normal to ask the question whether and how communication could be improved. In order to answer it, a stronger connection and collaboration between the domains needs to be fostered by performing research in an iterative way. For example, problems from environmental and soil sciences could inspire the development of new methods for data analysis, while data analysis could yield some new knowledge about the environmental system under observation.

A major attempt to address this problem has been made within the LANDMARK H2020 EU project (www.landmark2020.eu), where experts from the domains of soil and computer science collaborated in order to quantify the current and potential supply of the five main soil functions on agricultural land across the EU. Those functions are: (i) primary productivity; (ii) water regulation and purification; (iii) carbon sequestration and climate regulation; (iv) habitat for functional and intrinsic biodiversity and (v) nutrient cycling and provision (Schulte et al., 2014). The collaboration within the LANDMARK H2020 EU projects resulted in a decision support tool called the Soil Navigator (Debeljak et al., 2019). The Soil Navigator is an evidence-based DSS, which aims to assess and improve the supply of several soil functions simultaneously, using multi-criteria decision modeling with the Decision Expert (DEX) integrative methodology (Bohanec, 2014, 2017; Bohanec & Rajkovič, 1990). Instead of using expert knowledge only, our dissertation takes a complementary approach providing additional insights by using advanced machine learning methods for finding patterns and new knowledge from data.

The work performed in the dissertation lies at the intersection of artificial intelligence and ecological modeling. In order to attempt to bridge the gap between computer scientists on one and environmental and soil scientists on the other side, we have applied machine learning methodologies to real-world data. We apply existing approaches for multi-target regression in the domain of soil function modeling, handling the interdependencies among different aspects of soil functions or interdependencies among the main drivers of specific soil functions, by using predictive clustering trees (PCTs) or ensembles (random forests) of PCTs for multi-target regression. First, we study the problem of predicting multiple indicators of the primary productivity of soil. In particular, we address the prediction of herbage potential and nutrient uptake in Irish diary grassland farms. Our purpose is to propose potential management practices for increasing the grass yield production as the main feed for the dairy cattle and sheep (Nikoloski, Murphy, et al., 2019).

Second, we use semi-supervised PCTs and ensembles to learn from partially-labeled data about water quality in Irish grassland soils (Schulte et al., 2006). The data include three target attributes, i.e., biological water quality, phosphorus and nitrogen concentration in water, and only 50% of the data samples are 'completely' labeled. The remaining 50% are 'incomplete' (i.e., partially-labeled). We have learned local and global models by using complete data (partially-labeled in addition to labeled) and investigate whether the models learned from 'incomplete' data are accurate, understandable and interpretable from the domain perspective (Nikoloski et al., 2020).

Here we must note that the soil functions and their outcomes modeled in this dissertation are closely related to the soil functions considered in the project LANDMARK, but are not exactly the same. In the LANDMARK Soil Navigator tool for the assessment of soil functions, overall assessments are obtained by integrating the expert assessments of many different aspects of that soil function. In this dissertation, we model specific aspects and outcomes of a soil function, but not the complete soil function.

All in all, the thesis proposes the implementation and evaluation of new approaches for structuring the output spaces in MLC and MTR learning tasks. We use predictive clustering trees (PCTs) as the most adequate technology for developing our concept. PCTs are implemented in the CLUS software package (http://source.ijs.si/ktclus/clus-public). The existing MTR methods and the novel methods for structuring the output space have been applied to environmental and soil-related data obtained from various sources, mainly from TEAGASC, Ireland.

1.2 Goals, Hypotheses and Methodology

The main goal of the dissertation is to enhance the methods for MTP by addressing the MTP task through its hierarchical counterpart – hierarchical multi-target prediction. This is achieved by discovering hierarchical structures among the target attributes (binary labels in the case of MLC or continuous targets in the MTR case). With the newly designed framework, we will strive to improve the accuracy of predictive models, especially for large output spaces (>100 targets/labels). Next, we aim to investigate the potential of existing semi-supervised learning algorithms for MTR to exploit partially-labelled examples ('incomplete' examples where not all of the values for the target attributes are known). Finally, we aim to extensively analyse and evaluate the proposed methodology in various domains, either on existing benchmark datasets for MLC and MTR, or on new data appropriate for the tasks at hand.

We also aim to apply the methodology of MTR on environmental, and in particular soil, data provided by TEAGASC, Ireland. Our goal is to show whether the predictive clustering trees for multi-target regression could provide accurate and understandable models that will empower the domain experts to look for novel insights, relations and patterns in the data. Finally, the novel insights coupled with the existing body of knowledge might translate into useful recommendations to the practitioners from the specific domain under study – in our case, these would be the Irish farmers.

The main hypotheses investigated in this dissertation are:

Hypothesis 1. Transforming MTP tasks to hierarchical MTP tasks by using data-derived hierarchies obtained by structuring the output spaces in the MTP tasks will improve the predictive performance on the original MTP tasks (especially for large output spaces).

a. Structuring the label/output space in MLC tasks by using per label feature importance scores to describe labels will improve the predictive performance on the original MLC tasks.

b. Structuring the target space in MTR tasks by using the target values themselves or the feature importance scores for each target will translate MTR tasks to HMTR tasks and this will lead to improved predictive performance on the original MTR tasks.

Hypothesis 2. PCTs can learn predictive models with good predictive performance that contribute novel domain knowledge in estimating herbage potential and nutrient uptake from soil, management and environmental data about Irish dairy farms.

Hypothesis 3. Exploiting partially labeled examples brings additional value to the learned models in terms of predictive performance and understandability in the domain of modeling water quality across Irish agricultural land.

To achieve the research goals and to test all the hypotheses stated above, we use predictive clustering trees PCTs (Blockeel et al., 1998; Struyf & Džeroski, 2006) and ensembles thereof (Kocev et al., 2013). The PCT algorithm is one of the most appropriate modeling algorithms for solving the original problem of MTP (MLC and/or MTR), as well as the problem of hierarchical multi-target prediction (HMTP), given the use of PCTs for HMLC (Vens et al., 2008) (Hypothesis 1a.) and PCTs for HMTR (Mileski et al., 2017) (Hypothesis 1b.). In order to test the second hypothesis, we apply PCTs for MTR to a set of data collected at Irish dairy farms. We address the problem of predicting herbage accumulation and nutrient uptake from soil, management and environmental data. On one hand, we evaluate the predictive performance of the built models. On the other hand, we have domain experts inspect the built models and assess whether they are consistent with existing and contribute novel domain knowledge.

In order to test the third hypothesis, we investigate the use of SSL for MTR (Levatić et al., 2018) with PCTs for exploiting partially labeled examples. We use ensembles of PCTs (random forests of PCTs) (Kocev et al., 2013) for MTP in order to obtain models with state-of-the-art predictive performance. In the domain of modeling water quality across Irish agricultural land, we also use the learned ensembles of PCTs for generating and drawing accurate maps.

All of the proposed methods for structured output prediction will be developed in the CLUS framework (http://source.ijs.si/ktclus/clus-public).

In the proposed dissertation, we depart from the existing local and global modeling approaches for structured output prediction, by implementing a novel algorithm for structuring the output space in multi-target prediction tasks. We create data-derived hierarchies by clustering two different target-related representations, i.e., the representation of targets by their target values for each example and the representation of targets by the feature importance scores of each feature for that target. We use the obtained data-derived hierarchy and transform the original MLC/MTR tasks to hierarchical MLC/MTR tasks (HMLC/HMTR). Furthermore, we investigate the influence of the proposed methodology on predictive performance in the case of learning both PCTs and ensembles of PCTs. In order to obtain the feature importance scores for predicting individual targets/labels, we use the GENIE3 feature ranking method (Huynh-Thu et al., 2010). For hierarchy creation, i.e., clustering the targets from the output space, we use hierarchical agglomerative clustering (with complete and single linkage), balanced k-means, and predictive clustering trees (PCTs).

The performance of the methods for creating predictive models can vary across different domains. We evaluate our methods on various benchmark datasets from different domains (for MLC: text classification, movie clips and genre classification and biology, presented in Chapter 5.1 (Nikoloski et al., 2018); for MTR: socio-economic and environmental sciences, presented in Chapter 5.2). Moreover, we evaluate our methods on several real-world datasets, in the domain of environmental sciences, especially our methods for MTR. The datasets mostly have large output spaces. Therefore, structuring large output spaces in these datasets provides a clear picture of whether our proposed methods perform well (Nikoloski, Kocev, & Džeroski, 2019).

In order to investigate the usability of the existing and proposed methods for structured output prediction, we have developed two case studies in the domain of environmental and soil sciences. We use practically relevant data provided by TEAGASC, Environment Soils and Land-use Department, Ireland. TEAGASC – the Agriculture and Food Development Authority – is the national body providing integrated research, advisory and training services to the agriculture and food industry and rural communities of Ireland.

The first case study, presented in Chapter 6.1 of this dissertation, corresponds to our second hypothesis. In this case study, we have investigated the usability of predictive clustering trees for multi-target regression as well as random forests of PCTs for MTR in estimating herbage accumulation and nutrient uptake on Irish dairy farms. This task is related to the primary productivity soil function (Nikoloski, Murphy, et al., 2019). The second case study, presented in Chapter 6.2 of the dissertation, corresponds to our third hypothesis. In this case study, we have focused on using supervised PCTs (that handle only complete data) and semi-supervised PCTs (that handle incomplete data, in addition to the complete data) for MTR, as well as ensembles of PCTs (random forests) for MTR for assessing the water quality in Irish agricultural lands by simultaneous estimation of the biological water quality as well as phosphorus and nitrogen concentration in the water. This task includes specific aspects and outcomes of two soil functions: (1) water regulation and purification and (2) provision and cycling of nutrients.

1.3 Contributions

The work presented in this dissertation comprises several contributions to the field of computer science, especially in machine learning algorithms for structured output prediction and application of such methods in the field of environmental (soil) sciences. A complete list of publications related to this research are given in the Bibliography section. A summary of the thesis contributions is given as follows:

Contribution 1. *Improving the predictive performance on multi-target prediction (MTP) tasks by structuring the output space and using the obtained data-derived hierarchy in hierarchical multi-target prediction (HMTP) tasks.*

- *A novel method for MLC, which structures the output space by clustering labels represented by feature importance scores and uses the obtained hierarchy in HMLC tasks.*
- *A novel method for MTR, which structures the output space by clustering targets represented by their original target values (target space) or by feature importance scores (feature ranking space) and uses the obtained hierarchy in HMTR tasks.*
- *An extensive empirical evaluation of the novel methods on benchmark datasets, which shows their improved performance.*

This contribution is mainly related to a novel methodology for multi-target prediction (MLC/MTR) tasks that converts the original MLC/MTR tasks to hierarchical variants, i.e., HMLC/HMTR tasks, using data-derived hierarchies structuring large output spaces. Namely, from the output space, we generate two different target representations, one consisting of the original values of the target/label attributes and another consisting of the importance scores of the input attributes for each target/label. For multi-label classification tasks, we show that better performance can be achieved if the hierarchy of labels, obtained by clustering the feature ranking representation, is used. Improvements are achieved along the majority of the 13 different most commonly used evaluation measures (Chapter 5.1), confirming Hypothesis 1a. Next, we perform an extensive study on different benchmark datasets in the case of multi-target regression (MTR) task. We show that structuring the target space, i.e., using data-derived hierarchies, improves the predictive performance as compared to the original MTR task where no hierarchy is used. Moreover, we show that by using data-derived hierarchies in the HMTR task, there are improvements in predictive performance even over using expert-provided (i.e., existing) hierarchies, especially in datasets with large (>100 targets) output spaces (Chapter 5.2), confirming Hypothesis 1b. In both studies, the divisive clustering methods (balanced k-means and PCTs) proved to be better than agglomerative methods for structuring (i.e., clustering) the output space. Overall, the results confirm Hypothesis 1.

Contribution 2. *A case study of modeling primary productivity (total herbage production and nutrient uptake) of Irish dairy farms from existing soil, weather and management data by using machine learning algorithms for multi-target regression.*

This contribution concerns the problem of modeling total herbage production and nutrient uptake. To this end, we use existing soil, environmental (weather) and management data from 15 commercial dairy farms in Ireland. The study related to this scientific contribution is presented in Chapter 6.1. The goal of this study is to develop a predictive model that can easily explain the relations between primary productivity and controllable (and non-controllable) factors related to soil, weather and management practices.

For learning predictive models we use predictive clustering trees (PCTs) and random forests of PCTs for single- and multi-target regression. Moreover, we perform additional research on finding the most limiting nutrient uptake for the total yield produced. Our results confirm the stated hypothesis (Hypothesis 2.), i.e., we have learned easily explainable predictive models with good predictive performance for all target variables (total herbage production, N, P and K uptake) simultaneously.

The results in terms of the predictive performance of the obtained PCT models for MTR are in accordance with the expectations of the domain community. Our main contribution is related to the understandability of the model and its application in the domain. We have found that the number of grazing events is closely related to the soil drainage class. Therefore, we have performed predictive modeling on three data subsets (in addition to the complete dataset), defined by the soil drainage class, i.e., for well-drained, somewhat-poorly drained, and poorly-drained soils.

Overall, our results show that phosphorus (P) uptake was the most limiting nutrient for herbage production on these Irish dairy farms, followed by nitrogen (N) and potassium (K). The predictive rules embodied in the multi-target regression tree are in accordance with the existing expert knowledge. Moreover, they provide additional insights into the factors that drive the yield production and nutrient uptake the most. This is very important in guiding the process of further collecting of data. Namely, instead of collecting all the features gathered so far, which is a very complex and laborious process, it is enough to collect only the relevant features that appear in the predictive rules streamlining the process of data monitoring and collection.

Contribution 3. *A case study of modeling water quality across agricultural lands in Ireland, learning from partially-labeled data by using semi-supervised learning algorithms for multi-target regression.*

The contribution of this part of the thesis is the novel application of recent semi-supervised algorithms for multi-target regression in the field of environmental (soil) sciences. Besides the application domain point-of-view, there are two important methodological aspects of this contribution. The use case study related to this scientific contribution is presented in Chapter 6.2 of this dissertation.

First, to the best of our knowledge, semi-supervised learning approaches have not yet been applied in the domain of environmental sciences. In fact, semi-supervised learning approaches are typically evaluated in benchmark settings where the labels of all examples are known and missing labels are simulated by artificially deleting the known labels. Practical applications are not common, and where they exist they typically come from domains like text, image and multi-media classification, and possibly drug design/re-purposing.

Second, most methods for semi-supervised learning in the context of structured-output/ multi-target prediction cannot deal with partially-labeled examples. They can use fully-labeled examples, with known values for all targets. They can also use fully-unlabeled examples with no known values for any target. But they cannot make use of examples that have known values for some and unknown values for other targets: This is a unique capability of the approach employed in this thesis.

We use existing and already pre-processed water quality data, collected during the national monitoring program in the years 2001, 2002 and 2003, in the Republic of Ireland. The data consist of pressure and pathway-related variables, the main drivers of water quality. We build data-driven models learning to predict the following target variables: biological water quality (Q-value), phosphorus (P) and nitrogen (N) concentrations.

Note that not all of the target variables are measured in each of the 708 data samples (examples). Namely, the descriptive attributes are complete/labeled for all data samples and the three predictors (i.e., target attributes) are incomplete, i.e., partially-labeled. Therefore, we use semi-supervised predictive clustering trees (PCTs) for multi-target regression to learn from the partially-labeled data. We use single PCTs and ensembles (random forests) of PCTs as learning methods.

Our results show that more accurate models can be achieved when semi-supervised approaches are used. Global (i.e., multi-target) models predict all targets simultaneously and overfit less. The most accurate predictions are obtained by using ensemble (i.e., random forest) PCT models, but ensembles are not easily explainable. Finally, single semi-supervised PCTs for MTR are smaller and more easily explainable. Overall, these results confirm Hypothesis 3 in full.

1.4 Organization of the Thesis

The current chapter is introductory. So far, it has provided motivation and an overview of the current research on the topic.

Next, it has outlined the research goals of the thesis and the hypotheses. Also, the methodology, used to accomplish the set of goals has been described briefly. In the previous subsection, the main scientific contributions of the thesis have been explained. The remainder of the dissertation is organized as follows.

Chapter 2 describes the background of the thesis. It begins with predictive modeling, i.e., machine learning for structured output prediction. Descriptions and formal definitions are given of the tasks of multi-target prediction (MLC/MTR) and hierarchical multi-target prediction (HMLC/HMTR). Next, we discuss semi-supervised learning from partially/incompletely labeled data. Finally, we briefly describe the problem of modeling soil functions. Namely, we define the five main soil functions and briefly review the recent research on modeling these soil functions within the LANDMARK project (LANDMARK, 2019).

Chapter 3 describes in more detail related work concerning methods for SOP. First we present the state-of-the-art of methods for MLC and MTR, followed by the state-of-the-art methods for HMLC and HMTR. Finally, we end this chapter with a description of existing methods for structuring the output space in MTP.

Chapter 4 presents the state-of-the-art of using artificial intelligence methods for modeling soil functions. On the one hand, we present the existing uses of machine learning in this area. On the other hand, we review the use of decision support methods for modeling each soil function.

Chapter 5 describes the problem of structuring the output space in multi-target prediction and our first set of contributions to science. First, we present our method for structuring the output space in multi-label classification by clustering labels in terms of feature importance scores per label. We compare our results to the results obtained for the flat MLC tasks, where no hierarchy on the labels is used. We next present our study related to the problem of structuring the output space in MTR. We propose two algorithms for structuring the output space in MTP based on two different representations of the targets. The first algorithm represents the targets by the values of their attributes while the second represents them by the feature importance scores. We use the data-derived structure (hierarchy) in the hierarchical multi-target regression, HMTR reformulation of the original task.

Chapter 6 describes our case studies of modeling two specific soil functions: primary productivity, and water regulation and purification, by using predictive clustering trees (PCTs) and random forest of PCTs for multi-target regression. Our first case study comes from the domain of dairy science. Namely, we use PCTs and random forests of PCTs for multi-target regression for modeling the total herbage production and nutrient uptake on 15 commercial dairy farms in Ireland. We use existing soil, environmental and management data from these farms, collected in the years 2015 and 2016. Our second case study concerns the modeling of water quality in Irish agricultural catchments. The data describes 708 sites (i.e., examples) in terms of pressure-pathway descriptive attributes and three partially-labeled water quality indicators (targets): biological water quality (Q-value), as well as phosphorus (P) and nitrogen (N) concentrations. The data were collected during the national monitoring program in the years 2001, 2002 and 2003 and not all of the target attribute values have been measured. We use supervised and semi-supervised PCTs for MTR and ensembles thereof and compare the predictive performance and complexity, represented by model accuracy and model size in both local (single-target) and global (multi-target) predictive modeling scenarios.

Finally, Chapter 7 concludes the dissertation. It summarizes the scientific contributions of the thesis and outlines several possible directions for further work.

Chapter 2

Background

2.1 Structured Output Prediction (SOP)

In this section we will describe the main data constituents, which machine learning algorithms require to learn from. The data constituents are the *attributes* and the values for those attributes given as *data examples* or *data instances*. Technically, data example is a record with one or more qualities with certain constraints, which apply to the values that can be taken by data examples with respect to that given quality. Therefore, each data quality must specify its own constraints on the values which is done through the use of data types. Data types are fully described in OntoDT data type taxonomy, a part of OntoDM-ontology for data mining (Panov et al., 2016). We have two groups of data types: primitive and generated. Primitive data types are: *Boolean* data types that can be either false ($\perp$) or true ($\top$); *real* data type consisting of the values from the set of real numbers $\mathbb{R}$ and *discrete* data types, consisting of the values given in the predefined set of possible values (for example: apple, orange, pear and peach). Generated data types are constructed either fully or partially by primitive data types. There are various generated data types (Allison, 2003), but in this dissertation we will focus on generated data type called *tuple*. A tuple is an ordered list of data types. An example of a tuple that takes two values would be tuple (Boolean, discrete(apple, orange, pear)). The first value is from Boolean data type and could be either $\perp$ or $\top$, but the second value is discrete and can be taken from the set of values {apple, orange, pear}. Examples of such tuples are: ($\top$,apple), ($\perp$, pear), etc.

Therefore there are many different data representations. Machine learning methods are robust, because they have an ability to process any given data representation. One of the possible representations of the data is a matrix representation. The data matrix (i.e., *dataset*) consists of data examples in the rows. The columns in the matrix are called *attributes* and they denote the properties of a specific data example. In predictive modeling, data attributes are represented by an input set of attributes X, called *features* or *descriptors*, and an output set of attributes Y, named *targets* or *predictors*. In classical machine learning tasks (single-target), the output space Y consists of one target (i.e., predictor) of a primitive data type. However, in this dissertation we will focus on output space Y consisting of more than one attribute ($|Y| > 1$) that can be represented by complex data structures, such as tuples of values, sequences (including time series) and hierarchies (Bakır et al., 2007; Džeroski, 2006; Panov et al., 2016). This specific task concerned with learning from structured outputs is called *structured output prediction (SOP).*

2.2 Multi-Target Prediction (MTP)

We can distinguish among different machine learning tasks for SOP based on the representation of the output space Y. If Y is represented by tuples of values for the target attributes, the specific SOP task is known as the *multi-target prediction (MTP)* task. If the values in the tuple in the MTP task are from real/numeric type, the MTP task is called the *multi-target regression (MTR)* task. An illustrative example of the data set used for the MTR task is given in Figure 2.1. Likewise, if the values in the tuple are from a discrete data type, the MTP task is called the *multi-target classification (MTC)* task. Specifically, if the values in a tuple are from binary (0 or 1), i.e., Boolean type (i.e., $\bot$ or $\top$), the specific MTP task is known as the *multi-label classification (MLC)* task. An illustrative example of the dataset used for image classification using MLC is given in Figure 2.2. Moreover, if the target variables from the original MTP task are structured in a form of hierarchy, the ML task at hand is known as the *hierarchical multi-target prediction (HMTP)* task. Analogously, we can define *hierarchical multi-target regression (HMTR)* and *hierarchical multi-label classification (HMLC)* task as a variants of the HMTP task.

Image	Descriptive attributes				Target attributes						
	╲	∿	—	...	*traffic light*	*car*	*truck*	*building*	*traffic sign*	*bridge*	*tree*
	32	3	54	...	***5***	***5***	***1***	***3***	***4***	***0***	***5***
	55	43	1	...	***18***	***6***	***2***	***3***	***3***	***1***	***9***
	23	4	5	...	***6***	***2***	***0***	***4***	***2***	***0***	***2***
	23	4	5	...	***8***	***3***	***2***	***9***	***0***	***0***	***5***
	21	2	4	...	***4***	***1***	***0***	***0***	***3***	***0***	***3***
...	...	...	...	...	...	...	...	...	...	...	...

Figure 2.1: An illustrative example of a multi-target regression (MTR) dataset. The table contains a set of images described with visual features extracted from the images, such as the number of lines of different type (curved, horizontal, diagonal, etc.) and target attributes given as numerical values (numbers of objects of the given types).

The formal definition of the multi-target prediction (MTP) task, covering both multi-label classification (MLC) and multi-target regression (MTR) task, is the following (Džeroski, 2006; Kocev et al., 2013):
Given:

- A descriptive (input) space,a Cartesian product of D descriptive attributes, i.e., $X = X_1 \times X_2 \times \cdots \times X_D$;

  - In MLC: An output (label) space $Y = 2^{\mathcal{L}}$, which consists of all possible subsets of a finite set of disjoint labels $\mathcal{L} = \{\lambda_1, \lambda_2, \ldots, \lambda_Q\}$ and $Q > 1$; An alternative view of Y in MLC is $Y = Y_1 \times Y_2 \times \cdots \times Y_Q$ where $Y_i = \{0, 1\}$
  - In MTR: An output (target) space Y spanned by T continuous target variables, i.e., $Y = Y_1 \times Y_2 \times \cdots \times Y_T$;

- A set of examples I consisting of pairs of elements, one from the input and one from the output space, accordingly.

- A quality criterion q, which selects and chooses the models with the lowest predictive error.

Find:

- A function f that maximizes quality criterion q such that:
  - In MLC: $f : X \to 2^{\mathcal{L}}$;
  - In MTR: $f : X \to Y$.

Image	Descriptive attributes				Target attributes						
	╲	∫	—	…	*traffic light*	*car*	*truck*	*building*	*traffic sign*	*bridge*	*tree*
	32	3	54	…	⊤	⊤	⊤	⊤	⊤	⊥	⊤
	55	43	1	…	⊤	⊤	⊤	⊤	⊤	⊤	⊤
	23	4	5	…	⊤	⊤	⊥	⊤	⊤	⊥	⊤
	23	4	5	…	⊤	⊤	⊤	⊤	⊥	⊥	⊤
	21	2	4	…	⊤	⊤	⊥	⊥	⊤	⊥	⊤
…	…	…	…	…	…	…	…	…	…	…	…

Figure 2.2: An illustrative example of a multi-label classification (MLC) dataset. The table contains a set of images described with visual features extracted from the images, such as the number of lines of different type (curved, horizontal, diagonal, etc.) and target attributes denoting the presence ($\top$) or absence ($\bot$) of objects of different types.

2.3 Hierarchical Multi-Target Prediction (HMTP)

If the output space in the ML task for SOP is represented as a hierarchy of target attributes, the ML task at hand is known as the *hierarchical multi-target prediction (HMTP)* task.

The main difference with the definition of the MTP task is in the representation of the output space. In HMTP, the targets are structured in a hierarchical format and each node in the target hierarchy (*meta-label*) is a result of an aggregation function on their respective children. Therefore, the formal definition of the HMTP task is the following:
Given:

- A descriptive (input) space X spanned by D independent descriptive variables, $X = X_1 \times X_2 \times \cdots \times X_D$;

- A target (output) space Y spanned by T target variables, $Y = Y_1 \times Y_2 \times \cdots \times Y_T$. Note that $Y_i \subseteq \mathbb{R}$ for MTR and $Y_i = \{0, 1\}$ for MLC. We define a hierarchy $\mathcal{H} = (Y, \leq_h)$ for the variables from the output space Y. The relation " $\leq_h$ " represents a parent-child relationship between tree nodes ($\forall (Y_1, Y_2) \in \mathcal{H} : Y_1 \leq_h Y_2$ if and only if Y_2 is a parent (meta-label) of Y_1) and defines a *hierarchical constraint*. In the HMTP task, the parent-labels are the result of an *aggregation function* on their respective children, i.e $Y_k = agg\{Y_i | Y_i \leq_h Y_k\}$;

- Set of examples E consisting of pairs of elements, one from input and another from output space, accordingly, i.e., $E = \{(x_i, y_i) | x_i \in X, y_i \in Y, 1 \leq i \leq N\}$, where N is the number of examples and where the values of the target variables satisfy the hierarchical constraint " $\leq_h$ " i.e $\forall j : \exists i (y_i \leq_h y_j \implies y_j = agg\{y_i | y_i \leq_h y_j\})$;

- A quality criterion q which selects and chooses the models with the lowest predictive error and the highest accuracy.

Find:

- Function $f : X \to Y$, which maximizes the quality criterion q, and all predictions $\hat{y} = f(x)$ are satisfying the hierarchical constraint.

The difference from the task of MTP is in the definition of the output space: for HMTP, we have a set of classes/targets organized in a hierarchy instead of a flat tuple of classes/targets.

According to the general definition of the HMTP task given above, we can define different HMTP tasks based on the aggregation function. HMLC task definition is one variant of the HMTP definition, where tuples of target variables are represented as a set of labels/classes, structured in a form of hierarchy. The values of those classes/labels in each example are Boolean values ($\top$ or $\bot$) that can be represented with binary values (1 or 0), accordingly. Then, by instantiating the aggregation function as *logical OR*, we obtain hierarchical constraint as defined by Vens et al. (2008) i.e., we define the HMLC task. An example of a dataset for the HMLC task is given in Figure 2.3.

In HMTR task (Mileski et al., 2017), the output space is represented by tuples of T continuous target variables organized as a hierarchy. In HMTR, there are more possibilities for instantiating the aggregation function that will define the hierarchical constraint. Possible choices are sum, max, min, average, etc. The definition of the parent-child relationships (hierarchy constraint) states that the variable belonging to a given hierarchy node automatically contributes to all its parent nodes. Therefore, it is very important to notice that, during calculation of the meta-labels by using the aggregate functions, we have to carefully select the aggregation function and the (prototype) function for calculating the predictions, in order not to break the hierarchical constraint.

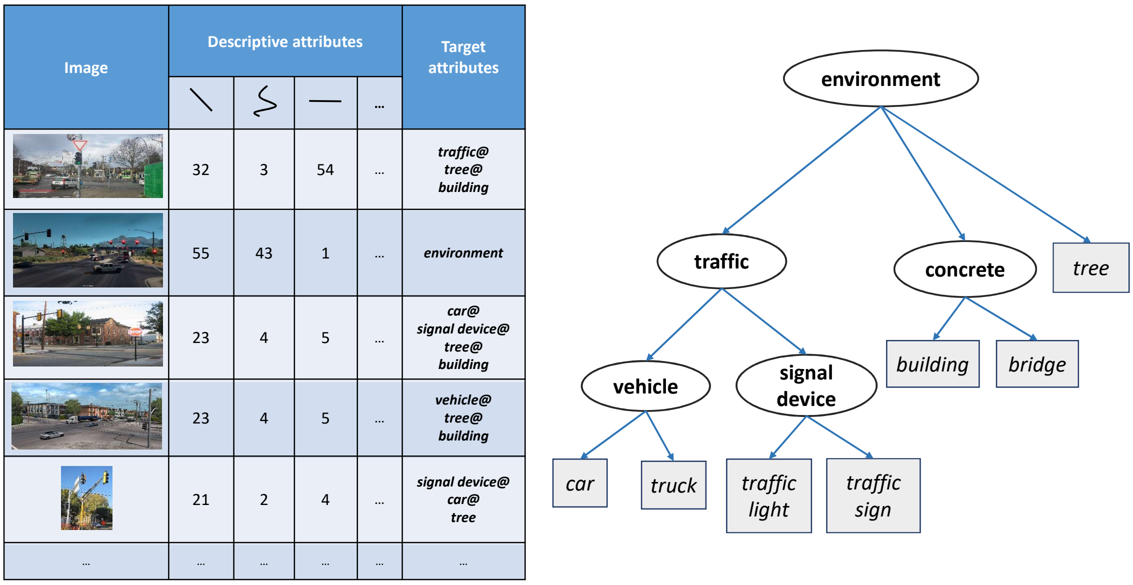


Image	Descriptive attributes				Target attributes
	╲	∫	—	...	
	32	3	54	...	traffic@ tree@ building
	55	43	1	...	environment
	23	4	5	...	car@ signal device@ tree@ building
	23	4	5	...	vehicle@ tree@ building
	21	2	4	...	signal device@ car@ tree
...	...	...	...	...	...

Figure 2.3: An illustrative example of a hierarchical multi-label classification (HMLC) dataset. The table contains a set of images described with visual features extracted from the images, such as the number of lines of different type (curved, horizontal, diagonal, etc.) and target values given as a hierarchical representation of labels, i.e., strings where child labels from different branches of the hierarchy are separated by the symbol '@'.

2.4 Semi-Supervised Learning for SOP

The specific tasks for structured output prediction covered in this dissertation are based on the data labeling given in Figure 2.4. The data example is *labeled* if all of the target attribute values are known, i.e., labeled. The data consisting of labeled data examples is called *labeled data*. The specific machine learning task for SOP dealing with labeled data is known as the *supervised* learning task. The data examples are *unlabeled* if all values of the target attributes are unknown, i.e., unlabeled. The data consisting of unlabeled examples is called *unlabeled data*. The specific ML task for SOP that handles only unlabeled data is known as the *unsupervised* learning task. In the SOP task, we have another type of data examples where not all of the target attributes are labeled, i.e., target attributes are partially-labeled. We called those data examples *partially-labeled* examples. The data consisting of partially-labeled examples is called *partially-labeled data*. We can distinguish between two task-dependent 'incomplete' examples. In the classical (i.e., single-target) learning tasks, as an 'incomplete' example can be considered the unlabeled examples i.e., the example where the class value is unknown, while in the SOP task, as an 'incomplete' example, beside the completely unlabeled examples, we can consider partially-labeled examples, where only a part of the target values are known/labeled.

Motivated by data incompleteness, i.e., differently labeled data, we define another machine learning task for SOP that is 'in-between' supervised and unsupervised ML task, known as the *semi-supervised* learning (SSL) task. The classical semi-supervised learning (SSL) tasks can handle both, unlabeled and labeled examples, while the SSL task for SOP can handle partially-labeled examples in addition to the labeled and unlabeled data examples. In practice, there is a large amount of 'incomplete' data which consists of partially-labeled and unlabeled data examples. Therefore, the very important advantage of the semi-supervised over supervised learning approaches is that the former considers the additional information of the 'incomplete' examples and the latter usually discards them from the learning process. As a result, the semi-supervised models achieve better predictive performance, use more examples for learning and could be better explained.

Next, we will describe two different semi-supervised learning variants, one that can learn from unlabeled and another that can learn from partially-labeled data, in addition to labeled data.

2.4.1 Semi-supervised learning (SSL) with labeled and unlabeled data

As the name suggests, *semi-supervised learning (SSL)* belongs somewhere between the supervised and unsupervised learning paradigm and the motivation behind this learning task is related to the high availability of additional information contained in the unlabeled data. In fact, it can be considered as an extension of either supervised or unsupervised learning. As the extension of the supervised learning, SSL is also called *semi-supervised prediction*, where the labeled training data is enriched by unlabeled training examples. An excerpt of the MTR dataset in the SSL setting is given in Table 2.1.

The goal of the semi-supervised prediction is to build a model from labeled and unlabeled data with better prediction quality than the model learned from labeled examples only. On the other hand, as the extension of unsupervised learning, SSL is also known as *constrained clustering*, where the data clusters consisting of unlabeled examples are enriched with additional constraint, i.e., "supervised information" about the clusters. The constraint is called *"must-link"* constraint if two data examples are close in the descriptive space, i.e., are in the same cluster or *"cannot-link"* if two data examples are in different clusters (Zhu & Goldberg, 2009).

In order to illustrate the semi-supervised learning with unlabeled in addition to labeled examples, we will present a very simple example shown in Figure 2.5. Let us assume that each example is represented as a one-dimensional feature $x \in \mathbb{R}$. There are two

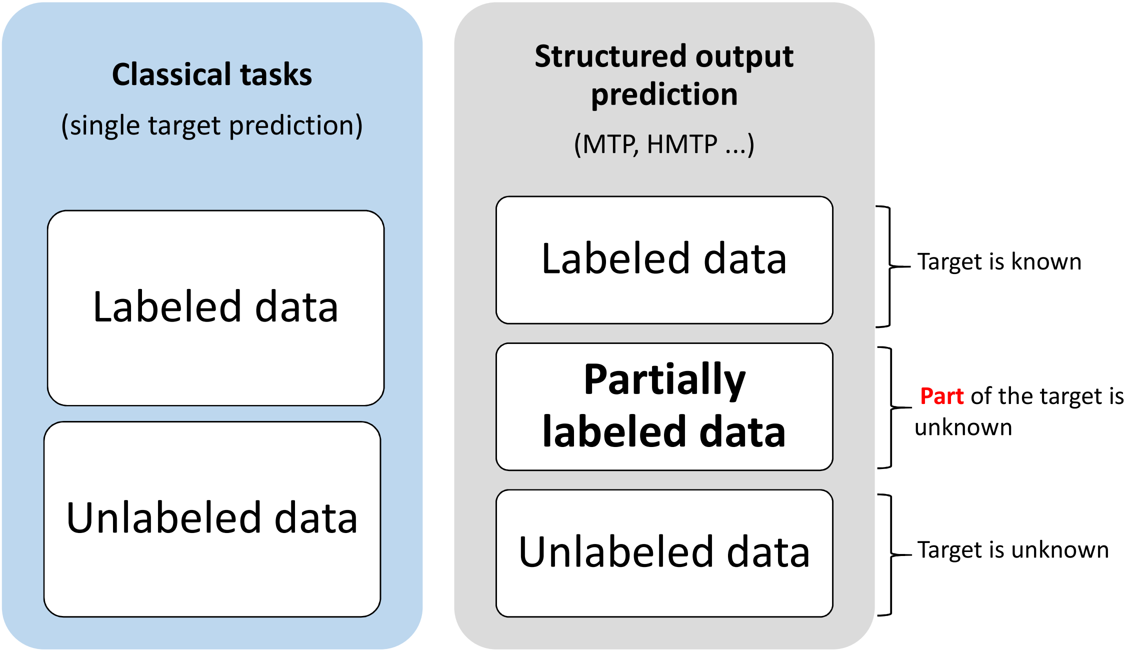


Figure 2.4: An illustration of the semi-supervised learning in structured output prediction with respect to different data labeling.

Table 2.1: A multi-target regression (MTR) dataset in the semi-supervised learning (SSL) setting. ’?’ denotes a missing value.

ID	**Description Space**				**Target Space**			
#	**Hi drain q1**	**drainage factor**	**Total N input**	**...**	**Q**	**P**	**N**	**Example type**
#1	3	0.35	121.26	...	**3.86**	**0.04**	**0.47**	Labeled examples
#2	2	0.25	119.86	...	**4.28**	**0.34**	**20.71**	
#3	2	0	120.4749	...	**4.35**	**0.024**	**0.43**	
#4	2	0.24	120.95	...	**3.47**	**0.155**	**0.15**	
#5	2	0.15	119.2568	...	?	?	?	Unlabeled examples
#6	3	0	121.0236	...	?	?	?	
#7	2	0.86	119.3698	...	?	?	?	
#8	2	0	115.3987	...	?	?	?	

classes: positive (+1) and negative (-1). In supervised learning, two examples are given, one positive, shown as a green circle, and one negative, shown as a red cross, respectively. The best estimate of the decision boundary is $\mathbf{x} = 0$ and if the class value of a new example is $x < 0$, then an example is considered as negative (-1), otherwise, if the class value of a new example is $x \geq 0$, it is considered as positive (+1).

In addition, a large number of unlabeled examples are also given there, shown as blue dots in Figure 2.5. We do not know the correct class assignment of those examples. Under the assumption that the unlabeled examples are normally distributed such that the examples from each class are centering around the central mean. Obviously, the prototype of the two labeled examples does not hold anymore for making a decision on the new class assignments. In semi-supervised learning, the estimate of the decision boundary should be somewhere between the two groups, i.e., $\mathbf{x} \approx 0.4$.

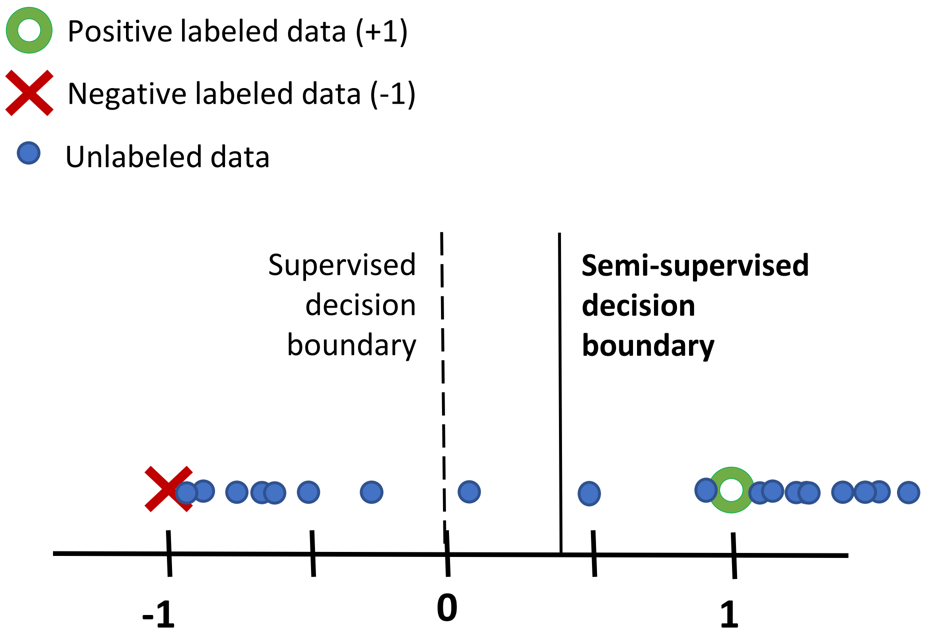

Positive labeled data (+1)
Negative labeled data (-1)
Unlabeled data
Supervised decision boundary
Semi-supervised decision boundary
-1
0
1

Figure 2.5: A simple example to demonstrate semi-supervised learning with labeled and unlabeled data.

2.4.2 Semi-supervised learning (SSL) with partially-labeled data

As we mentioned previously, machine learning methods which use unlabeled examples in addition to labeled ones, aiming to improve the performance of supervised methods, are called *semi-supervised learning* (SSL) methods (Chapelle et al., 2006; Levatić et al., 2018). Note that learning from partially-labeled data can be considered as semi-supervised learning for SOP.

An example excerpt from a SSL dataset with partially labelled examples is given in Table 2.2. It contains all three possible kinds of examples: unlabeled, partially labeled and (fully) labelled examples. The partially labeled examples include all six possible combinations of known and missing values of the three targets: three examples have only one target value known and three examples have only one target value unknown.

Table 2.2: A MTR dataset with partially labeled examples. '?' denotes a missing value.

ID	**Description Space**				**Target Space**			
#	**Hi drain q1**	**drainage factor**	**Total N input**	**...**	**Q**	**P**	**N**	**Example type**
#1	3	0.35	121.26	...	**3.86**	**0.04**	**0.47**	Labeled examples
#2	2	0.25	119.86	...	**4.28**	**0.34**	**20.71**	
#3	2	0	120.4749	...	**4.35**	**0.024**	**0.43**	
#4	2	0.24	120.95	...	**3.47**	**0.155**	**0.15**	
#5	2	0	120.26	...	**3.7**	**?**	**?**	Partially labeled examples
#6	3	0.75	121.26	...	**2.95**	**?**	**5.71**	
#7	2	0.565	120.47	...	**4.4**	**0.026**	**?**	
#8	3	0	116.86	...	**?**	**0.68**	**?**	
#9	3	0.547	116.56	...	**?**	**?**	**0.21**	
#10	2	0.65	118.36	...	**?**	**0.13**	**0.11**	
#11	2	0.15	119.2568	...	**?**	**?**	**?**	Unlabeled examples
#12	3	0	121.0236	...	**?**	**?**	**?**	
#13	2	0.86	119.3698	...	**?**	**?**	**?**	
#14	2	0	115.3987	...	**?**	**?**	?	

Usually, in SSL, the class information of unlabeled examples is entirely missing. Partially-labeled examples can also be considered as an additional source of information in the spirit of SSL. Namely, the known information of the descriptive attributes for unlabeled and partially-labeled examples can be exploited in order to improve the prediction quality and model itself. Here, we formally define the task of semi-supervised learning with partially-labeled data, as follows:

Given:

- A description (input) space $\mathcal{X}^D$ spanning D descriptive variables, i.e.,

$$\mathcal{X}^D = X_1 \times X_2 \times \cdots \times X_D,$$

where X_i is the set of possible values of the $i-th$ descriptive variable;

- A target (output) space $\mathcal{Y}^T$ spanning T target variables, i.e.,

$$\mathcal{Y}^T = (Y_1 \cup \{?\}) \times (Y_2 \cup \{?\}) \times \cdots \times (Y_T \cup \{?\}),$$

where Y_j is the set of possible values of the $j-th$ target variable, extended with potentially missing value (denoted as ?);

- A set I of N examples (x, y) where $x \in X^D$ and $y \in \mathcal{Y}^T$ and an example (x, y) is

$$\begin{cases} \text{fully- labeled,} & \text{if } \forall i \in \{1, 2, \ldots, T\} : y_i \in Y_i \\ \text{unlabeled,} & \text{if } \forall i \in \{1, 2, \ldots, T\} : y_i = ? \\ \text{partially- labeled,} & \text{otherwise} \end{cases} .$$

- A quality criterion q, which rewards the models with the lowest predictive error.

Find:

- a function $f : \mathcal{X}^D \rightarrow \mathcal{Y}^T$ by using the set of examples I, such that f maximizes the quality criterion q.

Depending on the example set I, changing with respect to the output (target) space $\mathcal{Y}^T$, we can distinguish between different MTR tasks. If there are labeled examples only in I, then we have the classical supervised multi-target regression task (MTR). We define the task of semi-supervised learning for multi-target regression (SSL for MTR) if we have only fully labeled and unlabeled examples in I in addition (Levatić et al., 2018).

We must note that, at first glance, it might seem paradoxical learning with the predictor $f : X \rightarrow Y$ from unlabeled and partially-labeled data, because the unlabeled data does not provide any examples of such mapping. The answer lies in the assumption of the link between the distribution of unlabeled data and the target label, i.e., the information given in the descriptive space of the unlabeled examples.

2.5 Modeling Soil Functions

In this section, we will describe the soil functions defined within the LANDMARK EU H2020 project (LANDMARK, 2019). The case studies presented in the dissertation (Chapter 6) are related to modeling two soil functions: primary productivity (i.e., estimation of total herbage production and nutrient uptake) and water purification and regulation (i.e., water quality assessment), by using the existing machine learning predictive modeling techniques for multi-target regression.

Soils are base providers of food, feed and fiber for humans and animals, respectively, and play a key role for functioning of terrestrial ecosystems. During the recent decades, a need to establish methods to evaluate the ability of soils to provide ecosystem services has moved towards the top of the agenda in soil science. The expansion of settlements and infrastructure, the development of industry and transport, the emergence of landfills, mining and intensive agriculture all affect the soils and their functioning. Deterioration of soil characteristics usually occurs as a result of human activity, and leads to degradation of one or more soil ecosystem services.

Scientific community pays considerable focus on the classification and valuation of individual soil ecosystem services with several resulting classification schemes, but limited consensus on a comprehensive framework (Jónsson & Davídsdóttir, 2016). Consequently, limited attention is devoted to understanding co-functionality of such ecosystem services and effects of coordinated management on total natural capital and long-term sustainable goals (Pereira et al., 2018). The most recent comprehensive endeavor of classification, valuation and simultaneous management of soil ecosystem services has been performed within the activities of EU H2020 project LANDMARK (LANDMARK, 2019). Project

LANDMARK addresses the necessity to strike a balance between interests of policy-making bodies on different spatial scales (local and regional) to achieve a harmonious use of soils and to limit the lowering of balance in the overall ecosystem by enforcing sustainable management.

To ease the presentation of applicability of our proposed methodology for structured output prediction with interpretable models on a real-world problem, i.e., modeling soil ecosystem services, the classification and valuation of soil ecosystem services, proposed within research activities of project LANDMARK (further referred to as *LANDMARK classification and/or valuation framework*), have been adopted.

LANDMARK classification framework (Schulte et al., 2014) focuses on functional capacity of soils to directly contribute towards the delivery of ecosystem services referred to as *soil functions* – being a 'demand' of particular service. Valuation, on the other hand, is defined as a 'supply' to a particular demand, i.e., soil function. The framework adopts five soil functions: primary productivity (agriculture), water purification and regulation, carbon sequestration and climate regulation, provision of functional and intrinsic biodiversity, and provision and cycling of nutrients. Composition of all five soil functions with emphasis on their inter-connectivity and cohabitation is given in Figure 2.6.

Primary productivity is the capacity of a soil to produce plant biomass for human use, providing food, feed, fiber and fuel within natural or managed ecosystem boundaries (white box).

Water purification and regulation is the capacity of a soil to remove harmful compounds from the water that it holds and to receive, store and conduct water for subsequent use and the prevention of prolonged droughts and flooding and erosion (blue box).

Carbon sequestration and climate regulation is the capacity of a soil to reduce the negative impact of increased greenhouse gas (i.e., CO_2, CH_4, and N_2O) emissions on climate (black box).

Figure 2.6: Illustrative representation of the five soil functions (Schulte et al., 2014).

Provision of functional and intrinsic biodiversity is the multitude of soil organisms and processes, interacting in an ecosystem, making up a significant part of the soil's natural capital, providing society with a wide range of cultural services and unknown services (green box).

Provision and cycling of nutrients is the capacity of a soil to receive nutrients in the form of by-products, to provide nutrients from intrinsic resources or to support the acquisition of nutrients from air or water, and to effectively carry over these nutrients into harvested crops (purple box).

Classification and valuation of soil ecosystem services into the five soil functions boosts the ability to improve their modeling both, individually and simultaneously. The former allows better understanding of the soil function, while the latter imposes better understanding of their competition and trade-offs with a goal to improve overall management on a local or regional scale.

Thus, for modeling purposes, LANDMARK valuation framework defines each soil function as an index with qualitative (indicative) value scale, composed from factors that can be valued with existing indicators or indices already familiar to the domain experts. Such lower level indicators or indices represent natural processes or phenomena underneath soil functions, which are grouped into three categories based on their nature: *soil (S)*-related (biological, chemical and physical), *environmental (E)* -related (humidity, temperature, hydrology, etc.) and *management (M)*-related (drainage, tillage, nutrient and pest management, etc.) processes or phenomena. Thus, mathematically, each soil function is a function of the soil (S), environmental (E) and management (M) factors, i.e., $SF = f(S \times E \times M)$.

For sake of clarity, each lower level indicator or index is referred to as *proxy-indicator* that itself can be further decomposed or dependent on other set of proxy-indicators. The hierarchical dependency ends with quantifiable factors that can be expressed quantitatively or qualitatively. In modeling setting, such factors are referred to as *attributes*.

Modeling activities within LANDMARK have been performed with knowledge-based and data-driven approaches of artificial intelligence (AI), using DEX (Decision EXpert) methodology for qualitative multi-criteria decision modeling (Bohanec, 2014, 2017; Bohanec & Rajkovič, 1990) and various machine learning (ML) methods, respectively.

Primary outcome of the modeling activities is an online decision support tool (DST), so-called, *Soil Navigator* for management and simultaneous optimization of all five soil functions on a field scale (Debeljak et al., 2019). The Soil Navigator DST operates on knowledge-based implementation through DEX methodology, where each soil function is integrated with two DEX models – one for each, cropland and grassland ecosystem, correspondingly. However, the experts have been using machine learning for modeling soil functions in order to better understand the integration of proxy-indicators and attributes, and consequently improve the DEX models.

The thesis follows the LANDMARK modeling framework and each soil function with defined hierarchical composition, in order to present the applicability of our proposed methodology and improvements in modeling certain indices, as well as proxy-indicators. Therefore, the modeling outcomes per soil function, achieved within the project LANDMARK, are described in Section 4. Along with modeling LANDMARK achievements, a related work with AI methods for modeling soil functions or subsequent proxy-indicators is given.

Chapter 3

Related Work

3.1 Methods for Multi-Target Prediction

Methods for the multi-target prediction (MTP) task have been extensively researched in the last decades since a lot of domains aim to estimate more than one output attributes simultaneously. We can distinguish between two categories of methods for solving the MTP task (Bakır et al., 2007; Borchani et al., 2015). One category is concentrated on the task itself and the other on the algorithm that solves the MTP task. In order to solve the MTP task, the former category of algorithms transforms the task itself and they are known as *problem transformation (i.e., local)* methods. Local methods construct t separate models for the t predictive variables which are combined to give the overall prediction for all the predictors. On the other hand, the latter category relies on tuning of the algorithm in order to solve the MTP task and they are known as *algorithm adaptation (i.e., global)* methods. Global methods build only one model for predicting all of the t predictive variables simultaneously.

The main drawback of local and global models is that they are not considering the target dependencies represented by the parent-child relationships in target attributes, but they deal with the target space as with a set of independent target attribute vectors. The methods that are filling the gap caused by the way the local and global MTP methods handle the target attributes are called methods for *structuring the output space*. The methods for structuring the output space can be considered as a task transformation methods, because they transform the classical MTP task to hierarchical MTP (HMTP) task and solve the MTP task by learning from hierarchical structured output space od labels/targets. The methods concerned with learning from hierarchically structured target attributes are called methods for the *hierarchical multi-target prediction (HMTP)* task.

In the next subsections, we will describe the related work for both, problem transformation and algorithm adaptation methods for solving the most representative MTP tasks, multi-label classification and multi-target regression as well as the currently known methods for solving the HMTP task. Finally, we will present the most prominent research attempts for solving the problem of structuring the output space.

3.1.1 State-of-the-art methods for multi-label classification

As we mentioned before, multi-label classification is a special case of multi-target prediction, where the output space is given by a set of vector rows with Boolean data type values representing the label co-occurrences (the "$\top$" value for current label means the occurrence of that specific label and the label value "$\perp$" shows that the label does not occur for a given arbitrary data example). Next, we will present the local and global state-of-the-art methods for solving the multi-label classification (MLC) task.

Local (problem transformation) methods for MLC The simplest problem transformation (i.e., local) method for solving the MLC task is called *binary relevance* (M.-L. Zhang et al., 2018). This approach decomposes the initial MLC task with t labels into t independent binary learning (single-target classification) tasks and each binary learning task is solved separately. Binary relevance method has a lot of crucial disadvantages such as limited usefulness in solving MLC problems with high-dimensional label spaces, not considering label dependencies, introducing imbalance in the example space (Zhou et al., 2012). However, Read et al. (2011) present some advantages of using this method such as linear scaling with the increase of labels as well as the possibility of parallelization of the training process because of independence of the binary tasks.

The next well-known problem transformation (i.e., local) method for solving the MLC task is *classifier chains* (B. Chen et al., 2016; Enrique Sucar et al., 2014; Read et al., 2011). This method is based on the binary relevance approach and during the learning process it generates chains of binary classifiers with permutation of labels. The predictions obtained from each single-label classifier (i.e., binary learning task) are added to the training set before learning the next classifier in the chain. Only in learning of the first binary classifier in the chain, the original input variables are used. Recently, some variants and improvements of classifier chains approach are proposed. Alali and Kubat (2015) present a method called PruDent, for chaining the layers of two binary classifiers. Similarly like in the classical classifier chain approach, the predictions obtained from binary relevance models in the first layer are used as an input to the classifier in the second layer in the chain, etc. Next, Read et al. (2015) address to the scalability issue of the classifier chain algorithm by proposing a large-scale approach for efficient learning from a large number od labels based on a hill climbing heuristic: the classifier trellis.

Another popular problem transformation (i.e., local) method for solving the MLC task is the *label power set* method (Boutell et al., 2004; Read et al., 2008; Tsoumakas & Vlahavas, 2008). This approach transforms the MLC problem into a multi-class classification problem. Each possible subset of labels in the data set becomes a new meta-class. Therefore, this approach has problems with scalability because the number of combinations of label subsets grows by adding a new label. The scalability issue has been tackled by Read et al. (2008) by pruning the set of all possible label subsets (i.e., label power set).

Binary pairwise is another problem transformation approach for solving the MLC task concerning with learning a binary classification model for each pair of labels (Fürnkranz et al., 2008). For each pair of labels (λ_1, λ_2), an artificial calibrated label λ_0 is created for each example, such as distinguish between a set of positive (i.e., examples labeled with λ_1) and negative (i.e., examples labeled with the λ_2) examples. Then, binary relevance classifier is used for learning from each augmented data set consisting of calibrated label λ_0 for each pair of labels that make this approach computationally expensive. Motivated by calibrated binary approach, Madjarov et al. (2012) propose a two-stage approach where in the first stage, binary relevance models are learned and the second stage consists of pairwise models with a calibrated label for each pair of labels.

Global (algorithm adaptation) methods for MLC In the last decades, some neural network-based problem adaptation algorithms were developed towards solving the MLC task. M.-L. Zhang and Zhou (2006) proposed an adaptation of the well-known backpropagation neural network algorithm (Riedmiller & Braun, 1993) by introducing a novel error function for label ranking according to their belonging to an instance, i.e., labels belonging to the instance are ranked higher than those that are not belonging. Next, M.-L. Zhang (2009) presents a neural network-based algorithm for multi-label learning induced by a well-known radial basis function (ML-RBF) and achieves competitive results compared to the other state-of-the-art methods for solving the MLC task.

Next, M.-L. Zhang and Zhou (2007) adapt the k-nearest neighbor (kNN) method proposed by Cover and Hart (1967) towards solving the MLC task. Using the proposed ML-kNN method, the label(s) for an unseen example were determined by using maximum a posteriori principle making this method superior against the other well-established MLC methods. Moreover, for solving the MLC task, kNN algorithm can be used in combination with binary relevance (Spyromitros-Xioufis et al., 2008) and logistic probability model (Cheng & Hüllermeier, 2009). Furthermore, Pugelj and Džeroski (2011) proposed a method for structured output prediction by using kNN which can be adapted for solving the MLC task.

Support vector machines (SVMs) (Cortes & Vapnik, 1995; Elisseeff & Weston, 2001; Hofmann et al., 2008) are very well-established MLC methods, especially in the domain of text classification (T. Gonçalves & Quaresma, 2004; Joachims, 1998). In order to solve the MLC task, many approaches exist that can combine SVMs with classifier chains (E. C. Gonçalves et al., 2013) and/or with binary classifiers (W.-J. Chen et al., 2016). Jayadeva et al. (2007) present an adapted SVMs algorithm, called twin SVMs, which is computationally efficient and discovers two dimensional projections of the data since it uses two non-parallel separating planes by solving two simpler SVMs tasks.

Predictive clustering trees (PCTs) are tree-based methods built within the predictive clustering framework (Blockeel et al., 1998). This framework learns decision trees called predictive clustering trees (PCTs) where the top node contains all of the training examples and then it recursively splits into lower partitions (clusters) of the whole train set. PCTs can be used for various multi-target prediction tasks, including MLC. Kocev et al. (2013) propose bagging and random forests of PCTs as ensemble methods for solving the MLC task and present the improvements of the predictive performance over the classical (i.e., single PCTs) scenario.

Decision trees can be used in combination with many of the well-established methods for solving the MLC task. Gjorgjevic et al. (2013) propose a hybrid method that generates a tree using decision trees and then use SVMs for making the predictions for the individual labels in the tree leaves. Next, Madjarov and Gjorgjevikj (2011) present an approach which combines multi-label model trees with single-label SVMs in such a way that first, they induce the ML model tree and then use binary relevance SVMs for the calculation of a prediction for each label. Moreover, Wu et al. (2015) propose the ML-TREE approach consisting of decision trees that internally use SVMs in such a way that at each test node, an individual one-vs-all SVM classifier is trained for the evaluation of node splits, which makes this approach computationally expensive. Furthermore, decision trees can be used in various domains, such as the domain of gene function prediction. Clare and King (2001) propose decision trees with a modified entropy function for the calculation of the splits, learned on gene expression data.

3.1.2 State-of-the-art methods for multi-target regression

Local (problem transformation) methods for MTR Since the local methods transform the problem into t separate single-target models, any known single target regression algorithm can be used to learn the single-target models. Prominent methods addressing the MTR task include: *ridge regression* (Hoerl & Kennard, 1970), *support vector regression machines* (W. Zhang et al., 2012), *regression trees* (Breiman, 1996) and *stochastic gradient boosting* (J. Friedman, 2002). Hoerl and Kennard (1970) proposed a separate ridge regression algorithm that deals with MTR problems.

Regressor chain (RC) (Spyromitros-Xioufis et al., 2012) is another problem transformation method motivated by the multi-label chain classifier (Read et al., 2011). During the training process, RC randomly selects a chain (permutation) of the target space, then builds a separate regression model for each target in consistence with the selected chain. Since RC uses the actual values of all previous targets in a chain, Spyromitros-Xioufis et al. (2012) also proposed regressor chain *corrected* (RCC) that uses cross-validation estimates instead of actual values. However, RC and RCC are sensitive to the selected chain ordering. In order to avoid this problem, Spyromitros-Xioufis et al. (2016), proposed ensemble approaches called ensemble of regressor chains (ERC) and ensemble of regression chains corrected (ERCC), where they randomly select as many models as the number of distinct label chains if the number of labels is less than 10. Otherwise, they randomly selected 10 chains and constructed an ensemble of chains.

Multi-target regressor stacking (MTRS) (Spyromitros-Xioufis et al., 2012) is another problem transformation method inspired by Godbole and Sarawagi (2004) where multi-label classification is performed by using stacked generalization. MTRS training is performed in two stages. First, t different single-target models are learned and then, instead of concatenating the t obtained predictions, MTRS includes an additional training stage, where a second collection of t separate single target meta-models are learned. At the end, the predictions are calculated from both stages. The predictions from the second stage use and adjust the predictions from the first stage.

W. Zhang et al. (2012) presented a new problem transformation method based on the multi-output support vector regression approach. Basically, they extend the actual feature space and represent the multi-output problem as equivalent single-output problems, that are solved using the single-output least squares SVRs (LS-SVR) algorithm. The multi-output model takes into account the target correlations by using the vector virtualization method.

Recently, J. Wang et al. (2019) propose a multi-target regression method (MTR-TSF) that embeds the intra-target relationships. First, by using hierarchical clustering on the output space, they reveal the correlation among the targets and create an additional feature vector X_{index} consisting of the indices of the nodes where specific instances belongs to. Next, they use a boosting regression algorithm to learn a similarity matrix for each target. Finally, by querying and clustering the similarity matrix, a target-specific feature vector X_{tsf} is created for all instances and is added to the original feature vector X. At the end, a prediction model per target is learned by considering the 'merged' feature space $X' = X \bigcup X_{index} \bigcup X_{tsf}$.

Global (algorithm adaptation) methods for MTR Algorithm adaptation methods learns a single model for all target variables and thus take into account the dependencies among the targets. There are many advantages over the local methods such as interpretability, better predictive performances, especially, if the targets are related (Kocev et al., 2013). Below, we briefly discuss various algorithm adaptation methods proposed in the literature.

The first attempt to deal with the prediction of multiple target variables are the statistical methods such as *reduced-rank regression* (Izenman, 1975). Furthermore, van der Merwe and Zidek (1980) proposed the general version of a multivariate regression problem of the James-Stein estimator, called *filtered canonical y-variate regression*. Next, lasso regression Tibshirani (1996) is a popular regression method for the estimation in linear models. It produces interpretable models while at the same time it is stable. Next, gaussian processes for MTR are based on the algorithm proposed by Rasmussen and Williams (2006). The most prominent statistical approach that deals with multiple targets is the *curds and whey*

(C & W) method (Breiman & Friedman, 1997).

Predictive clustering trees (PCTs) are tree-based methods built within the predictive clustering framework (Blockeel et al., 1998). Similarly for MLC, they can be used to solve MTR tasks. We presented a detailed description earlier. In addition, Appice and Džeroski (2007) presented an algorithm called multi-target step-wise model tree induction (MTSMOTI) for generating a multi-target model tree in a step-wise manner. The tree model is generated similarly as in PCTs, with the TDIDT algorithm. The difference is that each leaf in a tree model is associated with a set of linear models which generate the final target predictions. Conditional Inference Trees (CTrees) are non-parametric regression trees embedding tree-structured regression models into conditional inference procedures and estimate a regression relationship in a multi-target scenario (Hothorn et al., 2006).

A different type of the MTR algorithm is the rule-based algorithm called *FItted Rule Ensemble (FIRE)* method, proposed by Aho et al. (2009). This is a method for learning rule ensembles based on representing an ensemble of regression trees as a large collection of rules. FIRE uses an optimization procedure (minimization) to select the best (much smaller) set of rules and determine their respective weights. Furthermore, Breskvar et al. (2018) present an ensemble method with random output selection (ROS). Instead of using all target attributes, they randomly select subsets of target attributes when learning the base predictive models of the ensemble. This additional randomization improves the performance both in terms of time complexity and predictive accuracy.

The most famous non-parametric distance-based method for the regression task is the k-nearest neighbor method (Altman, 1992). It takes the average of the values of the k nearest examples as a prediction. K-nearest neighbour is a flexible algorithm, since it can use any distance function and any number k (nearest neighbours) (Pugelj & Džeroski, 2011).

Multiple-input multiple-output (MIMO) support vector regression method is a generalization of support vector machines (SVMs) (Cortes & Vapnik, 1995) for addressing the MTR task. The generalization is achieved by minimization of a Lagrangian equation which has multi-dimensional parameters that have to be optimized (Brouwer et al., 2014; Sánchez-Fernández et al., 2004).

Partial Least Squares Regression (PLS-PLSR) and Principal Component Regression (PLS-PCR) methods are other methods for multi-target regression which are implemented in the R software package *pls* (Mevik & Wehrens, 2007). These methods are commonly used in many natural sciences and are based on the calculation of the scores obtained by decomposition of the product matrix of orthogonal scores and loadings. Then regression coefficients are calculated using the scores.

Multivariate Adaptive Regression Splines (MARS) is a non-parametric regression method implemented in *EARTH* package in R. MARS, as a generalization of the step-wise linear regression, Hastie et al. (2001) constructs the dependencies between input and output variables by using a data-driven set of base functions and coefficients.

Another well-known and widely used method for MTR are the artificial neural networks (NN). They are designed based on human brain to recognize patterns in data. They can automatically model the non-linearity and can deal with multi-input multi-output problems. The most often used algorithm for training artificial neural networks is the *back-propagation* algorithm (Riedmiller & Braun, 1993). Back-propagation algorithm is a recursive and iterative method which efficiently optimizes the network weights by following the gradient descent method that exploits the chain rule. *Deep neural networks* (DNN) are artificial neural networks containing multiple hidden layers. They update the network weights by establishing the correlation between input (past events) and output (future events). There are several variants of DNNs designed based on the specific domains that they are used for. Convolutional deep neural networks (CNNs) (Krizhevsky et al., 2017) are used in the domain of computer vision. Recurrent neural networks (RNNs) are used in various cases of language modeling, such as handwriting and speech recognition (Graves et al., 2009; Sak et al., 2014).

3.1.3 State-of-the-art methods for semi-supervised MTR

Compared to the research interest for classical supervised MTR methods, the interest for semi-supervised learning methods is scanty. Therefore, in the following part, we will present the most prominent developed SSL methods for MTR.

One important attempt of introducing the SSL paradigm is in the field of multi-task learning. Multi-task learning differs from multi-target learning in the way that it considers many different single-target learning tasks with possibly different training spaces with different descriptive attributes. Y. Zhang and Yeung (2009) proposed the supervised multi-task regression method (SMTR) based on the Gaussian process (GP) with common shared parameters for the task kernels. Moreover, they extended the SMTR towards semi-supervised SMTR (SSMTR) by handling unlabeled data with data-driven adaptation of the GP kernel parameters. Next, Cardona et al. (2015) proposed a method which extends the GP regression in the spirit of SSL by using the process convolution covariance function combined with graph-based regularization and took an advantage of unlabeled data in order to improve the performance over supervised GP for regression. Another attempt to extend the Gaussian process towards the task of semi-supervised learning has been made by Navaratnam et al. (2007) in the field of computer vision on high-dimensional spaces by sampling a large amount of unlabeled data from the marginal distribution that are improving the model fitting.

Gönen and Kaski (2014) proposed an extended kernel-based matrix factorization (KBMF) method which includes the Bayesian network approach for selecting the most informative source from a set of different kernels. The approach is adjusted to work in a semi-supervised setting in the way that the unlabeled examples are treated as "labeled with the target mean value". They tested the approach in the field of prediction drug-protein interactions in two different data sets and showed promising results. On the other hand, Brouard et al. (2016) proposed a non-parametric kernel-based regression method called input output kernel regression (IOKR) for structured output prediction. IOKR method handles unlabeled data in the output space by considering the output space as a feature space associated to a chosen output kernel assigning a higher level of supervision to the feature space. Both, KBMF and IOKR method, are very complex and hard to understand by non-experts, and are not domain-specific since they are not tested on various domains (KBMF is trained on two datasets in the domain of protein-drug prediction and IOKR on one NCI-cancer dataset in the domain of e biological activities of molecules).

Next, prominent work in the field of semi-supervised learning on discrete outputs has been done by Brefeld (2008). They used the co-training paradigm in the process of creation semi-supervised supporting vector (co-SVMs) by maximizing the agreement among multiple independent hypotheses. Moreover, they presented a transductive semi-supervised SVMs, as an unsupervised variant of co-SVMs, where the maximization has been performed on the agreement among the hypotheses on the unlabeled data.

Another very prominent work in the field of semi-supervised learning for MTR has been proposed by Levatić et al. (2017). They proposed the iterative concept of self-training by using an ensemble of PCTs for MTR. Namely, the self-learning method uses the most reliable prediction from current iteration as additional data, i.e., target labels in the next iteration. The reliability of prediction is determined by defined threshold applied to the reliability scores. However, the problem of finding the proper prediction reliability measure is still not resolved. However, Levatić et al. (2017) presented an approach of automatic determination of reliability scores by exploiting out-of-bag error when learning the ensemble.

Motivated by introduced variance/heuristic function on descriptive and predictive space proposed by Blockeel et al. (1998), Levatić et al. (2018) proposed SSL for MTR by using PCTs with an adapted variance function in the PCT algorithm by introducing a parameter ω for controlling the level of supervision in both, descriptive and predictive space. They presented an extensive study on several benchmark MTR datasets by using scenarios with different percentages (5%, 10%, 20% and 30%) of labeled data in the learning set and showed that even with the 5% of labeled data, the availability of the model is guaranteed, since the descriptive/ clustering power of the learned model is obtained from the variation in the descriptive space. The predictive performance of these models is not satisfactory because of the small information in the predictive space, but possible knowledge discovery based on finding groups of data examples could be helpful in order to obtain some insight into the potential labels of such examples.

3.2 Methods for HMTP

In this section, we present the existing (state-of-the-art) methods, related to the task of hierarchical multi-target prediction, i.e., for hierarchical multi-label classification (HMLC) and hierarchical multi-target regression (HMTR), respectively.

3.2.1 State-of-the-art methods for HMLC

Rousu et al. (2006) presented Bayesian and kernel-based methods in the domain of hierarchical multi-label classification. Their approach is working on the large margin principle for structured output prediction. The modeling pipeline of this hierarchical approach does not require additional checking of the hierarchical constraint.

Next, Koller and Sahami (1997) propose a Bayesian classifier for learning hierarchically structured text documents, where the text documents are at each leaf of the hierarchy. The hierarchy constraint is ensured by predicting one topic at each level of the hierarchy until the documents reach the bottom level and the errors which are made in higher levels of the hierarchy can not be recovered in the lower levels.

An incremental approach for HMLC is presented by Cesa-Bianchi et al. (2006). They introduce a novel approach which incrementally learns a linear classifier for each node of the taxonomy. By defining a new hierarchical loss (H-loss) function, they evaluate classifiers for each training node in a top-down fashion.

Kiritchenko et al. (2006) present a method for hierarchical text categorization by enlarging the label sets of training examples in order to achieve consistency with the given hierarchy of classes. For learning they applied a multi-class learning algorithm and the labels that were misclassified were then re-labeled.

In the domain of gene function prediction, given gene ontology (GO), Barutcuoglu et al. (2006) present a Bayesian network approach for hierarchical classification, learning separate SVMs models for each class and then combining the predictions in the Bayesian network. The combination of these two methods (Bayesian network and SVMs) for solving the hierarchical classification task has proven to be useful because the SVMs margin outputs can be easily converted into conditional probabilities for the Bayesian network.

Clare and King (2003) present the C4.5H method for HMLC which is a hierarchical variant of the well-established C4.5 method for MTP (J. R. Quinlan, 1993). Moreover, Clare (2003) extends the C4.5 decision tree method from Clare and King (2001) toward hierarchical setting in the context of functional genomics.

The work of Blockeel et al. (2006) presents a comparative analysis between predictive clustering trees for HMLC and single-label classification (HSC) approach using tree-shaped hierarchies. Vens et al. (2008) extend the PCTs for HMLC towards hierarchies represented in a DAG (directed acyclic graph) format and show that global decision trees, which create one tree for predicting all classes, are superior than the local decision trees, separate tree for each class.

The naive Bayes approach was adapted towards hierarchical multi-label classification in the work presented by Silla Jr. and Freitas (2009). Moreover, Cerri et al. (2012) discover HMLC rules using a search heuristic. In another research work, Cerri et al. (2014) represent a label hierarchy as a sequence of chained artificial neural networks (ANNs) where the outcome of the first ANN is used as an input of the next network in the sequence.

Alaydie et al. (2012) present HiBLADE (Hierarchical multi-label Boosting with LAbel DEpendency) boosting-based algorithm for HMC, where for each classifier, at each iteration, the training set is selected in accordance with the given label hierarchy.

In the domain of protein functions prediction, Barros et al. (2012) propose the HMC-PC algorithm which is based on the probabilistic clustering method adapted for solving HMLC tasks where the hierarchies can be both, DAG and tree-shaped. In order to generate the predictions of the class vectors, it uses the probabilities of cluster membership for each label. Their results show comparable results with the state-of-the-art methods for HMLC.

3.2.2 State-of-the-art methods for HMTR

To begin with, multilevel analysis refers broadly to the methodology of research and data structures that deal with nested data, i.e., including more than one type of unit. This is directly related to involving several levels of aggregation. Consider an example from educational research, where students from different schools are considered, and their performance (e.g., grades) is being predicted. Then, a separate regression model can be fitted within each school, and the model parameters from these schools can be modeled as depending on each school properties (such as the socioeconomic status of the school's neighborhood, whether the school is public or private, and so on). The student-level regression and the school-level regression here are the two levels of a multilevel model. The lowest level is the student-level and each student belonging to this level can be linked with an appropriate class, and then each class to an appropriate school and so on. With this, some sort of dependency levels (i.e., a hierarchy) are created. Moreover, in the higher levels in the multilevel model, regression parameters (hyper-parameters) can be fitted for the regression model. That is the reason why in most of the research, the term "multilevel analysis" is mostly used interchangeably with "hierarchical linear modeling", although strictly speaking, they are distinct. Another application of the hierarchical linear modeling approach can be found in Kuo et al. (2000), where a two-level hierarchical linear model with multiple outputs was employed to analyze information obtained from two different groups of informants (child and parent participants) in order to assess the demographic risk factors on children's exposure to violence (ETV) and how these effects vary by informants. The main advantage of multilevel modeling is spreading of a residual components through each level of a hierarchy, thus the overall variance is partitioned and moreover, the predictors are included at each level. Hence, with application of multi-target regression at each level, the model can deal with between-level relations in the hierarchy. The latter makes multilevel modelling superior than regression modeling with respect to the model performance (Gelman, 2006). An extensive review for multilevel modeling is given by de Leeuw and Meijer (2008) and Snijders (2011).

Next, online analytical processing (OLAP) is a method which allows to extract and analyze data from multiple sources at the same time. The data is multidimensional, hence the extracted information can be compared in different ways. For example, a book store might compare their book sales in September with sales in August, then compare those results with the sales from another location, which might be stored in a different database. The OLAP data is stored in multidimensional databases and all attributes are considered as a separate dimension. Considering the multi-dimensionality, the OLAP data is structured in a hierarchical form by using some of the OLAP tools: consolidation (roll-up), drill-down, and slicing and dicing (O'Brien & Marakas, 2010). This structuring and hierarchical representation enables a complex calculations and manipulation of the data (trend analysis, data modeling) (Agrawal et al., 1997). The natural relationships in the data by using the OLAP method are also researched by Nguyen et al. (2000) by using a partially ordered set of levels (dimension schema).

Predictive clustering trees (PCTs) for the HMTR task were proposed recently by Mileski et al. (2017). The original PCTs for MTR are extended to the HMTR task with defining the prototype function and variance function. All operations for aggregation can be used as prototype functions, but keeping in mind that with some of them (for example, *minimum* or *maximum*) after averaging, the hierarchical constraint (parent-child relation within the hierarchy) can be violated. For the variance function, the weighted Euclidean distance is used where the weights are defined such that they decrease exponentially with the depth of the node in the hierarchy.

3.3 Methods for Structuring the Output Space in MTP

In contrast to the relevant work exploring local and global models for solving the MTP task, there are not so many well-established methods that explore the structure of the output space and possible dependencies in the space of target attributes. In the following section, we will present some of the few attempts made in that direction.

Tsoumakas and Vlahavas (2008) propose a transformation-based ensemble method for random k-labelsets (RAkEL) for MLC by using existing algorithms for MLC. The RAkEL algorithm creates an ensemble by random sampling a small subset with k labels for each base model. The sampled subsets are structured as a label powerset and a multi-class classifier is then used. Differently, Tsoumakas et al. (2008) present a data-driven approach called HOMER (Hierarchy Of Multilabel classifiERs) for effective and computationally efficient multi-label learning. The idea behind this approach is to structure the large label space into a tree-shaped hierarchy of smaller MLC tasks learned on a smaller subset of labels.

Joly et al. (2014) propose a method for dimensionality reduction of the output space by random projections of it, mainly focused on the MLC task. The projections are made in such a way that they preserve distances in the projected space. The reduction of the variance function is made on the projected space, while the predictions are made directly in the original output space using a decoding procedure. Similarly, Joly (2017), proposes a gradient boosting method for MTR which automatically adapts the target correlations by random projection of the output space.

Madjarov et al. (2016) present a comprehensive study of different data-derived methods for structuring the label space in the form of hierarchies for MLC. Namely, they use the label co-occurrence matrix to obtain a hierarchy of labels by using several clustering algorithms such as: agglomerative clustering with single and complete linkage, balanced k-means and PCTs. Their results say that divisive clustering methods (balanced k-means and PCTs) perform the best.

Next, Szymanski et al. (2016) present a study which addressed the question whether data-driven methods on a graph consisting of label co-occurrences are significantly better than a randomly generated graph of labels for MLC. This method is actually a data-driven version of the RAkEL method (Tsoumakas & Vlahavas, 2008). Their results show that in general, the data-driven approach is superior to randomly created graphs of labels.

Duivesteijn et al. (2012) propose a method called LeGo for structuring the label space in MLC by finding the local patterns (LeGo puzzles) capturing the intrinsic properties in the label space. The local patterns are represented with binary features which are further used as an input in the ML classifier.

S. Wang et al. (2014) propose a method for structuring the label space by representing the label dependencies in the form of a Bayesian network, where the nodes of the network are the labels and the edges are conditional probabilities. The proposed method has shown high flexibility over the existing MLC methods in terms of handling label incompleteness.

Wu et al. (2016) propose an ML-TREE ensemble approach where the actual relevant labels are kept for a given instance during the model training process. In the learning process, for each instance with multiple labels, they transfer the hierarchical tree of relevant labels in order to exploit the intrinsic label dependencies.

Most recently, Zhen et al. (2018) present a deep learning approach for considering the intra-target dependencies. Namely, they propose a multi-layer multi-target regression (MMR) method where intra-target dependencies are explicitly encoded by using matrix elastic nets (MEN) to create the structure of the target space (structure matrix S), which enables learning of the target correlations by minimization of the $rank(S)$. Then, the kernel trick is used in order to solve the problem of non-linearity in the representation of the target dependencies.

Chapter 4

State-of-the-art Methods for Modeling Soil Functions

The literature survey emphasizes that earlier modeling activities of soil functions have been based on classical mathematical models, including *physical-based* models that represent underlying soil functions as dynamic systems: DAISY (Abrahamsen & Hansen, 2000), AGROSIM model family (Mirschel & Wenkel, 2007); *mechanistic* models: DNDC (Gilhespy et al., 2014), STICS (Brisson et al., 2003), CENTURY (Parton & Rasmussen, 1994), and DayCent (Parton et al., 1998); and *large-scale GIS* models: EPIC (Balkovič et al., 2013).

The modeling setup in the project LANDMARK (LANDMARK, 2019) was focused on two concepts such as qualitative multi-criteria decision modeling (MCDM) as a concept of human intelligence and machine learning (ML) as a concept of artificial intelligence (AI). The former is emphasized as an approach due to the ability to consider expert knowledge and deliver knowledge-based support for decision-making. The latter, as shown in the remainder of the section, is able to enrich modeling efforts and improve overall predictive and descriptive performance on modeled phenomena. However, a very little effort towards using of ML methods in fitting of the decision support models was made. Motivated by this, in this dissertation we are more focused on potential ML uptake to the expert-based models in order to facilitate and improve the decision making process.

The need of backing up decision-making with data-driven approaches is mainly initialized due to limited performance of knowledge-based models and vast potential of new discoveries and better understandings, available from data collected recently with advances of the technology (Debeljak et al., 2007; Kocev et al., 2010; Trajanov et al., 2018). However, harmonization of data coming from different sources may be a hurdle in acquiring expert knowledge due to experts' unavailability or opposite understandings and conflicting opinions (Shaw & Woodward, 1990). Data harmonization and warehousing is beyond the scope of this thesis, but once the data harmonization is performed, machine learning gives a great opportunity to facilitate improved understanding of soil functions by either modeling them individually or simultaneously.

In order to have a clear overview of improvements in our research outcomes, literature on state-of-the-art methods for modeling soil functions including both, knowledge-based MCDM with decision support systems (DSS) and ML, is extensively reviewed in the following subsections for each soil function, separately.

4.1 Primary Productivity

Primary productivity is one of the most important soil functions for the farming communities, since it represents the capacity of soil to produce nutritional (caloric) biomass such as fuel, fiber, food and feed. Based on the UN predictions, the global agro-production must increases by 60% to fulfill the requirements of the inevitable growth of the world population by the year 2050 (WWAP, 2015). Therefore, farmers strive to find a proper tool for assessing the primary productivity by using existing attributes that can easily be measured or estimated.

Motivated by this fact, computer and data scientists attempt to satisfy the farmers' requirements by either structuring the collected knowledge using the MCDM models or mining the existing data that experts/farmers have a vast amount of and creating data-driven models using ML algorithms.

Primary (biomass) productivity is driven by many factors, but with a collection of proxy-indicators, it can be modeled to an acceptable level of accuracy. As we described previously, proxy indicators belong to either soil (S), environmental (E) and management (M) -related indicators. Mueller et al. (2010) and Toth et al. (2013) present the basic E and M attributes for primary productivity. The land management attributes are land-use types, pest management and fertilization, and environmental factors are climate (temperature and precipitation), geographical data (slope degree, altitude, longitude, latitude), etc. However, there are many other soil- and crop-related attributes that affect primary productivity: crop rotation and stocking rate, as well as other physical (soil structure), chemical (micro, macro and other soil constituents) and biological activity (pH, soil organic matter, etc.) attributes.

4.1.1 DSS for modeling primary productivity

The most recent study by Thoumazeau et al. (2019) presents *Biofunctool*® framework for assessing the quality of soil, based on an decision modeling integrative technique that takes into account physical, chemical and biological activities of soils. The selection of the 12 indicators for determining three soil functions: carbon transformation, nutrient cycling and structure maintenance, was performed by field visits and experiments through a "top down" approach and expert decisions (Bockstaller & Girardin, 2003; Griffiths et al., 2016). Structure maintenance is one of the aggregated proxy-indicators of primary productivity. The results of this study consist of indicator redundancy analysis and indicator sensitivity to the land management represented with correlation matrices.

Another multi-criteria decision support tool for qualitative assessment of different crop protection systems of apple production by aggregation of sustainability attributes is presented by Mouron et al. (2012). The sustainability is modeled by DEX methodology (Bohanec & Rajkovič, 1990) integrated in the DEXi decision support tool (Bohanec, 2014). In order to increase the efficiency of the innovation process, Pelzer et al. (2012) developed the DEXiPM (DEXi Pest Management) tool for *ex-ante* assessment of the sustainability of arable cropping systems. It relies on 75 indicators that describe the underlying cropping system. DEXiPM is evaluated on data about winter crop- and maize-based cropping systems from a French region.

Craheix et al. (2016) propose the MASC ®(Multi-criteria Assessment of the Sustainability of Cropping systems) decision support system for assessing crop sustainability. It is evaluated on 31 cropping systems from six different regions in France belonging to a variety of pedoclimatic zones. The MASC model is developed with DEX methodology (Bohanec, 2014) and conceptualizes the crop sustainability by breaking it down into three dimensions that define the sustainability: economic, social and environmental dimension.

The hierarchical structure of the qualitative multi-criteria decision model consists of 65 aggregated attributes.

Bohanec et al. (2017) propose SIGMO – a DEX model for assessing genetically modified (GM) crops. SIGMO model is combined with data-driven support for GMO crop species produced in different countries worldwide. The decisions are based on supply chains for potential presence of authorized and unauthorized GM organisms (GMOs). This research was a part of the DECATHLON FP7 EU project (DECATHLON, 2016).

Qualitative multi-criteria decision models on primary productivity, developed within the project LANDMARK, consists of 25 input attributes including soil properties, cropping specifications, environmental conditions and management practices (Sandén et al., 2019). Primary productivity soil function is decomposed into smaller and easier sub-problems: crop, soil, management and environment proxy-indicators, with quantified effect to the outcome of 20, 22, 28 and 30%, respectively. The model was validated on 399 sites from two different pedoclimatic zones across France (Metzger et al., 2005), showing overall accuracy of 40%.

4.1.2 ML for modeling primary productivity

As part of the LANDMARK project, Trajanov et al. (2018) apply classification trees for modeling primary productivity using data collected in France. The performance has shown to be significantly better compared to the previously built DEX model (Sandén et al., 2019). Upon validation, the aforementioned DEX model has been updated accordingly and its accuracy improved to 77%.

The rest of the related work in this segment is mainly concentrated on modeling net primary productivity (NPP), i.e., seasonal or annual yield expressed as a quantity produced from a particular arable area, as a proxy-value to valuation of primary productivity soil function. The literature cluster around modeling performed on either structured data about environmental conditions and agricultural management practices, or unstructured data, e.g. multi- and hyper-spectral images and remote sensed data.

Marinković et al. (2009) apply M5 model trees (R. J. Quinlan, 1992; Y. Wang & Witten, 1997) for predicting annual yield of 3 different crops: soy bean, maize and sugar beet. Furthermore, they use best-first search (Witten & Frank, 2005) and genetic algorithm (GA) for attribute selection (Goldberg & Holland, 1988). The data were collected in a period from 1999 to 2008, including annual yield in the Serbian province of Vojvodina, provided by the Faculty of Agriculture in Novi Sad, Serbia, and comprised weather attributes (minimal, maximal, average monthly temperature and precipitation), hydrometeorological attributes: evapotranspiration (potential and real) and hydrophitothermic index. They learn 3 single model trees for each crop and show that the data-driven models are at least in compliance with the existing crop production models. With attribute selection using GA they improve the soybean model by 5%, considering the correlation coefficient.

A research presented by Arumugam (2017) is based on the application of various data mining techniques in the prediction of the maximum yield of paddy crops. The dataset was constructed from 200 different questionnaires distributed to various farmers cultivating paddy along the Thamirabarani river basin in India, and comprises the following descriptive variables: soil type, crop variety, seed quality, seed rate, season, fertilizer type, amount of used fertilizer, amount of used pesticides, rainfall, land preparation method, sowing procedure, crop rotation, natural manure, soil fertility and temperature. In the study, the following ML techniques for classification are used: J48 (classification tree) (J. R. Quinlan, 1993), random forest (Breiman, 2001), decision stump and RepTree by using WEKA software (Witten & Frank, 2005) and the random forest has proved to be the best method with a predictive accuracy of 97.5%.

Ali et al. (2016) present a study for the application of machine learning on Irish grasslands data from 2 farms: Moorepark and Grange taken by *in-situ* measurements. The data consist of images taken from the following satellite families: ALOS-2, Radarsat2, Sentinel, TerraSAR-X, TanDEM-X/L. The problem they address is the estimation of grassland productivity. They developed multiple linear regression (MLR) (Tabachnick & Fidell, 2001), artificial neural networks (ANNs) (Hopfield & Tank, 1985) and adaptive neuro-fuzzy interface system (ANFIS) (Jang, 1993) models and showed that the ANFIS model provides the best estimation of the grassland productivity compared to the MLR and ANNs. Next, Wolanin et al. (2019) propose a hybrid machine learning approach combining process-based modeling and soil-canopy energy balance radioactive transfer model (SCOPE) (van der Tol et al., 2009) on optical data from Sentinel-2 and Landsat 8 satellites in order to predict crop gross primary productivity (GPP). They show good estimation accuracy of their hybrid model although they do not use environmental data on climate conditions.

Moreover, Pantazi et al. (2016) use the machine learning on remote sensing high-resolution data about factors affecting the crop growth and yield in UK. They use counter-propagation artificial neural networks (CP-ANNs) (Fjodorova et al., 2010), XY-fused Networks (XY-Fs) and Supervised Kohonen Networks (SKNs) (Melssen et al., 2006) as learning techniques. Their results show that SKNs are the best performing method for predicting wheat yield in a $22ha$ field in Bedfordshire, UK.

Furthermore, Chlingaryan et al. (2018) presents a review of studies for using machine learning in the last 15 years for crop yield and nitrogen status estimation. They compare different ML approaches for modeling productivity phenomena using remotely sensed data from precision agriculture (PA).

Crane-Droesch (2018) propose a semi-parametric variant of deep neural networks (SNNs) for prediction of corn yield in US Midwest. The data is temporal (daily basis) from the period 1979-2016 and consists of the variables including GIS information about the sites (longitude, latitude and county), minimum and maximum air temperature and relative humidity, precipitation, incoming shortwave radiation (sunlight), and average wind speed. They show that their approach outperforms the classical ordinary least squares (OLS) regression (Goldberger, 1964) and non-parametric deep NNs (Deutsch & Journel, 1998).

Kung et al. (2016) present an ensemble neural network (ENN) approach for predicting agriculture yield in Taiwan. Data is provided by The Council of Agriculture in Taiwan and consists of estimations of planted areas and yield produced. The ENN method consists of generating different networks, each with a different number of hidden layers and neurons. Those who are not satisfying the desired accuracy are discarded and the prediction is calculated by averaging the predictions from the existing networks of the ensemble. ENN method improves the predictive performance over the classical back-propagation NN (Riedmiller & Braun, 1993) and multiple regression analysis (MRA) (Aron & Aron, 1999) by 12.4%.

Ying-xue et al. (2017) propose the support vector machine-based open crop model (SBOCM) for predicting the rice production in China. The data is obtained from the Chinese Academy of Science from different weather stations and consists of meteorological variables and the outputs of rice development and yield records. The SVMs was used to investigate over four evaluating objectives, optimal kernel function, penalty coefficients, hyper-parameters. The most limiting objective for model optimization are the penalty coefficient, followed by kernel function and hyper-parameters.

4.2 Water Regulation and Purification

Water regulation and purification is substantial for providing clean water for drinking that meets the criteria of environmental pollution set by the EU Water Framework Directive (Directive 2000/60/EC of the European Parliament and of the Council) (WFD, 2000). Water regulation and purification is defined as the capacity of soil to receive, store and conduct water for subsequent use while minimizing the effects of prolonged droughts, flooding and erosion (Schröder et al., 2016). Water purification is one of the most important soil processes for reducing the pollution of natural water bodies and minimizing pollution spread to plants, animals and humans. In agriculture, water transports nutrients and sediments, including nitrogen (N) and phosphorus (P), into receiving water bodies, through water runoff and leaching pathways, and therefore, causes pernicious effects on the aquatic ecosystem. Agriculture is one of the main sources of nitrate and phosphate pollution of water bodies (FAO, 2003; OECD, 2001).

The water regulation and purification can be divided into three pathways : (1) *water runoff* (horizontal transport of water over soil surface, i.e., the maximal threshold to which a soil produces overland flow); (2) *water storage* (i.e., the capacity of soil to receive and store a water for reusing) and (3) *water percolation* (i.e, the degree of soil drainage, i.e., the extent to which soil profile allows the water to drain through it). Water movement is the main pathway of nitrogen (N) loss to groundwater and phosphorus (P) runoff (Daly et al., 2018). The pathways 1 and 3 are very critical for increasing of the nutrient transport, therefore, in the purification process, the erosion and sediments loss must be considered.

4.2.1 DSS for modeling water regulation and purification

Pierce et al. (2016) describe the need of a decision support tool for efficient groundwater management and present an overview of decision support systems and processes for assessing the groundwater regulation.

Le Page et al. (2012) present an integrated DSS for assessing groundwater based on remote sensing in semi-arid aquifer in Morocco. The DSS consists of two tools, one tool for the estimation of Agricultural Water Demand (SAMIR) from satellite images and the other integrates water resources planning (WEAP) including a groundwater model (MODFLOW). After the validation of estimates from the DSS model, satisfactory results are obtained.

Moura et al. (2011) propose a decision support tool for regulation of storm-water infiltration systems that reduce water flows in downstream sewers, minimize the overflows in surface waters and make it possible to recharge groundwater. Next, Hamouda et al. (2009) develop a decision support system for selecting the water treatment process and present an overview of the existing decision support methods which incorporate existing knowledge about water treatment systems.

Letcher (2005) present a decision support platform called Water Allocation Decision Support System (WAdss) developed for the management of trade-offs between stakeholders and policymakers with interests and concerns. WAdss assesses water storage across three water systems (unregulated, regulated and groundwater) and is validated on the data from two NSW catchments (Gwydir and Namoi). The WAdss tool is built on expert knowledge acquired from a large number of stakeholders.

Recio et al. (2005) propose a decision support tool for managing water resources in the agricultural environment. Objective of the DSS is to assist Eastern Mancha Central Irrigation Board representatives to evaluate water use policies in an efficient way and maintain the sustainability of natural resources in combination with regional economic development (primarily based on irrigation farming). The DSS incorporates two models, the econometric and the hydro-geological model. The former estimates the evolution of the regional crops map, the crop yields and the associated regional gross prices. The latter simulates the River Júcar basin and its associated aquifer model.

Navarro-Hellín et al. (2016) develop a soil irrigation decision support system (SIDSS) for managing irrigation systems in agriculture. Objective of SIDSS is to generate an irrigation plan and to optimize (i.e., minimize) water usage in an accurate way, by using soil and climate variables. Similarly, Giusti and Marsili-Libelli (2015) describe Fuzzy Decision Support System (FDSS) for smart irrigation planning and improving irrigation performances by reducing unnecessary water usages.

LANDMARK team of experts for modeling the soil function water purification and regulation, decomposes the function into three sub-phenomena: water storage, water runoff and water percolation – each being a proxy-indicator (Wall et al., 2019). Down the hierarchy, proxy-indicators are estimated with 48 attributes from soil (S), environment (E) and management (A) groups of attributes, mainly composed by an extensive survey of farmers' and policy-makers' knowledge (Bampa et al., 2019). The list comprises attributes on precipitation, evapotranspiration, soil organic matter, texture and dispersion, N and P surpluses, soil moisture deficit, water holding capacity, plant rooting depth, etc.

4.2.2 ML for modeling water regulation and purification

Related work on modeling water purification and regulation is mainly focused on modeling water loss with its pathways, sediment movements and water quality from biological and physical processes.

Dou and Yang (2018) present different machine learning approaches for modeling evapotranspiration (ET) in four main ecosystems on different spatial scales. They used both extreme learning machine (ELM) (Huang et al., 2012) and adaptive neuro-fuzzy inference system (ANFIS) (Jang, 1993) and the hybrids of ELM and ANFIS methods to estimate the daily ET. Three EML hybrids are considered based on the three activation functions. The hybrids of ANFIS algorithms are generated based on consequent parameter optimization with least squares and premise parameters with gradient descent. They achieved a significant difference regarding the modeling performance among four major ecosystem types.

Liu et al. (2018) present a study for the estimation of water quality dynamics, particularly by the association between spatial variability and catchment characteristics. They evaluated machine learning techniques on a dataset of nine water quality constituents collected from 32 monitoring sites for the period 2006 to 2016, across the Great Barrier Reef catchments (Queensland, Australia). Using dimensionality reduction with Principal Component Analysis (PCA) (Jolliffe, 2002) they identified four groups of sites with a similar spatial pattern, which can determine the key catchment characteristics helping to assess the water quality.

Soil erosion is a process that affects the environment negatively and the ability to model it increases its understanding. Therefore, Teng et al. (2019) propose the use of machine learning for modeling soil erosion in China. They use the Random forest algorithm for estimating soil erosion. The input (descriptive) data are satellite images, from which soil erodibility and a set of environment factors, were extracted. The results show an average erosion rate in China of $1.44t/ha/yr$. Moreover, Castrillo and García (2020) propose a study for the estimation of high frequency nutrient concentrations from water quality surrogates using the Random forests method for regression. The data were mostly collected by surrogate measures. They compare their results with the results obtained by linear regression models and obtain around 60% of improvements in the performance according to the RMSE measure.

In their study, Ahmed et al. (2019) predict water quality parameters consisting of pH, ammoniacal nitrogen (AN) and suspended solid (SS). They collected pretty noisy sensor data from monitoring stations in the Johor River Basin in Malaysia. They use different machine learning techniques including: Adaptive Neuro-Fuzzy Inference System (ANFIS) (Jang, 1993), Radial Basis Function Neural Networks (RBF-ANN) (Faris et al., 2017), and Multi-Layer Perceptron Neural Networks (MLP-ANN) (Hastie et al., 2009). Due to the presence of noise in the data, the predictions were unsatisfactory, however, they have applied an augmented wavelet de-noising variant of ANFIS (WDT-ANFIS) in order to obtain better predictions. The model was validated with field data from 2009-2010 and achieved a performance with a coefficient of determination $R^2 \geq 0.9$.

Furthermore, D. T. Bui et al. (2020) present a study about application of various well-known standalone machine learning techniques and novel hybrid methods for predicting water quality indices (WQI). They compiled six years (2012 to 2018) of monthly data from two water quality monitoring stations within the Talar catchment (Iran). The methods they used include four standalone methods: random forest (RF) (Breiman, 2001), M5P (R. J. Quinlan, 1992; Y. Wang & Witten, 1997), random tree (RT), and reduced error pruning tree (REPTree) (Witten & Frank, 2005), along with 12 novel hybrid methods as combinations of the standalone methods with bagging (BA) (Breiman, 1996), CV parameter selection (CVPS) (Kohavi, 1995) and randomizable filtered classification (RFC) (Witten & Frank, 2005). Using 10-fold CV and 70/30% division of train/test dataset, they obtained the best predictive performance with the Hybrid BA-RT method ($R^2 = 0.941, RMSE = 2.71, MAE = 1.87, NSE = 0.941$).

4.3 Carbon Storage and Climate Regulation

The carbon sequestration and climate regulation plays an important role in regulating atmospheric greenhouse gases (GHG) such as carbon dioxide (CO_2) and nitrous oxide (N_2O). Regulation is achieved by either improved sequestration or limited emission of particular gasses. Improvement of soil sequestration service relies on the optimization of soil property to absorb CO_2 through pathways naturally designed to convert it to soil organic carbon. Such improvement can be achieved by intervening on soil physical properties or adaptation of soil cover, i.e., forestation.

Unlike sequestration, climate regulation soil ecosystem service reacts in the opposite direction and aims at limiting emission of GHG to the atmosphere. Changes are mainly dictated by on-surface land management for agricultural purposes. The transformation of soil organic carbon (SOC) to CO_2 in soil ecosystems mainly occurs because of the conversion of nature (for example forests) to arable land. On the other hand, N_2O emission occurs because of the microbial transformation of applied fertilizer, which contains reactive nitrogen (N), to the agricultural land. N_2O emissions can occur indirectly as well, through ammonia (NH_4) and nitrate (NO_3^-), when reactive N is applied to other ecosystems.

Related work on processes that correspond to carbon sequestration and climate regulation is mainly framed around the management of SOC storage and its improvement by modeling effects of physical, chemical and biological soil properties on sequestration in different terrestrial ecosystems.

4.3.1 DSS for modeling carbon storage and climate regulation

Biofunctool® framework (Thoumazeau et al., 2019) is a decision support integrative tool for assessing soil quality. This tool includes physical, chemical and biological activities of soils and consists of three assessing protocols among which carbon transformation can be found. The indicators used to assess carbon transformation were: cast density (Cast), bait lamina (Lamina), Permanganate OXidizable Carbon (POXC), basal soil respiration (SituResp) and two litter indices (Fragment and Skeleton).

LANDMARK modeling activities on carbon sequestration and climate regulation result in a DEX model that decomposes the soil function index into three sub-processes (proxy-indicators): carbon sequestration, reduction of N_2O and CH_4 emissions (Van de Broek et al., 2019). The proxy indicators are estimated from the list of basic management and environmental attributes. Validation of this model is not completed due to lack of data related to this soil function in different climate zones. However, partial validation is performed based on evaluation of the proxy-indicators with data from European long-term experiments. The partial validation shows that the model is generally able to correctly assess the effect of different management practices on carbon sequestration and N_2O emissions.

4.3.2 ML for modeling carbon storage and climate regulation

Xu et al. (2018) use the statistical learning method generalized linear model (GLM) (Nelder & Wedderburn, 1972) for exploring the variance of carbon storage across different terrestrial ecosystems in China. The spatial patterns and the main drivers of C storage remain unclear due to the lack of data. However, GLM reveals some insights that climate, soil texture and nutrients are the main drivers of regulating spatial patterns of carbon storage. Similarly, Gardi et al. (2016) use GLM in the statistical analysis of their LUKAS data about determining the High Nature Value Farmland (HNVF), which supply the process of carbon sequestration. Considering soil organic carbon as a proxy for carbon storage, they compare HNVFs with soils that endure more conventional land management (nHNVFs) and study the consequences of diverse land uses and geographic regions as additional descriptive attributes. Their results show that soil organic carbon is higher in HNVF than in nHNVF at the European level and the difference is mainly affected by the geographic region and land use type.

E. Bui et al. (2009) presented a predictive modeling technique for predicting soil organic carbon (SOC) in agricultural regions in Australia by using decision trees. As an input, they use the national database of soil data, soil maps, digital surfaces of climate, elevation, and terrain variables, Landsat multi-spectral scanner data, lithology and land use. They found that despite the temperature, soil moisture level seemed to be the most important driver of SOC at the continental scale. Moreover, Mahmoudzadeh et al. (2020) present a study for the estimation of SOC for accurate monitoring of carbon sequestration. The data are collected from western Iran and comprise 865 soil samples and 101 input (descriptive) variables. They use five ML algorithms include: random forests (RF) (Breiman, 2001), Extreme Gradient Boosting (XGBoost) (T. Chen & Guestrin, 2016), Cubist (CU) (Kuhn et al., 2014), k-Nearest Neighbor (kNN) (Altman, 1992) and Support Vector Machines (SVMs) (Cortes & Vapnik, 1995). The best method for predicting the spatial distribution of SOC was the RF method ($RMSE = 0.35\%$ and $R^2 = 0.60$). In addition, the most important descriptive variables for predicting SOC are: rainfall, valley depth, terrain surface texture, air temperature, channel network base level and terrain vector roughness.

Schillaci et al. (2017) present a study for modeling topsoil SOC concentration from remote sensing data collected in a period 1998-2009 from cultivated areas in Sicily (Italy). The machine learning method they use is boosted regression trees (BRT) (Elith et al., 2008). Results show content performance, with coefficient of determination (R^2) between 0.61 and 0.69. Driving parameters of the SOC concentration have shown to be soil texture, land use, rainfall and topographic indices related to erosion and deposition.

Furthermore, Ottoy et al. (2017) applied four machine learning techniques for topsoil and subsoil SOC modeling on data from nature reserves in Flanders (Belgium). They apply multiple linear regression, boosted regression trees (Elith et al., 2008), artificial NNs (Hopfield & Tank, 1985) and least-squares SVMs (Suykens & Vandewalle, 1999). The best performing method is boosted regression trees, which has lowest cross-validation error and provides insight into the relative importance of predictors.

Taki et al. (2018) consider the phenomena of GHG emission as a non-linear process and perform a study on modeling intrinsic variables of GHG, which directly affect the carbon sequestration. The study comprises different non-linear ML algorithms for multi-target prediction such as: artificial neural networks (ANNs) (Hopfield & Tank, 1985), multi-layer perceptron (MLP) (Hastie et al., 2009), radial basis function (RBF) algorithm (Faris et al., 2017) and support vector machine (SVM) (Cortes & Vapnik, 1995). Modeling is performed on the remote sensing data from Shahreza city, Isfahan province, Iran, represented with the following sets of attributes: air, soil and plant temperatures (Ta, Ts, Tp) and energy exchange in a polyethylene greenhouse. RBF has the lowest predictive error compared to the rest of built models.

4.4 Provision of Functional and Intrinsic Biodiversity

Functional and intrinsic soil biodiversity enriches the interaction of sediments and minerals in soil, leading to overall improved soil processes and crop growth support. Consequently, it provides the society with a rich biodiversity source and contributes to a habitat for above-ground organisms. Processes that affect the development and maintenance of soil biodiversity are broadly classified as physical and chemical, including soil structure and hydrology, and nutritional content, respectively.

Gardi et al. (2009) present an overview about the main threats for the soil biodiversity and describe how the biodiversity indicators are developed and estimated.

4.4.1 DSS for modeling provision of functional and intrinsic biodiversity

Knowledge-based methodology prevails in decision modeling of soil of provision of functional and intrinsic biodiversity with a focus on assessment of abundance and richness of intrinsic communities across different spatial scales. Extensive work has been performed in the Netherlands and France along with recent endeavours within the LANDMARK project (Mulder et al., 2005; Rutgers et al., 2009).

Another research approach is focusing on negative effects on soil biota development and potential threats to soil biodiversity on different spatial scales. Such approach has recently been applied on a European scale (Orgiazzi et al., 2016).

Rutgers et al. (2009) and Mulder et al. (2005) perform soil biodiversity data monitoring within Dutch Soil Monitoring Network (NSMN). During the monitoring program in the period 1999-2003, biological and chemical soil attributes, as well as land use and management practices, were measured and analyzed. In total, they selected and visited 137 dairy farms because of availability of the data for management practices for those sites. Next, eight experts in soil biodiversity in the Netherlands were selected independently to fulfill a questionnaires and assess the main indicators for determining the soil biodiversity function. At the end, their ranks for the soil biodiversity indicators were included in the final decision support (i.e., expert assessment) tool in order to assess the final aggregated soil biodiversity estimate (Rutgers et al., 2009).

Similarly, in France, Cluzeau et al. (2012) performed a Soil Biodiversity Monitoring Network (RMQS-BioDiv), which is part of the French Soil Monitoring Network (RMQS), and measured and analyzed data from 2200 data sites from French Metropolitan Areas in Brittany (West France). The RMQS-BioDiv data was linked to the climate data averaged from the years 1990-2016. At the end, merged together with measured biological and other attributes, the final data consisted of 52 sites (29 grasslands, 23 croplands) with fully available values for the input attributes. Finally, the evaluation of biodiversity was made by expert judgment on those 52 sites using an a priori approach on biological and management attributes.

As part of LANDMARK modeling activities, van Leeuwen et al. (2019) developed a DEX model on the valuation of biodiversity soil function, by decomposition of the problem into a hierarchical model across four proxy-indicators (structure, hydrology, biology and nutrients). Proxy-indicators are decomposed down to 37 attributes that belong to a group of soil, management and environment ($S \times E \times M$) attributes (Schulte et al., 2014; Turbé et al., 2010; Vogel et al., 2018). The decision rules in the integration tables were fulfilled by the experts. In addition, van Leeuwen et al. (2019) performed a comparative analysis of assessments with aforementioned BioDiv-RMQS and NSMN expert models. Although the strategies for developing these three expert assessment approaches were different, the results show that all of them provided very correlated estimates for the valuation of provision of functional and intrinsic soil biodiversity.

Another knowledge-based assessment of the soil biodiversity function was proposed by Orgiazzi et al. (2016). They have ranked 13 potential threats to the soil biodiversity by surveying experts from around Europe. Assessment was based on quantification of three biodiversity components: soil microorganisms, fauna and biological function. Results imply that arable soils are exposed to threats, at most. However, despite the limitation of the proposed knowledge-based ranking, it is insightful for further monitoring and protection of soil biota.

4.4.2 ML for modeling provision of functional and intrinsic biodiversity

Debeljak et al. (2007) present a study for assessing effects of management practices to soil microorganisms, such as spring-tails and earthworms. The research is conducted on bi-annual data from Bt maize fields in Foulum (Denmark). They used regression trees (M5') (R. J. Quinlan, 1992; Y. Wang & Witten, 1997) for building predictive models. As input data they considered farming practices, soil parameters, the biological structure of soil communities, and the type and age of the crop at the time of sampling. Models were built to predict the abundance of different functional groups of species from both types, earthworms and spring-tails. They reported regression models for anecic worms and hemi-epiedaphic spring-tail with a performance of $R^2 = 0.83$ and $R^2 = 0.59$, respectively. All of the learned models did not find effects of the Bt maize crop on the specific soil microorganisms.

Another prominent example for predicting the abundance of different species occupying the same habitat is presented by Demšar et al. (2006). The data used comprise descriptors on agricultural events and soil biological parameters in order to estimate the effects of management practices, especially tillage on the abundance of spring-tails and mites and their biodiversity. They use regression and model trees for both, single- and multi-target regression, and obtain tree models with a promising predictive performance that can be used by decision makers.

Kocev et al. (2010) proposed a study for assessing the influence of the environmental conditions on the Lake Prespa diatom community. The data contain chemical and physical properties of the environment and the relative abundance of 116 different diatoms. They used different datasets, one with the whole 116 diatoms and another with the top 10 most abundant diatom taxa. As a learning algorithm, they used predictive clustering trees for single- and multi-target regression (Blockeel et al., 1998). The obtained tree models are consistent and extend the existing expert knowledge.

A more extensive study was proposed by Levatić, Kocev, Debeljak, et al. (2015) using the same data sets from the study of Demšar et al. (2006) plus two different data sets, one about organisms living in Slovenian rivers and another about the vegetation found in the State of Victoria, Australia. They used the classical methods for single- and multi-target prediction described before, however, this study also proposes the use of an advanced machine learning concept, hierarchical multi-label classification, because the species can be represented by an existing taxonomic hierarchical representation (Kocev et al., 2013; Silla & Freitas, 2011). Learning hierarchical models, additional information about the diatom's parent-child relation is considered. The additional information about the parent-child relations between the diatoms has contributed to improvement of the model performance of hierarchical over the classical multi-target prediction task.

Additionally, Guo et al. (2020) proposed a study about learning predictive models using the random forest algorithm (Breiman, 2001) for estimating the spatial distribution of the soil arthropods on a large scale. The data has been collected from cultivated land in Changtu County, Northeast China, and consists of the existing and freely available environmental variables. The resulting predictions calculated by random forest modeling were with the lower performance ($R^2 = 0.53$) due to data unavailability.

Very recently, Djerdj et al. (2020) proposed the use of deep learning algorithms for predicting the earthworm behavior, i.e. the activities of soil-dwelling organisms. The input data consisting image sequences and deep convolution neural networks (Krizhevsky et al., 2017) were used for learning algorithms. The model was validated by comparison with results of the standard avoidance test, using H_3BO_3, a standard pollutant.

4.5 Provision and Cycling of Nutrients

Provision and cycling of nutrients is the last soil ecosystem service conceptualized as a soil function for extensive management of soil performance in arable ecosystems. It reflects a capacity of soil to receive nutrients in the form of by-products, to provide nutrients from intrinsic resources or to support the acquisition of nutrients from air or water, and to effectively carry over these nutrients to harvested crops (Schröder et al., 2016). The cycling of nutrients is an important natural process in protecting the environment and organic nature, as well as enhancing the production of nutritional (caloric) commodities. Lack of nutrient circulation in soil requires manual application of 'new' nutrient that will compensate the deficit. Introduced nutrients are sustained by finite sources such as mined phosphorus (P) rock, potassium (K), and mineral nitrogen (N). Therefore, in modeling the nutrient cycling soil function, the dynamic and intrinsic properties of the soil need to be taken into account.

4.5.1 DSS for modeling provision and cycling of nutrients

Aarts et al. (2015) presented the annual nutrient cycling assessment (ANCA) decision support tool for quantifying the main performance of nutrient cycling indicators, including: excretion, use efficiency of feed by the herd, ammonia loss, crop yields, use efficiency of fertilizers by crops, soil and farm surplus, nitrate leaching, losses of GHG and use efficiency of farm, as a whole. The tool is developed for the purpose of Dutch dairy farms. The ANCA model is used as a tool for the calculation of nitrogen or phosphorus excretion. If the assessed value is below the national standard, authorities accept the farm-specific value, allowing the farmer to reduce the amount of manure previously expected to export.

Turunen et al. (2018) present a Multi-Attribute Value Theory (MAVT)-based decision support tool (DST) for facilitating sludge treatment decisions by assigning preference scores to each sludge treatment. The DST was validated with data from two municipal wastewater treatment plants (WWTP) in Finland. In the first case study, the tool output a preference score of 0.629 for preferred sludge pyrolysis. The decision support tool, in this study, proved to be adaptable to decision makers and improved transparency, understandability and comprehensibility of decision-making processes.

Djodjic et al. (2002) proposed a decision support system for phosphorus management at a watershed level. This DSS could be used to recommend the most optimal and proper management practices as well as to identify critical source areas and test the most proper best management practices. The DSS was developed by using the nutrient attributes including: Maryland Phosphorus Index (PI), diagnosis expert system (ES), prescription ES, and a nonpoint-source pollution model, Ground Water Loading Effects of Agricultural Management Systems (GLEAMS). The proposed decision support tool was applied to agricultural watersheds in southern Sweden.

Sulaeman et al. (2012) developed a Phosphorus and Potassium Decision Support System (PPDSS) tool used for determining the fertilizer requirement for specific crops based on soil testing. As input, PKDSS uses 14 soil properties divided into two protocols: first protocol to select the desired recommendation and second protocol to assess the correction factor. There soil properties are provided by a legacy soil database stored at the Indonesian Center for Agricultural Land Resource Research and Development (ICALRD) and can be used for the creation of quantitative soil property maps by using digital soil mapping techniques. The created maps can then be used from PKDSS for making a recommendations for fertilizer input and can serve to the agriculture policy makers in Indonesia.

Recently, within the project LANDMARK, the soil function on provision and cycling of nutrients is decomposed down to four proxy-indicators that valuate the cycling of nutrients: (1) *accommodation value* (AV), capacity of soil to receive products that contain nutrients (for example, manure); (2) *nutrient fertilizer replacement value* (NFRV), amount of nutrient availability from those products called manufactured fertilizers; (3) *apparent nutrient recovery* (ANR), the amount of nutrient taken up form the crops, and (4) *harvest index* (HI), the amount of nutrients taken up by crops, exported in harvests for consumption or upstream processing. These proxy-indicators are derived from a set of 54 basic soil (S), environment (E) and management (M) attributes (Schröder et al., 2016).

4.5.2 ML for modeling provision and cycling of nutrients

Suchithra and Pai (2020) defined the problem of soil nutrient classification to a multi-label classification task and applied a fast classification learning technique called Extreme Learning Machine (ELM) (Huang et al., 2012) with different activation functions such as: Gaussian radial basis, sine-squared, hyperbolic tangent, triangular basis and hard limit. The goal is to optimize fertilizer inputs and improve soil and environmental quality. Input data consist of soil features like village-wise soil fertility indices of Available Phosphorus (P), Available Potassium (K), Organic Carbon (OC) and Boron (B), and pH. Best accuracy of more than 80% is achieved using ELM with the Gaussian radial basis activation function.

Moreover, Ransom et al. (2019) propose a study for corn nitrogen recommendation using soil (S) and environmental (E) attributes as an input. They use eight well-known machine learning algorithms: stepwise (Efroymson, 1960), ridge regression (Hoerl & Kennard, 1970), least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996), elastic net regression (Zou & Hastie, 2005), principal component regression (PCR) (Jolliffe, 1982), partial least squares regression (PLSR) (Wold, 1997), decision trees (Breiman et al., 1984) and random forest (Breiman, 2001) on data collected from 49 sites in the U.S. Midwest. The best algorithm for corn N recommendation, in regard to predictive performance, is random forest (R^2 between 0.72 and 0.84). However, decision trees have been reported to require a minimal set of input attributes for generating the prediction.

Hosseinzadeh et al. (2020) applied artificial neural network (ANN) and multiple linear regression (MLR) models in predicting nutrient recovery from solid waste under different vermicompost treatments. The dataset contains 7 biological and chemical indices and as predictors, total nitrogen (TN) and total phosphorus (TP) recovery. The results show that the best prediction is obtained by ANN models (Hopfield & Tank, 1985), i.e., TN and TP with $R^2 = 0.9983$ and $R^2 = 0.9991$ respectively, compared to MLR models (Tabachnick & Fidell, 2001) with $R^2 = 0.834$ and $R^2 = 0.729$.

Finally, Dong et al. (2020) presented a new technique for precision fertilization of maize by applying a combination of wavelet analysis (Akansu & Haddad, 1992) and back propagation neural network (BPNN) (Riedmiller & Braun, 1993) and traditional SVMs (Cortes & Vapnik, 1995) and Random forests (Breiman, 2001). The data used for modeling came from the published "3414" experiments. The results showed that the best performance was achieved using the model that combines wavelet analysis with the BP neural network. Moreover, Wavelet-BPNN model has important practical significance because it provides the most optimal recommendation for precision fertilization considering the increase of maize production, during reduction of production cost and agricultural pollution.

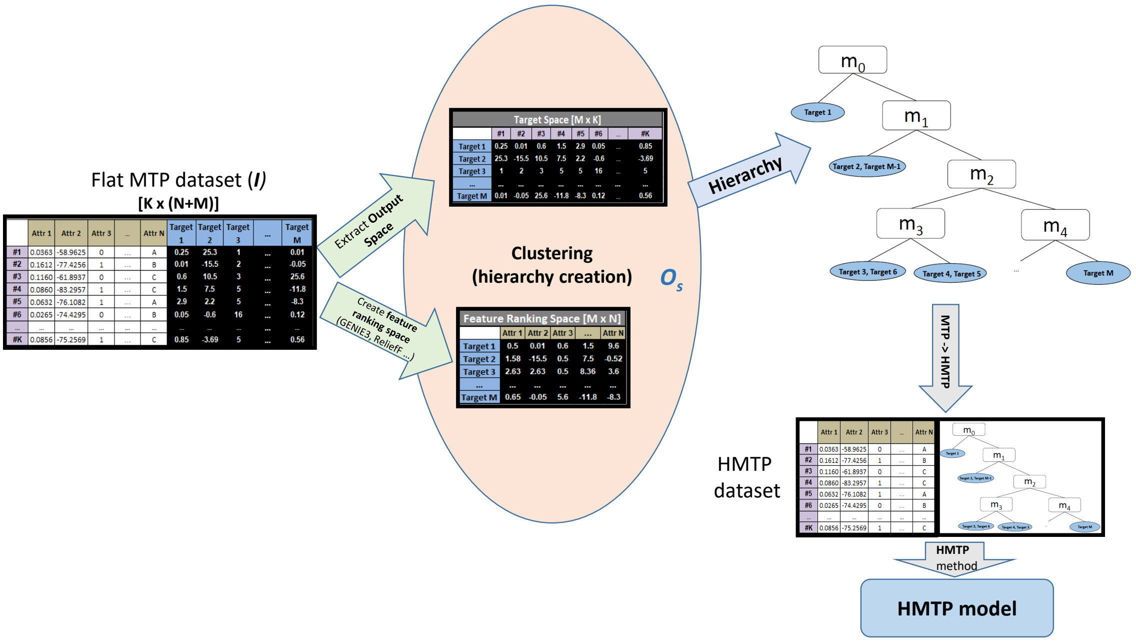
Chapter 5

Structuring the Output Space in Multi-Target Prediction

The main advantage of classical algorithms for multi-target prediction (MTP) is that they can learn one model for predicting multiple output variables simultaneously, instead of learning many single-target models, one for each target attribute. However, those classical MTP models do not consider the interconnections among the target attributes of the output space.

Motivated by this limitation of the classical MTP paradigm, in this chapter, we present an adaptation of classical MTP algorithms, by using different representations for data-driven structuring of the original output space, discovering target relationships in the form of a hierarchy, as shown in Figure 5.1. We describe the methodology by giving the pseudocode of the algorithm for structuring the output space in MTP, cf. Algorithm 5.1. The input to the algorithm is the original (i.e., flat) MTP dataset I. Next, we define the representation of the output space O_s (targets) for structuring: (1) *feature ranking representation*, consisting of feature importance scores for each target/label; (2) *target space representation*, consisting of the original values for the output/target variables. Once we obtain the representation of the output space O_s, we cluster this representation of the targets with an arbitrary hierarchical clustering algorithm and obtain the hierarchy (i.e., structure) of target attributes. The leaves in the hierarchy contain the original output/target variables. Using the obtained *hierarchy*, we transform the original MTP dataset to its hierarchical variant (HMTP) and by using any machine learning method for HMTP, we will create $Model_{HMTP}$. The remaining part of the pipeline consists of calculation of the predictions related to the targets/labels that are in the hierarchy leaves. Finally, using those predictions, we evaluate the predictive performance.

The remainder of the chapter contains our work on structuring the output space, both for the case of multi-label classification and multi-target regression. First, we present our algorithm for structuring the label space in MLC, using the representation that consists of the feature importance scores for each label. Next, we present our extensive study for structuring the output space for the MTR task where we consider both the target representation consisting of the original values for the target variables and the representation of the targets consisting of the feature importance scores for each target. In both studies, the results show improvements in the predictive performance if the data-derived structure on the output space is used, especially in the case of large output spaces.



	Attr 1	Attr 2	Attr 3	...	Attr N	Target 1	Target 2	Target 3	...	Target M
#1	0.0363	-58.9625	0	...	A	0.25	25.3	1	...	0.01
#2	0.1612	-77.4256	1	...	B	0.01	-15.5	2	...	-0.05
#3	0.1160	-61.8937	0	...	C	0.6	10.5	3	...	25.6
#4	0.0860	-83.2957	1	...	C	1.5	7.5	5	...	-11.8
#5	0.0632	-76.1082	1	...	A	2.9	2.2	5	...	-8.3
#6	0.0265	-74.4295	0	...	B	0.05	-0.6	16	...	0.12
...	...	...	...	...	...	...	...	...	...	...
#K	0.0856	-75.2569	1	...	C	0.85	-3.69	5	...	0.56

Target Space [M x K]

	#1	#2	#3	#4	#5	#6	...	#K
Target 1	0.25	0.01	0.6	1.5	2.9	0.05	...	0.85
Target 2	25.3	-15.5	10.5	7.5	2.2	-0.6	...	-3.69
Target 3	1	2	3	5	5	16	...	5
...	...	...	...	...	...	...	...	...
Target M	0.01	-0.05	25.6	-11.8	-8.3	0.12	...	0.56

Feature Ranking Space [M x N]

	Attr 1	Attr 2	Attr 3	...	Attr N
Target 1	0.5	0.01	0.6	1.5	9.6
Target 2	1.58	-15.5	0.5	7.5	-0.52
Target 3	2.63	2.63	0.5	8.36	3.6
...	...	...	...	...	...
Target M	0.65	-0.05	5.6	-11.8	-8.3

	Attr 1	Attr 2	Attr 3	...	Attr N
#1	0.0363	-58.9625	0	...	A
#2	0.1612	-77.4256	1	...	B
#3	0.1160	-61.8937	0	...	C
#4	0.0860	-83.2957	1	...	C
#5	0.0632	-76.1082	1	...	A
#6	0.0265	-74.4295	0	...	B
...	...	...	...	...	...
#K	0.0856	-75.2569	1	...	C

Figure 5.1: An illustration of the proposed framework for structuring the output space in MTP. We consider two target representations for clustering: the representation that uses original output space, i.e., the values of each target/label for each example and the feature ranking representation. The importance scores for each feature with respect to a given target/label are used as a representation of the targets. We thus transform the original MTP tasks into a HMTP task, for which we build a HMTP model.

Algorithm 5.1: The algorithm for structuring the output space in multi-target prediction (MTP).

```
Data: $I$ - MTP dataset
Result: $Model_{HMTP}$
if case: feature ranking representation then
    FimpPath = CreateFimp($I$);
    $O_s$ = CreateFeatureRankingRep(FimpPath);
end
if case: target space representation then
    $O_s$ = ExtractOutputSpaceRep($I$);
end
hierarchy = Clustering($O_s$);
$I_H$ = TransformMTP2HMTP($I$, hierarchy);
$Model_{HMTP}$ = HMTPMethod($I_H$);
```

5.1 Structuring the Label Space in Multi-Label Classification Using a Feature Ranking Representation of the Labels

In this chapter, we present the adaptation of the classical task of MLC by considering additional information provided by a hierarchy on the label space. Namely, we transform the original multi-label classification (MLC) task to its hierarchical (HMLC) variant, by structuring the label space, we use a representation consisting of the feature importance scores per label. We perform the structuring by creating a data-derived hierarchy of labels and using this hierarchy in the HMLC task. The data-derived hierarchy is generated using existing and commonly used hierarchical clustering techniques such as agglomerative clustering (single and complete linkage), balanced k-means and clustering with PCTs. The evaluation has been performed on 8 diverse benchmark datasets using 13 well-known MLC evaluation metrics.

Related to this topic, Madjarov et al. (2016) have presented an extensive study of different data-driven methods for generating label hierarchies for MLC by using the label co-occurrence space. More precisely, the hierarchies were constructed using the four clustering algorithms that we also used in our study. Next, Szymanski et al. (2016) have addressed the question whether the data-driven approach of using the label co-occurrence graph is better than taking random partitions of the label space for MLC as performed by RAkELd. Their results have shown that in almost all cases, data-driven partitioning outperforms the baseline RAkELd in all evaluation measures but Hamming loss.

In this section, we present an extension of the study performed by Madjarov et al. (2016), where the data-derived hierarchies are obtained by clustering the label space using a label representation consisting of label co-occurrences. We compare the results obtained by our data-derived hierarchies generated by structuring the label space using the label representation consisting of feature importance scores for each label with the results obtained using the data-derived hierarchies generated by clustering the representation consisting of label co-occurrences. Improvements are achieved if we use the data-derived hierarchies obtained by clustering the label space representation consisting of feature importance scores per label. In general, the use of the data-derived hierarchies improves the predictive performance over the flat MLC task. Furthermore, the divisive methods (balanced k-means and PCTs) generate the most accurate hierarchies, i.e., are the best clustering algorithms for structuring the label space.

The paper included in this section is:

- NIKOLOSKI, Stevanche, KOCEV, Dragi, DŽEROSKI, Sašo. Structuring the output space in multi-label classification by using feature ranking. In: APPICE, Annalisa (Ed.). New frontiers in mining complex patterns: 6th International Workshop, NFMCP 2017, in conjunction with ECML-PKDD 2017, Skopje, Macedonia, September 18-22, 2017: revised selected papers, (Lecture notes in computer science, ISSN 0302-9743, Lecture notes in artificial intelligence, LNCS 10785). Cham: Springer. 2018, LNCS 10785, pp. 151-166, doi:"10.1007/978-3-319-78680-3_11.

The contributions of Stevanche Nikoloski to this paper are as follows. SN contributed to the adaptation of the existing computer code for structuring the output space in MLC, especially developing computer code for designing a label representation, consisting of the label's feature importance scores. He also participated in designing the experiments, carried out the experiments and analyzed their results. He wrote the paper draft and revised it according to the relevant comments from the co-authors and reviewers.

Structuring the Output Space in Multi-label Classification by Using Feature Ranking

Stevanche Nikoloski[2,3](✉), Dragi Kocev[1,2], and Sašo Džeroski[1,2]

[1] Department of Knowledge Technologies, Jožef Stefan Institute,
Ljubljana, Slovenia
{dragi.kocev,saso.dzeroski}@ijs.si
[2] Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
stevanche.nikoloski@ijs.si
[3] Teagasc, Environment Soils and Land-Use Department,
County Wexford, Ireland

Abstract. Motivated by the increasing interest for the task of multi-label classification (MLC) in recent years, in this study we investigate a new approach for decomposition of the output space with the goal to improve the predictive performance. Namely, the structuring of the output/label space is performed by constructing a label hierarchy and then approaching the MLC task as a task of hierarchical multi-label classification (HMLC). Our approach is as follows. We first perform feature ranking for each of the labels separately and then represent each of the labels with its corresponding feature ranking. The construction of the hierarchy is performed by the (hierarchical) clustering of the feature rankings. To this end, we employ four clustering methods: agglomerative clustering with single linkage, agglomerative clustering with complete linkage, balanced k-means and predictive clustering trees. We then use predictive clustering trees to estimate the influence of the constructed hierarchies, i.e., we compare the predictive performance of models without exploiting the hierarchy and models using hierarchies constructed using label co-occurrences or per label feature rankings. Moreover, we investigate the influence of the hierarchy in the context of single models and ensembles of models. We evaluate the proposed approach across 8 datasets. The results show that the proposed method can yield predictive performance boost across several evaluation measures.

Keywords: Multi-label classification · Hierarchy construction
Feature ranking · Structuring of the label space

1 Introduction

Nowadays, the number of new applications of multi-label learning is steadily increasing, hence, the researchers are very interested to develop novel methods and new ideas related to multi-label classification and structuring of the

A. Appice et al. (Eds.): NFMCP 2017, LNAI 10785, pp. 151–166, 2018.
https://doi.org/10.1007/978-3-319-78680-3_11

label/output space. Multi-label classification (MLC) is the task of learning from data examples where each example can be associated with multiple labels. MLC deals with a label dependencies and relations which is orthogonal with existing traditional methods which take into account label independence and learn independent functions from mapping from input space to the output (label) space. The different application problems include video and image annotations (new movie clips, genres), predicting genes and proteins (functional genomics), classification of a tweets and music into emotions, text classification (web-pages, bookmarks, e-mails,...) and others.

The MLC task is typically approached either by decomposing the MLC problem into multiple single class classification problems (i.e., problem transformation methods) or by modifying the algorithms to consider the multiple classes jointly (i.e., algorithm adaptation methods) [12]. In an extensive experimental study Madjarov et al. [7] show that the landscape of MLC methods is not simple: on some datasets problem transformation methods achieve top performance while on other datasets the algorithm adaptation methods are top performing. Furthermore, the study recommends the use of two algorithms for benchmarking: RF-PCT (Random forests of predictive clustering trees, an algorithm adaptation method) [5] and HOMER (Hierarchy Of Multi-label learnERs, a problem transformation method) [13]. The latter divides the label space into subspaces and then constructs classifiers for each of the subspace (e.g., label power set classifiers). This hints that the best performance might be obtained in between the spectrum of the algorithm adaptation and problem transformation methods. In other words, state-of-the-art MLC performance might be obtained by transforming the original MLC problem into several MLC problems and then learn predictive models (preferably using algorithm adaptation methods).

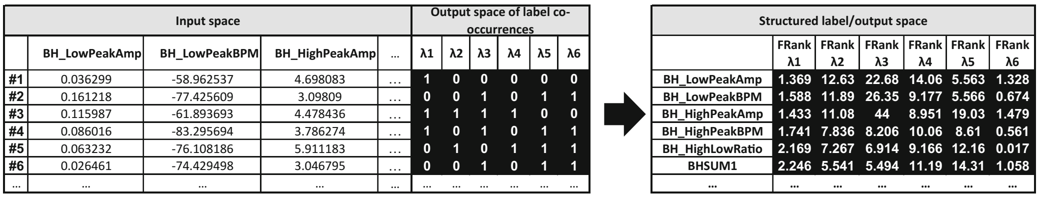
A crucial step in developing methods for output decomposition for MLC is the creation of the subspaces. More specifically, the goal is to find a dependency structure and consider jointly the labels that are inter-dependent. The construction of the output structure of the labels can be very tedious and expensive process, especially if domain experts are needed to complete the task. Moreover, selection of the representation language of the dependencies can be complicated task on its own. Typically, these dependencies are represented as hierarchies of labels [6]. The hierarchies can then be constructed in a data-driven manner using the descriptive space and/or the label space. This presents automatic and relatively efficient process to obtain the representation of the potential dependencies in the label space.

Madjarov et al. [6] present an extensive study of different data-driven methods for constructing label hierarchies for multi-label classification by using the label co-occurence matrix. More precisely, the hierarchies are constructed using four clustering algorithms, agglomerative clustering with single and complete linkage, balanced k-means and predictive clustering trees applied on the label co-occurrences (see Fig. 1, left table).

Next, Szymansky et al. [11] address the question whether data-driven approach using label co-occurrence graph is significantly better than a random choice in label space division for multi-label classification as performed by RAkELd. Their results show that in almost all cases data-driven partitioning outperforms the baseline RAkELd in all evaluation measures, but Hamming loss.

In this study, we build upon the idea of decomposition of the output space and we present a different approach for data-driven structuring of label space in multi-label classification. Our approach constructs the label hierarchy by clustering the per label feature rankings. Namely, instead of using the original label space consisting of label co-occurrences (see Fig. 1, left table), we calculate a feature importance/ranking scores of the features for each label by using the GENIE3 method for feature importance calculation coupled with the random forest ensemble learning method [1,3] (see Fig. 1, right table).

The obtained structure is then used as the label hierarchy and the MLC task is addressed as hierarchical multi-label classification (HMLC) [5,15]. We thus evaluate whether considering the dependency in the label space can provide better predictive performance than addressing MLC as a flat problem. In other words, we investigate whether considering the MLC task as a hierarchical MLC task can yield better predictive performance. Our approach is illustrated through the example in Fig. 1. The table on the left hand-side shows the construction of the label hierarchy using the label co-occurrence (as performed in [6,11]), while the table on the right hand-side shows our proposed method for constructing the label hierarchy.



Input space					Output space of label co-occurrences					
	BH_LowPeakAmp	BH_LowPeakBPM	BH_HighPeakAmp	...	λ1	λ2	λ3	λ4	λ5	λ6
#1	0.036299	-58.962537	4.698083	...	1	0	0	0	0	0
#2	0.161218	-77.425609	3.09809	...	0	0	1	0	1	1
#3	0.115987	-61.893693	4.478436	...	1	1	1	1	0	0
#4	0.086016	-83.295694	3.786274	...	1	0	1	0	1	1
#5	0.063232	-76.108186	5.911183	...	0	1	0	1	1	1
#6	0.026461	-74.429498	3.046795	...	0	0	1	0	1	1
...	...	...	...	...	...	...	...	...	...	...

Structured label/output space						
	FRank λ1	FRank λ2	FRank λ3	FRank λ4	FRank λ5	FRank λ6
BH_LowPeakAmp	1.369	12.63	22.68	14.06	5.563	1.328
BH_LowPeakBPM	1.588	11.89	26.35	9.177	5.566	0.674
BH_HighPeakAmp	1.433	11.08	44	8.951	19.03	1.479
BH_HighPeakBPM	1.741	7.836	8.206	10.06	8.61	0.561
BH_HighLowRatio	2.169	7.267	6.914	9.166	12.16	0.017
BHSUM1	2.246	5.541	5.494	11.19	14.31	1.058
...	...	...	...	...	...	...

Fig. 1. Excerpt from the original *emotions* dataset showing the output space consists of label co-occurrences (*left* table) and the space consists of ranks of the features for each of the labels, separately (*right* table). The former is obtained with structuring the original label set using feature ranking.

We perform an experimental evaluation using 8 benchmark datasets from different domains: text, image, music and video classification, and gene function prediction. The predictive performance of the methods is assessed using 13 different evaluation measures used in the context of MLC (6 threshold dependent and 7 threshold independent).

The obtained results indicate that using the methods for creating the hierarchies using feature ranking can yield a better predictive performance as compared to the original flat MLC methods without the hierarchy. Moreover, using the hierarchy constructed by structuring of the output space using the feature rankings of the labels gives better predictive performance compared to using the hierarchy obtained using the label co-occurrences.

The reminder of this paper is organized as follows. Section 2 presents the background work, i.e., discussion on the tasks of multi-label classification and hierarchical multi-label classification methods. Then, in Sect. 3, we present the structuring of the output space using feature ranking. In Sect. 4, we show the experimental design. The results obtained from the experiments are presented and discussed in Sect. 5. Finally, Sect. 6 concludes this paper.

2 Background

In this section, we first define the task of multi-label classification and then the task of hierarchical multi-label classification. Multi-label learning considers learning from examples which are associated to more than one label coming from a predefined set of labels containing all possible labels. There are two types of multi-label learning tasks: multi-label classification and multi-label ranking. The main goal of multi-label classification is to create a predictive model that will output a set of relevant labels for a given, previously unseen example. Multi-label ranking, on the other hand, can be understood as learning a model that, for each unseen examples, associates a list of rankings (preferences) on the labels from a given set of possible labels and a bipartite partition of this set into relevant and irrelevant labels. An extensive bibliography of methods for multi-label learning can be found in [7,14] and the references therein.

The task of multi-label learning can be defined as follows [5]. The input space $\mathcal{X}$ consists of vectors of values of nominal or numeric data types i.e., $\forall x_i \in \mathcal{X}, x_i = (x_{i1}, x_{i2}, \ldots x_{iD})$, where D is a number of descriptive attributes. The output space $\mathcal{Y}$ consists of a subset of a finite set of disjoint labels $\mathcal{L} = \{\lambda_1, \lambda_2, \ldots, \lambda_Q\}$ (Q > 1 and $\mathcal{Y} \subseteq \mathcal{L}$). Given this, each example is a pair of a vector and a set from the input and output space, respectively. All of the examples then form the set of examples (i.e., the dataset) E. The goal is then to find a function $h : \mathcal{X} \rightarrow 2^{\mathcal{L}}$ such that from the input space assigns a set of labels to each example.

The main difference between multi-label classification and hierarchical multi-label classification (HMLC) is that in the latter the labels from the label space are organized into a hierarchy. A given example labeled with a given label it is also labeled with all its parent labels (known as the hierarchy constraint). Furthermore, an example can be labeled with multiple labels, simultaneously. That means a several paths can be followed from the root node in order to arrive at a given label.

Here, the output space $\mathcal{Y}$ is defined with a label hierarchy $(\mathcal{L}, \leq_h)$, where $\mathcal{L}$ is a set of labels and $\leq_h$ is a partial order parent-child relationship structured as a tree $(\forall \lambda_1, \lambda_2 \in \mathcal{L} : \lambda_1 \leq_h \lambda_2$ if and only if λ_1 is a parent of $\lambda_2)$ [5]. Each example from the set of examples E is a pair of a vector and a set from the input and output space respectively, where the set satisfies the hierarchy constraint, i.e., $E = \{(x_i, \mathcal{Y}_i) | x_i \in \mathcal{X}, \mathcal{Y} \subseteq \mathcal{L}, \lambda \in \mathcal{Y}_i \Rightarrow \forall \lambda' \leq_h \lambda : \lambda' \in \mathcal{Y}_i, 1 \leq i \leq N\}$, where N is a number of examples in E. Same conditions as in multi-label classification should be satisfied for the quality criterion q (high predictive performance and low computational cost). In [9], an extensive bibliography is given, where the HMLC task is presented across different application domains.

3 Structuring of Label Spaces Using Feature Ranking

In this section, we explain our method for structuring the label space using feature ranking and we describe the different clustering algorithms used in this work. Our proposed method for label space structuring is outlined in procedure *StructuringLabelSpaceFR* in Table 1. First, we take the original training dataset D^{train} and using random forest method with GENIE3 feature importance, we create feature rankings for each label separately. We then construct a dataset D^{ranks} consisting of the feature rankings. Next, we obtain a hierarchy using one of the clustering algorithms described bellow. The hierarchy is then used to preprocess the datasets and obtain their hierarchical variants D_H^{train} and D_H^{test}. At the end, we learn the HMLC predictive models.

Table 1. The algorithm for structuring the label space using feature rankings per label.

```
procedure StructuringLabelSpaceFR($D^{train}$, $D^{test}$) returns performance
 1: // create feature importance (.fimp) file with Random forest (GENIE3)
 2: FimpPath = CreateFimp($D^{train}$);
 3: // Create new arff with feature ranks from fimp file
 4: $D^{ranks}$ = CreateArffFromFimp(FimpPath);
 5: hierarchy = Clustering($D^{ranks}$);
 6: //transform multi-label dataset to hierarchical multi-label one
 7: $D_H^{train}$ = MLC2HMC($D^{train}$, hierarchy);
 8: $D_H^{test}$ = MLC2HMC($D^{test}$, hierarchy);
 9: //solve transformed hierarchical multi-label problem by using approach for HMC
10: HMCModel = HMCMethod($D_H^{train}$);
11: //generate HMC predictions using CLUS platform
12: predictions = HMCModel($D_H^{test}$);
13: //Extract predictions only for the leaves from the HMC predictions
14: P = ExtractLeavesPredictionsFromHMCPredictions(predictions);
15: return Evaluate(P)
```

In our approach, described in a procedure *StructuringLabelSpaceFR* (Table 1), we can see that additional step, compare to the algorithm given by Madjarov et al. [6], is the function *CreateFimp* at line 4, which increases the theoretical complexity of the procedure. According to the dimensionality of the space which is going to be clustered using the function *Clustering* at line 5, one dimension in the space consists of label co-occurrences is the number of examples (instances) which means that in case of more complex datasets with large number of examples, the clustering procedure will take more of the time in order to create a hierarchy. From the other side, the procedure of creating the hierarchy using feature rankings has a dimension which depends of the feature space cardinality. Typically, the feature space cardinality is much smaller than the number of examples. It means that clustering of the rankings will finish faster than clustering of the label co-occurrences for datasets with large number of examples but small number of features, which is a case in most of the benchmarks datasets available. Consequently, although we have additional function in our procedure of structuring of the output space, for more complex datasets with high number of examples and smaller number of features, the clustering procedure, i.e., the hierarchy creation will be completed in a reasonable time, thus compensating for obtaining the feature rankings.

We next describe the procedures for obtaining the feature rankings. Random forests as ensemble method for predictive modeling are originally proposed by Breiman [1]. The empirical analysis of their use as feature ranking methods has been studied by Verikas et al. [16]. The random forests are constructed by first performing bootstrap sampling on the data and then building a decision tree for each bootstrap sample. The decision trees are constructed by taking the best split at each level, from a randomly selected feature subset.

Huynh-Thu et al. [3] propose to use the reduction of the variance in the output space at each test node in the tree (the resulting algorithm is named GENIE3). Namely, the variables that reduce the variance of the output more are, consequently, more important than the ones that reduce the variance less. Hence, for each descriptive variable we measure the reduction of variance it produces when selected as splitting variable. If a variable is never selected as splitting variable then its importance will be 0.

The GENIE3 algorithm has been heavily evaluated for single-target regression tasks (e.g., for gene regulatory network reconstruction). The basic idea adopted for future ranking is the same of that proposed in GENIE3, but we use random forest of predictive clustering trees (PCTs) for building the ensemble. The result is a feature ranking algorithm that works for different types of structure output prediction tasks (including MLC and HMLC).

Furthermore, we discuss the different clustering methods used to obtain the hierarchies of the labels. For achieving a good performance of the HMLC methods, it is critical to generate label hierarchies that more closely capture the relations among the labels. The only constraint when building the hierarchy is that we should take care about the leaves of the label hierarchies. They need to define the original MLC task. In particular, the labels from the original MLC problem represent the leaves of the label hierarchy, while the labels in internal nodes of the tree are so-called meta-labels. Meta-labels model the potential relations among the original labels.

For obtaining the hierarchies, we use four different clustering methods (two agglomerative and two divisive):

– agglomerative clustering with single linkage;
– agglomerative clustering with complete linkage;
– balanced k-means clustering (*divisive*) and
– predictive clustering trees (*divisive*).

Agglomerative clustering algorithms consider each example as separate cluster at the beginning and then iteratively merge pairs of clusters based on their distance metric (linkage). If we use the maximal distance of two examples from the clusters C_1 and C_2, then this type of agglomerative clustering is using *complete* linkage, i.e., $\max\{dist(c_1, c_2) : c_1 \in C_1, c_2 \in C_2\}$. If we use the minimal distance between two clusters, then the agglomerative clustering approach is with *single* linkage i.e., $\min\{dist(c_1, c_2) : c_1 \in C_1, c_2 \in C_2\}$.

Balanced k-means is top-down approach for clustering. First, all labels from the label space $\mathcal{L}$ are in one common cluster at the top node of the hierarchy. Then, the procedure consecutively divides (splits) this cluster into k disjoint sub-clusters ($k < |\mathcal{L}_n|$) using k-means clustering. The division also is concerned with the number of examples in each cluster: the algorithm outputs clusters with approximately equal size [13]. The procedure recursively is repeated on each sub-cluster (meta-label) until we have n different clusters consisting of one label from the label space $\mathcal{L}$. In other words, our label space $\mathcal{L}$ is covered by leaves of the hierarchy obtained by the balanced k-means clustering approach.

We also use predictive clustering trees to construct the label hierarchies. More specifically, the setting from the predictive clustering framework used in this work is based on treating the target space as descriptive space, i.e., the target space is also a descriptive space. Descriptive/target variables are used to provide descriptions for the obtained clusters. Here, the focus is using predictive clustering framework on the task of clustering instead of predictive modelling [2,4]. The obtained hierarchies using agglomerative clustering (single and complete linkage) and using predictive clustering trees for *emotions* dataset are shown in Fig. 2.

We next present the predictive clustering trees (PCTs) - the modelling framework we used throughout this work. PCTs are a generalization of decision trees towards the tasks of predicting structured outputs, including both MLC and HMLC. In order to apply PCTs to the task of HMLC, Vens et al. [15] define the variance and the prototype as follows. First, the set of labels for each example is represented as a vector of binary components. If the example belongs to the class c_i then the i'th component of the vector is 1 and 0, otherwise. The variance of a set of examples E is thus defined as follows:

$$Var(E) = \frac{1}{|E|} \cdot \sum_{i=1}^{|E|} dist(\Gamma_i, \overline{\Gamma})^2 \tag{1}$$

where $\overline{\Gamma} = \frac{1}{|E|} \cdot \sum_{i=1}^{|E|} \Gamma_i$.

In other words, the variance $Var(E)$ in (1) represents the average squared distance between each example's class vector (Γ_i) and the mean class vector of the set ($\overline{\Gamma}$). When we talk about HMC, then the similarity at higher levels of the hierarchy are more important than the similarity at lower levels. This is reflected with the distance term used in (1), which is weighted Euclidean distance:

$$dist(\Gamma_1, \Gamma_2) = \sqrt{\sum_{s=1}^{|\Gamma|} \theta(c_s) \cdot (\Gamma_{1,s} - \Gamma_{2,s})^2}$$

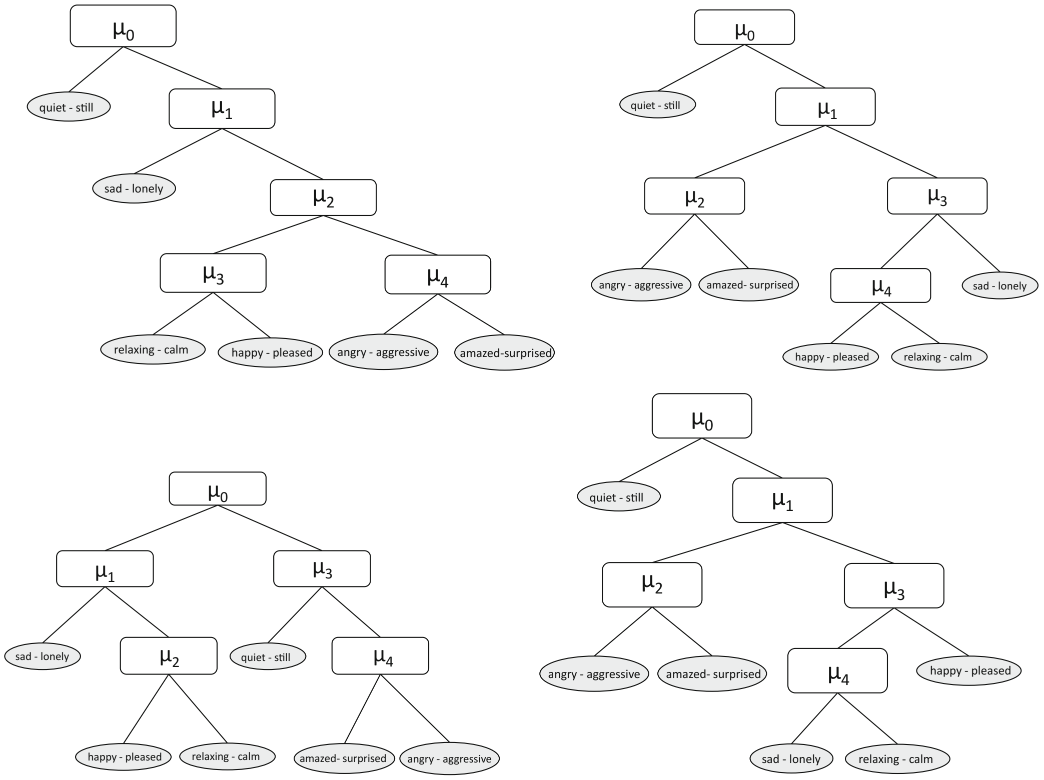


Fig. 2. Hierarchies obtained using agglomerative single (top-left), agglomerative complete (top-right), balanced K-means clustering (bottom - left) and PCTs (bottom - right) clustering methods for *emotions* dataset.

where $\Gamma_{i,s}$ is the s'th component of the class vector Γ_i of the instance E_i, $|\Gamma|$ is the size of the class vector, and the class weights $\theta(c) = \theta_0 \cdot \operatorname{avg}_j\{\theta(p_j(c))\}$, where $p_j(c)$ is j'th parent of the class c and $0 < \theta_0 < 1$. The class weights $\theta(c)$ decrease with the depth of the class in the hierarchy thus making the differences in the lower parts of the hierarchy less influential to the overall score.

Random forests of PCTs for HMLC are considered in the same way as the random forest of PCTs for MLC. In the case of HMLC, the ensemble is a set of PCTs for HMLC. A new example is classified by taking a majority vote from the combined predictions of the member classifiers. The prediction of the random forest ensemble of PCTs for HMLC follows the hierarchy constraint (if the example is labeled with a given label then is automatically labeled with all its ancestor-labels).

4 Experimental Design

The aim of our study is to address the following questions:

(i) Whether feature ranking on the label (output) space in the MLC task can be used to construct good label hierarchies?

(ii) Which clustering method yields better hierarchy?
(iii) How this scales from single model to ensemble of models?
(iv) Can we achieve better predictive models with using a hierarchies obtained by structuring the feature ranking or co-occurrences space?

In order to answer the above questions, we use eight multi-label classification benchmark problems from different domains. We have 3 datasets from text classification, 4 datasets from multimedia, includes movie clips and genres classification and 1 dataset from biology. All datasets are predefined by other researchers (typically the data owners) and divided into train and test subsets. The basic information and statistics about these datasets are given in Table 2.

Table 2. Statistics of used benchmark tasks in terms of application domain (*domain*), number of training examples ($\#tr.e$), testing examples ($\#t.e$), number of descriptors (D), total number of labels (L) and number of labels per example.

Dataset	Domain	$\#tr.e$	$\#t.e$	D	L	l_c
emotions	multimedia	391	202	72	6	1.87
scene	multimedia	1211	1159	294	6	1.07
yeast	biology	1500	917	103	14	4.24
tmc2007	text	21519	7077	500	22	2.16
medical	text	645	333	1449	45	1.25
enron	text	1123	579	1001	53	3.38
mediamill	multimedia	30993	12914	120	101	4.38
corel5k	multimedia	4500	500	499	374	3.52

In our experiments, we use 13 different evaluation measures, as presented in [7,14]. These are divided into two groups: 6 threshold dependent/example based measures (*hamming loss, accuracy, precision, recall*, F_1 *score*) and 7 threshold independent measures out of which three ranking-based (*one-error, coverage* and *ranking-loss*) and four areas under ROC and PRC curves (*AUROC, AUPRC, wAUPRC* and *pooledAUPRC*). The threshold independent measures are typically used in HMLC and they do not require a (pre)selection of thresholds and calculating a prediction [15]. All of the above measures offer different viewpoints on the results from the experimental evaluation.

Hamming loss is an example-based evaluation measure that evaluate how many times a pair of example and its label are misclassified. *One-error* is a ranking-based evaluation measure that evaluates how many times the top-ranked label does not exist in a set of relevant labels of the example. *Coverage* evaluates how far, on average, we need to go down the list of label ranks in order to cover all relevant labels of given example. *Ranking loss* evaluates the average fraction of the label pairs that are reversely ordered for the given example. Precision and recall are very important measures defined for binary classification tasks with classes of positive and negative examples. *Precision* is a proportion of positive prediction that are correct, and *recall* is the proportion of positive examples that correctly predicted as positive. F_1 *score* is the harmonic mean between precision and recall. *Accuracy* for each instance is defined as the proportion of correctly predicted labels over total number of labels for that instance. Overall accuracy is the average across all instances. A precision-recall curve (PR curve) is a curve that represent the precision as a function of its recall. $AUPRC$ (area under the PR curve) is the area between the PR curve and the recall axis. $wAUPRC$ evaluates the weighted average of the areas under the individual (per class) PR-curves. If choosing some threshold, we transform the multi-label problem into binary problems with considering binary classifier as a couple (instance, class) and predicting whether that instance belongs to that class, we can obtain PR curves that differ depend of the varying the threshold. The area under the average PR curve (from all different threshold curves) is called $pooledAUPRC$. From the other side, if we consider the space of true positive rates (sensitivity) versus false positive rates (fall-out) then the curve considers the sensitivity as a function of the fall-out is called ROC-curve. The are under this ROC-curve is the evaluation measure called $AUROC$.

The majority of our experiments are performed using the CLUS software package (https://sourceforge.net/projects/clus/), which implements the predictive clustering framework, including PCTs, random forests of PCTs and feature ranking [5,10]. A hierarchical tree defined by the used clustering methods in HMLC setting are defined as tree shaped hierarchies. We use the same values for k in balanced k-means clustering algorithm, as suggested in [7].

For obtaining a hierarchy using the agglomerative clustering method we use the R software package (function *agnes()* from the *cluster* package. For more info, see https://stat.ethz.ch/R-manual/R-devel/library/cluster/html/agnes.html). We use the MATLAB software package to create hierarchies with balanced k-means clustering which is based on Hungarian (Munkres') assignment algorithm to assign the examples to the clusters [8]. We use Euclidean distance metric in all our algorithms that require distance. Moreover, for random forest for feature ranking we use GENIE3 as a feature importance method based on variable selection with ensembles of PCTs [3].

In order to make a comparative analysis with the results obtained by the study by Madjarov et al. [6], we repeated their experiments on the same experimental setting with the experiments we perform for feature ranking.

5 Results

In this section, we present the obtained results from the experiments we performed using our novel proposed method for structuring the output space. In our study, as an output space, we consider the space consisting of label co-occurrences (as presented by Madjarov et al. [6]) and the space consisting of feature ranks for each label, respectively. We compare the following methods for hierarchy construction:

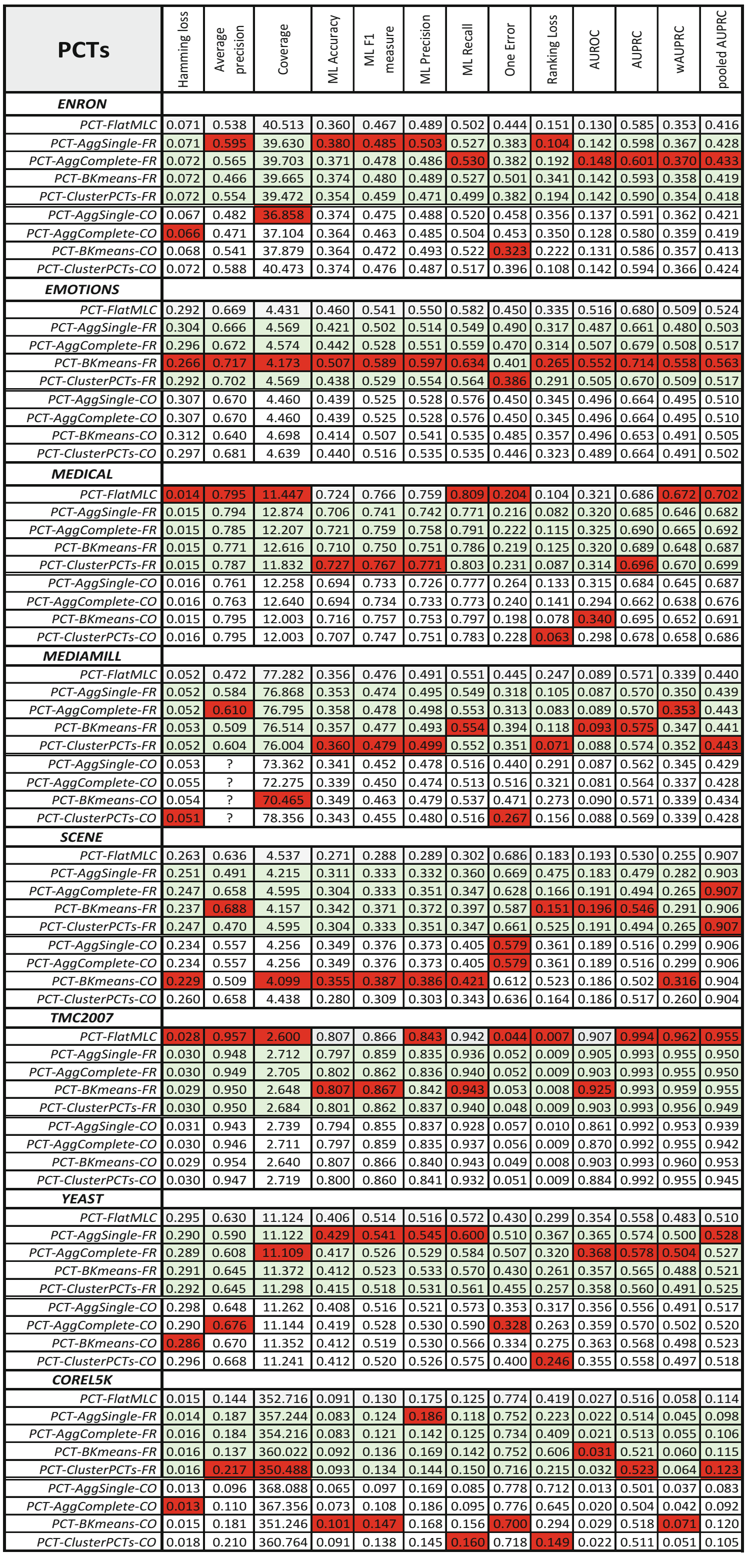
- flat MLC problem without considering a hierarchy in the label space (*FlatMLC*);
- agglomerative clustering with single linkage (*AggSingle*);
- agglomerative clustering with complete linkage (*AggComplete*)
- balanced k-means clustering (*BKmeans*)
- clustering using predictive clustering trees (*ClusPCTs*).

Since we have two different models (single PCTs model and random forest of PCTs) and two different structured output spaces, we show separately the results for single PCTs (Fig. 3) and random forest of PCTs (Fig. 4). In order to distinguish between using either single tree or random forest of PCTs and different methods of structuring the output space (label co-occurrences and feature rankings), we use prefixes (*PCT-* and *RF-*) and suffixes (*-CO* and *-FR*) before and after the hierarchy construction method name, respectively. For example, *RF-AggComplete-CO* refers to the agglomerative clustering method with complete linkage of the output space of label co-occurrences using random forest of PCTs for model creation. Then, *PCT-ClusPCTs-FR* refers to the clustering method with PCTs of the output space consists of feature rankings per label using single PCTs for model creation, etc.

Observing the results obtained using single PCTs (Fig. 3), we can note that there is no clear winner across all evaluation measures and datasets. In the case of threshold independent measures, such as *AUPRC, AUROC, wAUPRC* and *pooledAUPRC*, we can see that hierarchies created using clustering of the output space consisting of feature rankings perform the best for enron, emotions, mediamill and yeast datasets. Considering the scene and corel5k datasets, we can observe that they perform the best according to *AUROC, AUPRC* and *pooledAUPRC*, but not for *wAUPRC*. *PCT-BKmeans-FR* outperforms the other algorithms for hierarchy creation in the emotions dataset according to the most of the evaluation measures but not according to one-error. Moreover, the hierarchies created clustering the feature rankings outperform the other algorithms considering the ML performance measures (*ML F1 measure, ML accuracy, ML precision* and *ML recall*) in 5 out of the 8 datasets.

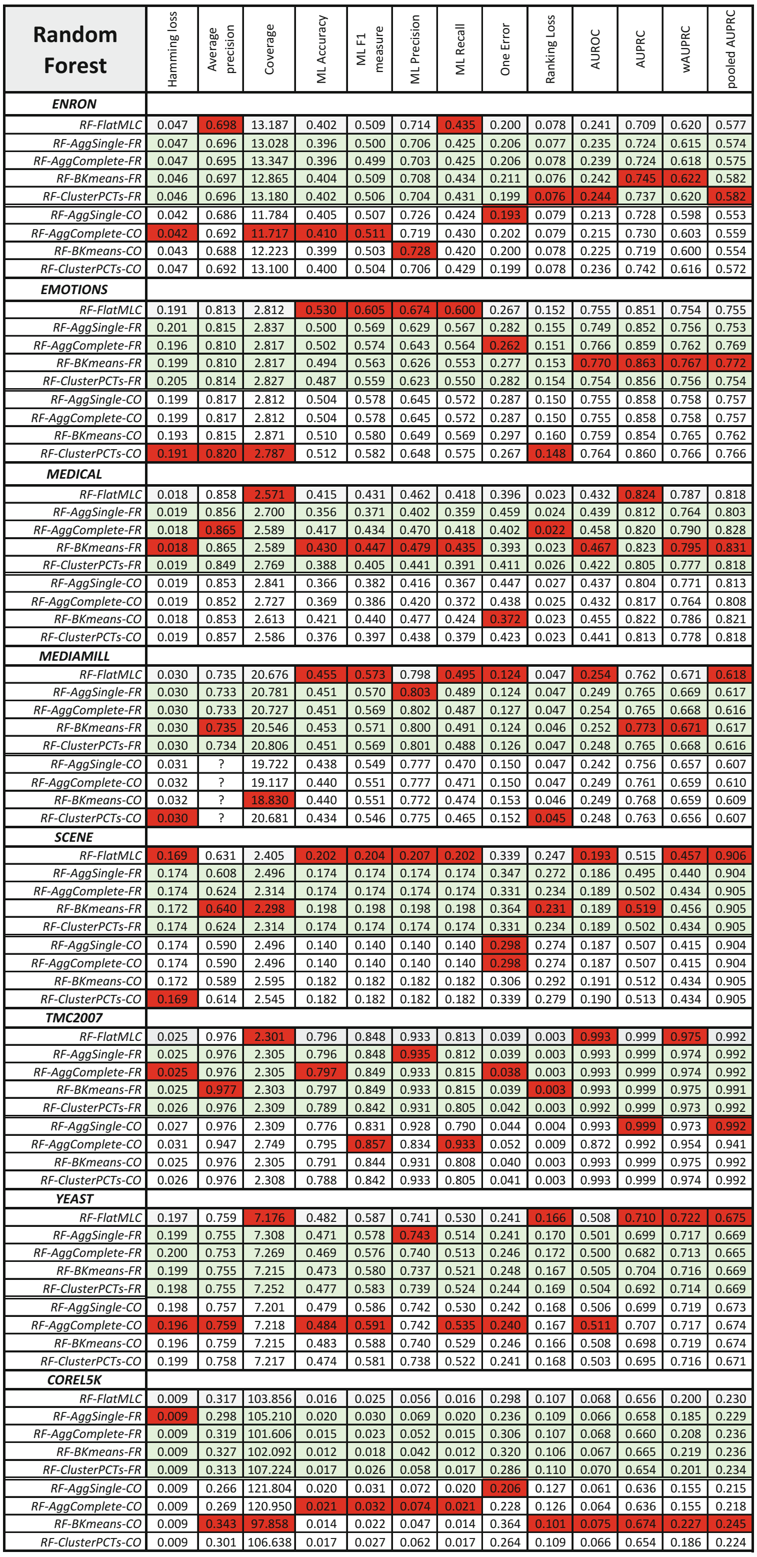
Generally, structuring the output space consisting of feature rankings for each label yields better predictive performance compared to the structuring the output space consisting of label co-occurrences considering most of the evaluation measures in almost all datasets. For the corel5k dataset only, we can see that both have similar performance. If we consider medical and tmc2007 datasets, we can see that structuring the output space does not improve the performance as compared to the flat MLC task, where there is no hierarchy considered. All in all, we can conclude that using the hierarchies, the predictive performance can be improved.

The results obtained when random forests are used as predictive models are given in Fig. 4. These results present a different situation as compared to the results obtained when single PCTs are used as predictive models. First of all, the predictive performance is improved as compared to the single PCTs for large majority of the cases. Most notably, the performance for the threshold



PCTs	Hamming loss	Average precision	Coverage	ML Accuracy	ML F1 measure	ML Precision	ML Recall	One Error	Ranking Loss	AUROC	AUPRC	wAUPRC	pooled AUPRC
ENRON													
PCT-FlatMLC	0.071	0.538	40.513	0.360	0.467	0.489	0.502	0.444	0.151	0.130	0.585	0.353	0.416
PCT-AggSingle-FR	0.071	0.595	39.630	0.380	0.485	0.503	0.527	0.383	0.104	0.142	0.598	0.367	0.428
PCT-AggComplete-FR	0.072	0.565	39.703	0.371	0.478	0.486	0.530	0.382	0.192	0.148	0.601	0.370	0.433
PCT-BKmeans-FR	0.072	0.466	39.665	0.374	0.480	0.489	0.527	0.501	0.341	0.142	0.593	0.358	0.419
PCT-ClusterPCTs-FR	0.072	0.554	39.472	0.354	0.459	0.471	0.499	0.382	0.194	0.142	0.590	0.354	0.418
PCT-AggSingle-CO	0.067	0.482	36.858	0.374	0.475	0.488	0.520	0.458	0.356	0.137	0.591	0.362	0.421
PCT-AggComplete-CO	0.066	0.471	37.104	0.364	0.463	0.485	0.504	0.453	0.350	0.128	0.580	0.359	0.419
PCT-BKmeans-CO	0.068	0.541	37.879	0.364	0.472	0.493	0.522	0.323	0.222	0.131	0.586	0.357	0.413
PCT-ClusterPCTs-CO	0.072	0.588	40.473	0.374	0.476	0.487	0.517	0.396	0.108	0.142	0.594	0.366	0.424
EMOTIONS													
PCT-FlatMLC	0.292	0.669	4.431	0.460	0.541	0.550	0.582	0.450	0.335	0.516	0.680	0.509	0.524
PCT-AggSingle-FR	0.304	0.666	4.569	0.421	0.502	0.514	0.549	0.490	0.317	0.487	0.661	0.480	0.503
PCT-AggComplete-FR	0.296	0.672	4.574	0.442	0.528	0.551	0.559	0.470	0.314	0.507	0.679	0.508	0.517
PCT-BKmeans-FR	0.266	0.717	4.173	0.507	0.589	0.597	0.634	0.401	0.265	0.552	0.714	0.558	0.563
PCT-ClusterPCTs-FR	0.292	0.702	4.569	0.438	0.529	0.554	0.564	0.386	0.291	0.505	0.670	0.509	0.517
PCT-AggSingle-CO	0.307	0.670	4.460	0.439	0.525	0.528	0.576	0.450	0.345	0.496	0.664	0.495	0.510
PCT-AggComplete-CO	0.307	0.670	4.460	0.439	0.525	0.528	0.576	0.450	0.345	0.496	0.664	0.495	0.510
PCT-BKmeans-CO	0.312	0.640	4.698	0.414	0.507	0.541	0.535	0.485	0.357	0.496	0.653	0.491	0.505
PCT-ClusterPCTs-CO	0.297	0.681	4.639	0.440	0.516	0.535	0.535	0.446	0.323	0.489	0.664	0.491	0.502
MEDICAL													
PCT-FlatMLC	0.014	0.795	11.447	0.724	0.766	0.759	0.809	0.204	0.104	0.321	0.686	0.672	0.702
PCT-AggSingle-FR	0.015	0.794	12.874	0.706	0.741	0.742	0.771	0.216	0.082	0.320	0.685	0.646	0.682
PCT-AggComplete-FR	0.015	0.785	12.207	0.721	0.759	0.758	0.791	0.222	0.115	0.325	0.690	0.665	0.692
PCT-BKmeans-FR	0.015	0.771	12.616	0.710	0.750	0.751	0.786	0.219	0.125	0.320	0.689	0.648	0.687
PCT-ClusterPCTs-FR	0.015	0.787	11.832	0.727	0.767	0.771	0.803	0.231	0.087	0.314	0.696	0.670	0.699
PCT-AggSingle-CO	0.016	0.761	12.258	0.694	0.733	0.726	0.777	0.264	0.133	0.315	0.684	0.645	0.687
PCT-AggComplete-CO	0.016	0.763	12.640	0.694	0.734	0.733	0.773	0.240	0.141	0.294	0.662	0.638	0.676
PCT-BKmeans-CO	0.015	0.795	12.003	0.716	0.757	0.753	0.797	0.198	0.078	0.340	0.695	0.652	0.691
PCT-ClusterPCTs-CO	0.016	0.795	12.003	0.707	0.747	0.751	0.783	0.228	0.063	0.298	0.678	0.658	0.686
MEDIAMILL													
PCT-FlatMLC	0.052	0.472	77.282	0.356	0.476	0.491	0.551	0.445	0.247	0.089	0.571	0.339	0.440
PCT-AggSingle-FR	0.052	0.584	76.868	0.353	0.474	0.495	0.549	0.318	0.105	0.087	0.570	0.350	0.439
PCT-AggComplete-FR	0.052	0.610	76.795	0.358	0.478	0.498	0.553	0.313	0.083	0.089	0.570	0.353	0.443
PCT-BKmeans-FR	0.053	0.509	76.514	0.357	0.477	0.493	0.554	0.394	0.118	0.093	0.575	0.347	0.441
PCT-ClusterPCTs-FR	0.052	0.604	76.004	0.360	0.479	0.499	0.552	0.351	0.071	0.088	0.574	0.352	0.443
PCT-AggSingle-CO	0.053	?	73.362	0.341	0.452	0.478	0.516	0.440	0.291	0.087	0.562	0.345	0.429
PCT-AggComplete-CO	0.055	?	72.275	0.339	0.450	0.474	0.513	0.516	0.321	0.081	0.564	0.337	0.428
PCT-BKmeans-CO	0.054	?	70.465	0.349	0.463	0.479	0.537	0.471	0.273	0.090	0.571	0.339	0.434
PCT-ClusterPCTs-CO	0.051	?	78.356	0.343	0.455	0.480	0.516	0.267	0.156	0.088	0.569	0.339	0.428
SCENE													
PCT-FlatMLC	0.263	0.636	4.537	0.271	0.288	0.289	0.302	0.686	0.183	0.193	0.530	0.255	0.907
PCT-AggSingle-FR	0.251	0.491	4.215	0.311	0.333	0.332	0.360	0.669	0.475	0.183	0.479	0.282	0.903
PCT-AggComplete-FR	0.247	0.658	4.595	0.304	0.333	0.351	0.347	0.628	0.166	0.191	0.494	0.265	0.907
PCT-BKmeans-FR	0.237	0.688	4.157	0.342	0.371	0.372	0.397	0.587	0.151	0.196	0.546	0.291	0.906
PCT-ClusterPCTs-FR	0.247	0.470	4.595	0.304	0.333	0.351	0.347	0.661	0.525	0.191	0.494	0.265	0.907
PCT-AggSingle-CO	0.234	0.557	4.256	0.349	0.376	0.373	0.405	0.579	0.361	0.189	0.516	0.299	0.906
PCT-AggComplete-CO	0.234	0.557	4.256	0.349	0.376	0.373	0.405	0.579	0.361	0.189	0.516	0.299	0.906
PCT-BKmeans-CO	0.229	0.509	4.099	0.355	0.387	0.386	0.421	0.612	0.523	0.186	0.502	0.316	0.904
PCT-ClusterPCTs-CO	0.260	0.658	4.438	0.280	0.309	0.303	0.343	0.636	0.164	0.186	0.517	0.260	0.904
TMC2007													
PCT-FlatMLC	0.028	0.957	2.600	0.807	0.866	0.843	0.942	0.044	0.007	0.907	0.994	0.962	0.955
PCT-AggSingle-FR	0.030	0.948	2.712	0.797	0.859	0.835	0.936	0.052	0.009	0.905	0.993	0.955	0.950
PCT-AggComplete-FR	0.030	0.949	2.705	0.802	0.862	0.836	0.940	0.052	0.009	0.903	0.993	0.955	0.950
PCT-BKmeans-FR	0.029	0.950	2.648	0.807	0.867	0.842	0.943	0.053	0.008	0.925	0.993	0.959	0.955
PCT-ClusterPCTs-FR	0.030	0.950	2.684	0.801	0.862	0.837	0.940	0.048	0.009	0.903	0.993	0.956	0.949
PCT-AggSingle-CO	0.031	0.943	2.739	0.794	0.855	0.837	0.928	0.057	0.010	0.861	0.992	0.953	0.939
PCT-AggComplete-CO	0.030	0.946	2.711	0.797	0.859	0.835	0.937	0.056	0.009	0.870	0.992	0.955	0.942
PCT-BKmeans-CO	0.029	0.954	2.640	0.807	0.866	0.840	0.943	0.049	0.008	0.903	0.993	0.960	0.953
PCT-ClusterPCTs-CO	0.030	0.947	2.719	0.800	0.860	0.841	0.932	0.051	0.009	0.884	0.992	0.955	0.945
YEAST													
PCT-FlatMLC	0.295	0.630	11.124	0.406	0.514	0.516	0.572	0.430	0.299	0.354	0.558	0.483	0.510
PCT-AggSingle-FR	0.290	0.590	11.122	0.429	0.541	0.545	0.600	0.510	0.367	0.365	0.574	0.500	0.528
PCT-AggComplete-FR	0.289	0.608	11.109	0.417	0.526	0.529	0.584	0.507	0.320	0.368	0.578	0.504	0.527
PCT-BKmeans-FR	0.291	0.645	11.372	0.412	0.523	0.533	0.570	0.430	0.261	0.357	0.565	0.488	0.521
PCT-ClusterPCTs-FR	0.292	0.645	11.298	0.415	0.518	0.531	0.561	0.455	0.257	0.358	0.560	0.491	0.525
PCT-AggSingle-CO	0.298	0.648	11.262	0.408	0.516	0.521	0.573	0.353	0.317	0.356	0.556	0.491	0.517
PCT-AggComplete-CO	0.290	0.676	11.144	0.419	0.528	0.530	0.590	0.328	0.263	0.359	0.570	0.502	0.520
PCT-BKmeans-CO	0.286	0.670	11.352	0.412	0.519	0.530	0.566	0.334	0.275	0.363	0.568	0.498	0.523
PCT-ClusterPCTs-CO	0.296	0.668	11.241	0.412	0.520	0.526	0.575	0.400	0.246	0.355	0.558	0.497	0.518
COREL5K													
PCT-FlatMLC	0.015	0.144	352.716	0.091	0.130	0.175	0.125	0.774	0.419	0.027	0.516	0.058	0.114
PCT-AggSingle-FR	0.014	0.187	357.244	0.083	0.124	0.186	0.118	0.752	0.223	0.022	0.514	0.045	0.098
PCT-AggComplete-FR	0.016	0.184	354.216	0.083	0.121	0.142	0.125	0.734	0.409	0.021	0.513	0.055	0.106
PCT-BKmeans-FR	0.016	0.137	360.022	0.092	0.136	0.169	0.142	0.752	0.606	0.031	0.521	0.060	0.115
PCT-ClusterPCTs-FR	0.016	0.217	350.488	0.093	0.134	0.144	0.150	0.716	0.215	0.032	0.523	0.064	0.123
PCT-AggSingle-CO	0.013	0.096	368.088	0.065	0.097	0.169	0.085	0.778	0.712	0.013	0.501	0.037	0.083
PCT-AggComplete-CO	0.013	0.110	367.356	0.073	0.108	0.186	0.095	0.776	0.645	0.020	0.504	0.042	0.092
PCT-BKmeans-CO	0.015	0.181	351.246	0.101	0.147	0.168	0.156	0.700	0.294	0.029	0.518	0.071	0.120
PCT-ClusterPCTs-CO	0.018	0.210	360.764	0.091	0.138	0.145	0.160	0.718	0.149	0.022	0.511	0.051	0.105

Fig. 3. Results with the 13 performance measures for *single PCTs* from experiments performed on 8 different datasets. The best results obtained per measure per dataset are highlighted.



Random Forest	Hamming loss	Average precision	Coverage	ML Accuracy	ML F1 measure	ML Precision	ML Recall	One Error	Ranking Loss	AUROC	AUPRC	wAUPRC	pooled AUPRC
ENRON													
RF-FlatMLC	0.047	**0.698**	13.187	0.402	0.509	0.714	**0.435**	0.200	0.078	0.241	0.709	0.620	0.577
RF-AggSingle-FR	0.047	0.696	13.028	0.396	0.500	0.706	0.425	0.206	0.077	0.235	0.724	0.615	0.574
RF-AggComplete-FR	0.047	0.695	13.347	0.396	0.499	0.703	0.425	0.206	0.078	0.239	0.724	0.618	0.575
RF-BKmeans-FR	0.046	0.697	12.865	0.404	0.509	0.708	0.434	0.211	0.076	0.242	**0.745**	**0.622**	0.582
RF-ClusterPCTs-FR	0.046	0.696	13.180	0.402	0.506	0.704	0.431	0.199	**0.076**	**0.244**	0.737	0.620	**0.582**
RF-AggSingle-CO	0.042	0.686	11.784	0.405	0.507	0.726	0.424	**0.193**	0.079	0.213	0.728	0.598	0.553
RF-AggComplete-CO	**0.042**	0.692	**11.717**	**0.410**	**0.511**	0.719	0.430	0.202	0.079	0.215	0.730	0.603	0.559
RF-BKmeans-CO	0.043	0.688	12.223	0.399	0.503	**0.728**	0.420	0.200	0.078	0.225	0.719	0.600	0.554
RF-ClusterPCTs-CO	0.047	0.692	13.100	0.400	0.504	0.706	0.429	0.199	0.078	0.236	0.742	0.616	0.572
EMOTIONS													
RF-FlatMLC	0.191	0.813	2.812	**0.530**	**0.605**	**0.674**	**0.600**	0.267	0.152	0.755	0.851	0.754	0.755
RF-AggSingle-FR	0.201	0.815	2.837	0.500	0.569	0.629	0.567	0.282	0.155	0.749	0.852	0.756	0.753
RF-AggComplete-FR	0.196	0.810	2.817	0.502	0.574	0.643	0.564	**0.262**	0.151	0.766	0.859	0.762	0.769
RF-BKmeans-FR	0.199	0.810	2.817	0.494	0.563	0.626	0.553	0.277	0.153	**0.770**	**0.863**	**0.767**	**0.772**
RF-ClusterPCTs-FR	0.205	0.814	2.827	0.487	0.559	0.623	0.550	0.282	0.154	0.754	0.856	0.756	0.754
RF-AggSingle-CO	0.199	0.817	2.812	0.504	0.578	0.645	0.572	0.287	0.150	0.755	0.858	0.758	0.757
RF-AggComplete-CO	0.199	0.817	2.812	0.504	0.578	0.645	0.572	0.287	0.150	0.755	0.858	0.758	0.757
RF-BKmeans-CO	0.193	0.815	2.871	0.510	0.580	0.649	0.569	0.297	0.160	0.759	0.854	0.765	0.762
RF-ClusterPCTs-CO	**0.191**	**0.820**	**2.787**	0.512	0.582	0.648	0.575	0.267	**0.148**	0.764	0.860	0.766	0.766
MEDICAL													
RF-FlatMLC	0.018	0.858	**2.571**	0.415	0.431	0.462	0.418	0.396	0.023	0.432	**0.824**	0.787	0.818
RF-AggSingle-FR	0.019	0.856	2.700	0.356	0.371	0.402	0.359	0.459	0.024	0.439	0.812	0.764	0.803
RF-AggComplete-FR	0.018	**0.865**	2.589	0.417	0.434	0.470	0.418	0.402	**0.022**	0.458	0.820	0.790	0.828
RF-BKmeans-FR	**0.018**	0.865	2.589	**0.430**	**0.447**	**0.479**	**0.435**	0.393	0.023	**0.467**	0.823	**0.795**	**0.831**
RF-ClusterPCTs-FR	0.019	0.849	2.769	0.388	0.405	0.441	0.391	0.411	0.026	0.422	0.805	0.777	0.818
RF-AggSingle-CO	0.019	0.853	2.841	0.366	0.382	0.416	0.367	0.447	0.027	0.437	0.804	0.771	0.813
RF-AggComplete-CO	0.019	0.852	2.727	0.369	0.386	0.420	0.372	0.438	0.025	0.432	0.817	0.764	0.808
RF-BKmeans-CO	0.018	0.853	2.613	0.421	0.440	0.477	0.424	**0.372**	0.023	0.455	0.822	0.786	0.821
RF-ClusterPCTs-CO	0.019	0.857	2.586	0.376	0.397	0.438	0.379	0.423	0.023	0.441	0.813	0.778	0.818
MEDIAMILL													
RF-FlatMLC	0.030	0.735	20.676	**0.455**	**0.573**	0.798	**0.495**	**0.124**	0.047	**0.254**	0.762	0.671	**0.618**
RF-AggSingle-FR	0.030	0.733	20.781	0.451	0.570	**0.803**	0.489	0.124	0.047	0.249	0.765	0.669	0.617
RF-AggComplete-FR	0.030	0.733	20.727	0.451	0.569	0.802	0.487	0.127	0.047	0.254	0.765	0.668	0.616
RF-BKmeans-FR	0.030	**0.735**	20.546	0.453	0.571	0.800	0.491	0.124	0.046	0.252	**0.773**	**0.671**	0.617
RF-ClusterPCTs-FR	0.030	0.734	20.806	0.451	0.569	0.801	0.488	0.126	0.047	0.248	0.765	0.668	0.616
RF-AggSingle-CO	0.031	?	19.722	0.438	0.549	0.777	0.470	0.150	0.047	0.242	0.756	0.657	0.607
RF-AggComplete-CO	0.032	?	19.117	0.440	0.551	0.777	0.471	0.150	0.047	0.249	0.761	0.659	0.610
RF-BKmeans-CO	0.032	?	**18.830**	0.440	0.551	0.772	0.474	0.153	0.046	0.249	0.768	0.659	0.609
RF-ClusterPCTs-CO	**0.030**	?	20.681	0.434	0.546	0.775	0.465	0.152	**0.045**	0.248	0.763	0.656	0.607
SCENE													
RF-FlatMLC	**0.169**	0.631	2.405	**0.202**	**0.204**	**0.207**	**0.202**	0.339	0.247	**0.193**	0.515	**0.457**	**0.906**
RF-AggSingle-FR	0.174	0.608	2.496	0.174	0.174	0.174	0.174	0.347	0.272	0.186	0.495	0.440	0.904
RF-AggComplete-FR	0.174	0.624	2.314	0.174	0.174	0.174	0.174	0.331	0.234	0.189	0.502	0.434	0.905
RF-BKmeans-FR	0.172	**0.640**	**2.298**	0.198	0.198	0.198	0.198	0.364	**0.231**	0.189	**0.519**	0.456	0.905
RF-ClusterPCTs-FR	0.174	0.624	2.314	0.174	0.174	0.174	0.174	0.331	0.234	0.189	0.502	0.434	0.905
RF-AggSingle-CO	0.174	0.590	2.496	0.140	0.140	0.140	0.140	**0.298**	0.274	0.187	0.507	0.415	0.904
RF-AggComplete-CO	0.174	0.590	2.496	0.140	0.140	0.140	0.140	**0.298**	0.274	0.187	0.507	0.415	0.904
RF-BKmeans-CO	0.172	0.589	2.595	0.182	0.182	0.182	0.182	0.306	0.292	0.191	0.512	0.434	0.905
RF-ClusterPCTs-CO	**0.169**	0.614	2.545	0.182	0.182	0.182	0.182	0.339	0.279	0.190	0.513	0.434	0.905
TMC2007													
RF-FlatMLC	0.025	0.976	**2.301**	0.796	0.848	0.933	0.813	0.039	0.003	**0.993**	0.999	**0.975**	0.992
RF-AggSingle-FR	0.025	0.976	2.305	0.796	0.848	**0.935**	0.812	0.039	0.003	0.993	0.999	0.974	0.992
RF-AggComplete-FR	**0.025**	0.976	2.305	**0.797**	0.849	0.933	0.815	**0.038**	0.003	0.993	0.999	0.974	0.992
RF-BKmeans-FR	0.025	**0.977**	2.303	0.797	0.849	0.933	0.815	0.039	**0.003**	0.993	0.999	0.975	0.991
RF-ClusterPCTs-FR	0.026	0.976	2.309	0.789	0.842	0.931	0.805	0.042	0.003	0.992	0.999	0.973	0.992
RF-AggSingle-CO	0.027	0.976	2.309	0.776	0.831	0.928	0.790	0.044	0.004	0.993	**0.999**	0.973	**0.992**
RF-AggComplete-CO	0.031	0.947	2.749	0.795	**0.857**	0.834	**0.933**	0.052	0.009	0.872	0.992	0.954	0.941
RF-BKmeans-CO	0.025	0.976	2.305	0.791	0.844	0.931	0.808	0.040	0.003	0.993	0.999	0.975	0.992
RF-ClusterPCTs-CO	0.026	0.976	2.308	0.788	0.842	0.933	0.805	0.041	0.003	0.993	0.999	0.974	0.992
YEAST													
RF-FlatMLC	0.197	0.759	**7.176**	0.482	0.587	0.741	0.530	0.241	**0.166**	0.508	**0.710**	**0.722**	**0.675**
RF-AggSingle-FR	0.199	0.755	7.308	0.471	0.578	**0.743**	0.514	0.241	0.170	0.501	0.699	0.717	0.669
RF-AggComplete-FR	0.200	0.753	7.269	0.469	0.576	0.740	0.513	0.246	0.172	0.500	0.682	0.713	0.665
RF-BKmeans-FR	0.199	0.755	7.215	0.473	0.580	0.737	0.521	0.248	0.167	0.505	0.704	0.716	0.669
RF-ClusterPCTs-FR	0.198	0.755	7.252	0.477	0.583	0.739	0.524	0.244	0.169	0.504	0.692	0.714	0.669
RF-AggSingle-CO	0.198	0.757	7.201	0.479	0.586	0.742	0.530	0.242	0.168	0.506	0.699	0.719	0.673
RF-AggComplete-CO	**0.196**	**0.759**	7.218	**0.484**	**0.591**	0.742	**0.535**	**0.240**	0.167	**0.511**	0.707	0.717	0.674
RF-BKmeans-CO	0.196	0.759	7.215	0.483	0.588	0.740	0.529	0.246	0.166	0.508	0.698	0.719	0.674
RF-ClusterPCTs-CO	0.199	0.758	7.217	0.474	0.581	0.738	0.522	0.241	0.168	0.503	0.695	0.716	0.671
COREL5K													
RF-FlatMLC	0.009	0.317	103.856	0.016	0.025	0.056	0.016	0.298	0.107	0.068	0.656	0.200	0.230
RF-AggSingle-FR	**0.009**	0.298	105.210	0.020	0.030	0.069	0.020	0.236	0.109	0.066	0.658	0.185	0.229
RF-AggComplete-FR	0.009	0.319	101.606	0.015	0.023	0.052	0.015	0.306	0.107	0.068	0.660	0.208	0.236
RF-BKmeans-FR	0.009	0.327	102.092	0.012	0.018	0.042	0.012	0.320	0.106	0.067	0.665	0.219	0.236
RF-ClusterPCTs-FR	0.009	0.313	107.224	0.017	0.026	0.058	0.017	0.286	0.110	0.070	0.654	0.201	0.234
RF-AggSingle-CO	0.009	0.266	121.804	0.020	0.031	0.072	0.020	**0.206**	0.127	0.061	0.636	0.155	0.215
RF-AggComplete-CO	0.009	0.269	120.950	**0.021**	**0.032**	**0.074**	**0.021**	0.228	0.126	0.064	0.636	0.155	0.218
RF-BKmeans-CO	0.009	**0.343**	**97.858**	0.014	0.022	0.047	0.014	0.364	**0.101**	**0.075**	**0.674**	**0.227**	**0.245**
RF-ClusterPCTs-CO	0.009	0.301	106.638	0.017	0.027	0.062	0.017	0.264	0.109	0.066	0.654	0.186	0.224

Fig. 4. Results with the 13 performance measures for *Random Forest* from experiments performed on 8 different datasets. The best results obtained per measure per dataset are highlighted.

independent measures (*AUPRC, AUROC, wAUPRC* and *pooledAUPRC*) for the mediamill and tmc2007 datsets are improved for almost twice, which is consistent to the notion from the literature that ensembles of PCTs improve the performance over single predictive models. Hierarchies created with clustering of the space consisting of feature rankings outperform both hierarchies obtained using label co-occurrences and flat MLC for the threshold independent measures on the medical, enron and emotions datasets. *RF-BKmeans-FR* performs the best for medical dataset in seven evaluation measures. Considering the hierarchies obtained with clustering the space of label co-occurrences, we can note that they outperform the other methods for the corel5k dataset. Using hierarchies (i.e., label dependences) rather than flat multi-label task improves the predictive performance generally for most of the evaluation measures, but not for (*ML F1 measure, ML accuracy, ML precision* and *ML recall*) in the emotions and scene datasets.

Finally, in our study we also considered training errors i.e., the errors made in the learning phase. There, in a large majority of the cases, the original *FlatMLC* method performed the best. This means that other methods we use for constructing the hierarchies do not overfit as the original one. This is another advantage of methods for construction the hierarchies identified from the obtained results.

6 Conclusions and Further Work

In this work, we have presented an approach for hierarchy construction and structuring the output (label) space by using feature ranking. More specifically, we cluster the feature rankings to obtain a hierarchical representation of the potential relations existing among the different labels. We then address the task of MLC as a task of HMLC. Moreover, we compare our approach with the approach of clustering the space consisting of label co-occurrences [6].

We investigated four clustering methods for hierarchy creation, agglomerative clustering with single and complete linkage, balanced k-means and clustering using predictive clustering trees (PCTs). The resulting problem was then approached as a HMLC problem using PCTs and random forests of PCTs for HMLC. We used eight benchmark datasets to evaluate the performance.

The results reveal that the best methods for hierarchy construction are agglomerative clustering methods and balanced k-means. Compared to the original MLC method where there is no hierarchy this improves the performance in most of the datasets. In four datasets, the hierarchies obtained by clustering the label space consisting of feature rankings improve the predictive performance compared to the hierarchies obtained by clustering the space consisting of label co-occurrences. Similar conclusions, but to a lesser extent, can be made for the random forests of PCTs for HMLC - in many of the cases (datasets and evaluation measures) the predictive models exploiting the hierarchy of labels yield better predictive performance. Finally, by considering the training error performance, we find that original MLC models overfit more than the HMLC models.

For further work, we plan to make more extensive evaluation on more datasets with diverse properties and to try more different feature ranking methods. Furthermore, we assume that potential improvement of the performance can be achieved with cutting the hierarchies based on some conditions such as density, distribution or distance between nodes. Moreover, we plan to include a comparison to network approaches given by Szymanski et al. [11]. Finally, we plan to extend this approach to other tasks, such as multi-target regression.

Acknowledgment. We would like to acknowledge the support of the European Commission through the project MAESTRA - Learning from Massive, Incompletely annotated, and Structured Data (Grant number ICT-2013-612944), the project LANDMARK - Land management, assessment, research, knowledge base (H2020 Grant number 635201) and Teagasc Walsh Fellowship Programme.

References

1. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
2. Dimitrovski, I., Kocev, D., Loskovska, S., Džeroski, S.: Fast and scalable image retrieval using predictive clustering trees. In: International Conference on Discovery Science, pp. 33–48 (2013)
3. Huynh-Thu, V.A., Irrthum, Wehenkel, L., Geurts, P.: Inferring regulatory networks from expression data using tree-based methods. PLos One **5**(9) (2010)
4. Kocev, D.: Ensembles for predicting structured outputs. Ph.D. thesis, IPS Jožef Stefan, Ljubljana, Slovenia (2011)
5. Kocev, D., Vens, C., Struyf, J., Džeroski, S.: Tree ensembles for predicting structured outputs. Pattern Recogn. **46**(3), 817–833 (2013)
6. Madjarov, G., Dimitrovski, I., Gjorgjevikj, D., Džeroski, S.: Evaluation of different data-derived label hierarchies in multi-label classification. In: International Workshop on New Frontiers in Mining Complex Patterns, pp. 19–37 (2014)
7. Madjarov, G., Kocev, D., Gjorgjevikj, D., Džeroski, S.: An extensive experimental comparison of methods for multi-label learning. Pattern Recogn. **45**(9), 3084–3104 (2012)
8. Malinen, M.I., Fränti, P.: Balanced K-means for clustering. In: Fränti, P., Brown, G., Loog, M., Escolano, F., Pelillo, M. (eds.) S+SSPR 2014. LNCS, vol. 8621, pp. 32–41. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-662-44415-3_4
9. Silla, C.N., Freitas, A.: A survey of hierarchical classification across different application domains. Data Min. Knowl. Disc. **22**, 31–72 (2011)
10. Struyf, J., Džeroski, S.: Constraint based induction of multi-objective regression trees. In: Bonchi, F., Boulicaut, J.-F. (eds.) KDID 2005. LNCS, vol. 3933, pp. 222–233. Springer, Heidelberg (2006). https://doi.org/10.1007/11733492_13
11. Szymanski, P., Kajdanowicz, T., Kersting, K.: How is a data-driven approach better than random choice in label space division for multi-label classification? Entropy **18**, 282 (2016)
12. Tsoumakas, G., Katakis, I.: Multi label classification: an overview. Int. J. Data Warehouse Min. **3**(3), 1–13 (2007)
13. Tsoumakas, G., Katakis, I., Vlahavas, I.: Effective and efficient multilabel classification in domains with large number of labels. In: Proceedings of the ECML/PKDD Workshop on Mining Multidimensional Data, pp. 30–44 (2008)

14. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 667–685. Springer, Boston (2010). https://doi.org/10.1007/978-0-387-09823-4_34
15. Vens, C., Struyf, J., Schietgat, L., Džeroski, S., Blockeel, H.: Decision trees for hierarchical multi-label classification. Mach. Learn. **73**(2), 185–214 (2008)
16. Verikas, A., Gelzinis, A., Bacauskiene, M.: Mining data with random forests: a survey and results of new tests. Pattern Recogn. **44**(2), 330–349 (2011)

5.2 Structuring the Output Space in Multi-Target Regression Using Different Representations of the Targets

In this chapter, we present an extension of the classical methods for multi-target regression (MTR) that considers the interconnections between target attributes, i.e., discovers and exploits structure on the target space in the form of a hierarchy by using known hierarchical clustering methods.

As we described in the previous chapter, similar studies have proved to be successful for the task of multi-label classification (Madjarov et al., 2016; Nikoloski et al., 2018). Motivated by the improvements achieved for the MLC task, we adapt the methodology to the task of MTR. We then perform an extensive study on 16 different data sets from various domains, mostly from the environmental domain (8 out of 16), representative the task of multi-target regression.

We use two different representations of the targets to structure the output space: the *target representation*, consisting of the actual numerical values of the target variables, and the *feature rankings representation*, consisting of the feature importance scores for each target attributes. The modeling techniques we use are predictive clustering trees (PCTs) and ensembles (i.e., random forests) of PCTs for MTR. We compare the results of the original flat MTR models, hierarchical MTR (HMTR) models which consider data-derived hierarchies (from both the target and feature ranking representation) and HMTR models which consider the already known (i.e., expert created) hierarchies.

We use relative root mean square error (RRMSE) as an evaluation measure, as this is the most commonly used measure for the multi-target regression task. Depending on the size of the target space, we present two separate statistical evaluations, one for the bundle of datasets with small/medium target space ($<$ 100 target attributes) and another for two datasets with large target spaces ($>$ 100 target attributes). The two datasets in our collection with large target spaces both belong to the domain of environmental sciences. For datasets with small target spaces, we use the non-parametric Friedman test (M. Friedman, 1940) with the correction recommended by Iman and Davenport (1980), Nemenyi (1963), and for datasets with large target spaces, we use per-target performances.

Our results show that significant improvements in the performance can be achieved if the data-derived hierarchy of the target attributes is used, especially for datasets with large target spaces ($>$ 100 targets). Similar, but weaker conclusions can be made by using ensembles of PCTs, where structuring the output space does not improve the predictive performance significantly. The best structuring (i.e., clustering) methods are the divisive methods (balanced k-means and unsupervised PCTs).

Another significant finding from this study is that the data-derived hierarchies are a better choice than expert created hierarchies, which implies that we could obtain a good structure of the target space if we discover the knowledge from the data directly, rather than using the structure based on some domain expert pre-defined relations, i.e., hierarchies.

The paper included in this section is:

- NIKOLOSKI, Stevanche, KOCEV, Dragi, DŽEROSKI, Sašo. (2019), Data-driven structuring of the output space improves the performance of multi-target regressors. *IEEE Access*, 7:145177-145198, doi:10.1109/ACCESS.2019.2945084.

The contributions of Stevanche Nikoloski to this paper are as follows. SN contributed to the implementation/computer code for structuring the output space in MTR. He also participated in designing the experiments and carried out the experiments, as well as processed, evaluated and statistically compared their results. He drafted the paper and revised it based on co-author's and reviewer's feedback.

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 12, 2019, accepted September 29, 2019, date of publication October 2, 2019, date of current version October 17, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2945084

Data-Driven Structuring of the Output Space Improves the Performance of Multi-Target Regressors

STEVANCHE NIKOLOSKI[1,3], **DRAGI KOCEV**[1,2], **AND SAŠO DŽEROSKI**[1,2]

[1]Department of Knowledge Technologies, Jožef Stefan Institute, 1000 Ljubljana, Slovenia
[2]Jožef Stefan International Postgraduate School, 1000 Ljubljana, Slovenia
[3]Environment Soils and Land-use Department, Teagasc, Johnstown Castle, Y35 Y521 Ireland

Corresponding author: Stevanche Nikoloski (stevanche.nikoloski@ijs.si)

This work was supported in part by the European Commission through the project MAESTRA—Learning from Massive, Incompletely annotated, and Structured Data under Grant ICT-2013-612944, in part by the project LANDMARK—Land management, assessment, research, knowledge base (H2020) under Grant 635201, and in part by the Teagasc Walsh Fellowship Programme.

ABSTRACT The task of multi-target regression (MTR) is concerned with learning predictive models capable of predicting multiple target variables simultaneously. MTR has attracted an increasing attention within research community in recent years, yielding a variety of methods. The methods can be divided into two main groups: problem transformation and problem adaptation. The former transform a MTR problem into simpler (typically single target) problems and apply known approaches, while the latter adapt the learning methods to directly handle the multiple target variables and learn better models which simultaneously predict all of the targets. Studies have identified the latter group of methods as having competitive advantage over the former, probably due to the fact that it exploits the interrelations of the multiple targets. In the related task of multi-label classification, it has been recently shown that organizing the multiple labels into a hierarchical structure can improve predictive performance. In this paper, we investigate whether organizing the targets into a hierarchical structure can improve the performance for MTR problems. More precisely, we propose to structure the multiple target variables into a hierarchy of variables, thus translating the task of MTR into a task of hierarchical multi-target regression (HMTR). We use four data-driven methods for devising the hierarchical structure that cluster the real values of the targets or the feature importance scores with respect to the targets. The evaluation of the proposed methodology on 16 benchmark MTR datasets reveals that structuring the multiple target variables into a hierarchy improves the predictive performance of the corresponding MTR models. The results also show that data-driven methods produce hierarchies that can improve the predictive performance even more than expert constructed hierarchies. Finally, the improvement in predictive performance is more pronounced for the datasets with very large numbers (more than hundred) of targets.

INDEX TERMS Clustering, feature ranking, hierarchy, multi-target regression, target space.

I. INTRODUCTION

In supervised learning, the main goal is to learn, from a set of examples with known output (target) values, a function predicting the target value of a previously unseen example. The task where the examples refer to one target is called single target prediction and if the examples refer to more than one target is called multi-target prediction. In certain studies, the target components are considered independently and predictive models are built for each component separately. The overall prediction is then generated as a combination of all per-target predictions. In this way, the potential relations between the target components are not taken into account and the gap that is left with this is directly related with the quality of the obtained models.

Considering the t components of the output space, we can distinguish between single ($t = 1$) and multi-target prediction ($t > 1$). If the target space consists of

The associate editor coordinating the review of this manuscript and approving it for publication was Shagufta Henna.

continuous/numeric variables then the task at hand is multi-target regression (MTR). Likewise, if the target space consists of discrete/nominal variables then the task is called multi-target classification. The multi-label classification can be treated as a special case of multi-target classification [1]. Namely, multi-label classification (MLC) is the task of learning from data examples where each example can be associated with multiple labels, which belong to a predefined set of labels. The point of interest in our study is the multi-target regression task.

In many real life problems, for instance, in ecology (predicting the abundance of different species occupying the same habitat [2], estimating different vegetation quality indices for the same site [3] and predicting the composition of a community of organisms [4]), the target space is structured, meaning that there are some internal relations and dependences (e.g., hierarchical structure) among the targets. Finding those potential dependencies/relations is one of the most challenging problems in machine learning [5].

The methods for multi-target prediction can be categorized into two groups: (1) local methods (problem transformation methods), that create an individual model per target, and then combine the separate models in order to obtain an overall prediction and (2) global methods (known as big-bang methods or algorithm adaptation methods), that predict all targets at once [6], [7]. The main advantage of the global over the local methods is that the latter exploit the potential dependencies among the targets during the learning phase to obtain predictive models with better predictive performance.

A drawback of global models is that they ignore the local modularity in the connections among the target components such as parent-child, siblings relationships etc. In order to address this challenge, we focus on identifying some potential target relations by structuring the output space using a data-driven approach. Here, we approach the problem of structuring the output space by looking into two different spaces coupled with using different clustering approaches (balanced k-means, agglomerative and predictive clustering). First, we cluster the original output space that consists of the target values for each example. We then cluster the space consisting of the feature ranks for each component. At the end, we transform a flat multi-target regression problem into a hierarchical one using the hierarchy obtained by one of the cluster-based approaches. In other words, we translate the MTR task into a hierarchical multi-target regression (HMTR) task. The main research question is to investigate whether a predictive model learned on the transformed problem can achieve better predictive performance compared to a predictive model learned from the flat multi-target regression problem.

The predictive models that we use in the study are predictive clustering trees (PCTs). We selected PCTs since they are global models that can be used for different structured output prediction tasks (including MTR and HMTR) and they are constructed very efficiently. They are able to make a predictions for several types of structured outputs such as tuples of numerical/discrete values, time series, and hierarchies of variables. More details can be found in [8]–[13]. PCTs can be considered as a generalization of standard decision trees towards predicting structured outputs. But the change in just a few of the training examples can sometimes drastically change the structure of the tree. To improve their predictive performance, the predictive models can be combined into an ensemble [14]. An ensemble is a set of single (base) predictive models whose predictions are combined. For basic classification and regression tasks, it is widely accepted that ensemble learners improve the predictive performance of single tree learners [6].

More specifically, we use single PCTs and ensemble of PCTs for both MTR and HMTR setting. We perform an extensive empirical evaluation of the proposed methods on 16 MTR benchmark datasets. Most of the datasets (11 out of 16 datasets) are also used in [15]: The remaining datasets from [15] have small number of targets (2 or 3) and there is not much point in learning hierarchies in such small output spaces. For hierarchy creation, we use agglomerative clustering methods with single and complete linkage, balanced k-means, and predictive clustering trees (PCTs). In order to make our study more comprehensive, we perform experiments on two large datasets (with 111 and 492 targets) thus exploring the effect of including structures in large output spaces.

The results from the evaluation reveal that better predictive performance can be achieved by using data-driven approaches to construct the hierarchies rather than considering either, the flat multi-target regression task, or the pre-defined hierarchy created by a domain expert. Moreover, for large datasets, the results are in line with teh results for MLC [16], [17]: divisive hierarchy creation algorithms (balanced k-means and PCTs for clustering) are the best methods for clustering large output spaces. All in all, constructing a hierarchy of the target variables improves the predictive performance of the predictive models.

The reminder of this paper is organized as follows. In Section 2, we present the related work on the topic of multi-target regression and hierarchical multi-target regression. In Section 3, we show the data-driven approaches for structuring the target space and the space created from feature ranks of the targets for MTR. Furthermore, in this section we present the learning methodology used to create predictive models. Computational complexity is also discussed at this point. In Section 4, we present the experimental design, where we describe out data, point out the addressed experimental questions and instantiate the parameters used in our study, present the evaluation measures and the used statistical validation as well as the explanation on how the expert hierarchies are created for each data set. Experimental results are given and discussed in Section 5, while Section 6 concludes this paper.

II. BACKGROUND AND RELATED WORK

A. FORMAL DEFINITION OF MULTI-TARGET REGRESSION (MTR)

In our study, we focus on the task of multi-target regression that can be formally defined as follows [6], [18].

Given is:

- A description (input) space X covered by tuples of D independent descriptive instances (examples) i.e., $X = \{X_1, X_2, \ldots, X_D\}$;
- A target (output) space Y covered by tuples of T continuous target variables i.e., $Y = \{Y_1, Y_2, \ldots, Y_T\}$;
- Set of examples E consisting of a pairs of elements, one from input and another from output space, accordingly i.e., $E = \{(x_i, y_i)|x_i \in X, y_i \in Y, 1 \leq i \leq N\}$, where N is a number of examples;
- A quality criterion q, which selects and chooses the models with the lowest predictive error.

Find:

- A function $f : X \rightarrow Y$ which maximizes quality criterion q.

In our study, f is represented with predictive clustering trees (PCTs) or ensembles thereof.

B. METHODS FOR MULTI-TARGET REGRESSION

As mentioned above, we distinguish two groups of MTR methods: local (*problem transformation*) and global (*algorithm adaptation*) methods [6], [7], [19]. Local methods construct t separate models for the t target variables, which are combined to give the overall prediction for all the targets. From the other side, global methods build only one model for predicting all of the t target variables simultaneously. We next present the state-of-the-art MTR algorithms from both groups of methods.

1) LOCAL (PROBLEM TRANSFORMATION) METHODS

Since the local methods transform the problem into t separate single-target models, any known single target regression algorithm can be used to learn the single-target models. Prominent methods addressing the MTR task include: *ridge regression* [20], *support vector regression machines*, *regression trees* [14] and *stochastic gradient boosting* [21]. Reference [20] proposed a separate ridge regression algorithm that deals with MTR problems.

Regressor chain (RC) [22] is another problem transformation method motivated by the multi-label chain classifier [23]. During the training process, RC randomly selects a chain (permutation) of the target space, then builds a separate regression model for each target in consistence with the selected chain. Since RC uses the actual values of all previous targets in a chain, [22], also proposed regressor chain *corrected* (RCC) that uses cross-validation estimates instead of actual values. However, RC and RCC are sensitive to the selected chain ordering. In order to avoid this problem, [15], proposed an approach called ensemble of regressor chains (ERC) and ensemble of regression chains corrected (ERCC), where they randomly select as many models as the number of distinct label chain if the number of labels is less than 10. Otherwise, they randomly selected 10 chains and construct an ensemble of chains.

Multi-target regressor stacking (MTRS) [22] is another problem transformation method inspired by [24] where multi-label classification is performed by using stacked generalization. MTRS training is performed in two stages. First, t different single-target models are learned and then, instead of concatenating the t obtained predictions, MTRS includes additional training stage, where a second collection of t separate single target meta-models are learned. At the end, the predictions are calculated from both stages. The predictions from the second stage use and adjust the predictions from the first stage.

Zhang et al.(2012) [25], presented a new problem transformation method based on multi-output support vector regression approach. Basically, they extend the actual feature space and represent the multi-output problem as equivalent single-output problems, that are solved using the single-output least squares SVRs (LS-SVR) algorithm. The multi-output model takes into account the target correlations by using the vector virtualization method.

Recently, Wang *et al.* [26] propose a multi-target regression method (MTR-TSF) that embeds the intra-target relationships. First, by using hierarchical clustering on the output space, they reveal the correlation among the targets and create an additional feature vector X_{index} consisting of the indices of the nodes where specific instances belongs to. Next, they use a boosting regression algorithm to learn a similarity matrix for each target. Finally, by querying and clustering of the similarity matrix, a target specific feature vector X_{tsf} is created for all instances and is added to the original feature vector X. At the end, a prediction model per target is learned by considering the 'merged' feature space $X' = X \bigcup X_{index} \bigcup X_{tsf}$.

2) GLOBAL (ALGORITHM ADAPTATION) METHODS

Algorithm adaptation learns a single model for all target variables and thus take into account the dependencies among the targets. There are many advantages over the local methods such as interpretability, better predictive performances, especially, if the targets are related [6]. Below, we briefly discuss various algorithm adaptation methods proposed in the literature.

First attempt to deal with prediction of multiple target variables are the statistical methods such as *reduced-rank regression* [27]. Furthermore, [28] proposed the general version of a multivariate regression problem of the James-Stein estimator, called as *filtered canonical y-variate regression*. Next, lasso regression [29] is a popular regression method for estimation in linear models. It produces interpretable models while at the same time it is stable. Next, gaussian process for MTR are based on the algorithm proposed by [30]. The most prominent statistical approach that deals with multiple targets is the *curds and whey (C & W)* method [31].

Predictive clustering trees (PCTs) are tree-based methods built within the predictive clustering framework [8]. This framework learns decision trees called predictive clustering trees (PCTs) where the top node contains all of the training examples and then it recursively splits into lower partitions (clusters) of the whole train set. PCTs can be used for classical machine learning tasks (clustering, classification and regression), but also, can be applied to multi-target prediction. PCTs can deal with structured outputs prediction, such as vectors, time series, sequences or hierarchies [9]–[13].

In addition, [32] presented an algorithm called multi-target step-wise model tree induction (MTSMOTI) for generation a multi-target model tree on a step-wise manner. The tree model is generated similarly as in PCTs, with TDIDT algorithm. The difference is that each leaf in a tree model is associated with a set of linear models which generate the final target predictions. Conditional Inference Trees (CTrees) are non-parametric regression trees embedding tree-structured regression models into conditional inference procedures and estimate a regression relationship in a multi-target scenario [33].

A different type of MTR algorithm is the rule based algorithm called *FItted Rule Ensemble (FIRE)* method, proposed by [34]. This is a method for learning rule ensembles based on representing an ensemble of regression trees as a large collection of rules. FIRE uses an optimization procedure (minimization) to select the best (much smaller) set of rules and determine their respective weights.

Furthermore, Breskvar *et al.* [35] present an ensemble method with random output selection (ROS). Instead of using all target attributes, they randomly select subsets of target attributes when learning the base predictive models of the ensemble. This additional randomization improves the performance both in terms of time complexity and predictive accuracy.

The most famous non-parametric distance-based method for regression task is the k-nearest neighbour method. It takes the average of the values of the k nearest examples as a prediction. K-nearest neighbour is a flexible algorithm, since it can use any distance function and any number k (nearest neighbours) [36].

Multiple-input multiple-output (MIMO) support vector regression method is a generalization of support vector machines (SVMs) for addressing the MTR task. The generalization is achieved by minimization of a Lagrangian equation which has multi-dimensional parameters that have to be optimized [37], [38].

Partial Least Squares Regression (PLS-PLSR) and Principal Component Regression (PLS-PCR) methods are another methods for multi-target regression which are implemented in the R software package *pls* [39]. These methods are commonly used in many natural sciences and are based on calculation of the scores obtained by decomposition of the product matrix of orthogonal scores and loadings. Then regression coefficients are calculated using the scores.

Multivariate Adaptive Regression Splines (MARS) is a non-parametric regression method implemented in *EARTH* package in R. MARS, as a generalization of step-wise linear regression [40] constructs the dependencies between input and output variables by using a data-driven set of base functions and coefficients.

Another well-known and widely used method for MTR are the artificial neural networks (NN). They are designed based on human brain to recognize patterns in data. They can automatically model the nonlinearity and can deal with multi-input multi-output problems. The most often used algorithm for training artificial neural networks is *backpropagation* algorithm [41]. Backpropagation algorithm is recursive and iterative method which efficiently optimize the network weights by following the gradient descent method that exploits the chain rule. *Deep neural networks* (DNN) are artificial neural networks containing multiple hidden layers. It update the network weights by establishing the correlation between input (past events) and output (future events). There are several variants of DNNs designed based on the specific domains that are used for. Convolutional deep neural networks (CNNs) are used in the domain of computer vision. Recurrent neural networks (RNNs) are used in various cases of language modelling, such as handwriting and speech recognition [42], [43]. Zhen *et al.* [44] present a deep learning approach for considering the intra-target dependencies. Namely, they propose a multi-layer multi-target regression (MMR) method where intra-target dependencies are explicitly encoded by using matrix elastic nets (MEN) to create the structure of the target space (structure matrix S), which enables learning of the target correlations by minimization of the $rank(S)$. Then, the kernel trick is used in order to solve the problem of non-linearity in the representation of the target dependencies.

C. FORMAL DEFINITION OF HIERARCHICAL MULTI-TARGET REGRESSION (HMTR)

We follow similar guidelines as for defining the task of MTR to formally define the task of hierarchical multi-target regression [13]:

Given is:

- A description (input) space X covered by tuples of D independent descriptive instances (examples) i.e., $X = \{X_1, X_2, \ldots, X_D\}$;
- A target (output) space $\mathcal{Z}$ covered by tuples of T continuous target variables i.e., $\mathcal{Z} = \{\mathcal{Z}_1, \mathcal{Z}_2, \ldots, \mathcal{Z}_T\}$. We define a hierarchy $\mathcal{H} = (\mathcal{Z}, \leq_p)$ for the variables from the output space $\mathcal{Z}$. The relation " $\leq_p''$ represents a parent-child relationship between tree nodes $(\forall(\mathcal{Z}_1, \mathcal{Z}_2) \in \mathcal{H} : \mathcal{Z}_1 \leq_p \mathcal{Z}_2$ if and only if $\mathcal{Z}_2$ is a parent (meta-label) of $\mathcal{Z}_1$) and is called *hierarchical constraint*. The meta-labels are result of an aggregation function (for example, sum or average) on their respective children i.e $\mathcal{Z}_k = agg\{\mathcal{Z}_i | \mathcal{Z}_i \leq_p \mathcal{Z}_k\}$;
- Set of examples E consisting of pairs of elements, one from input and another from output space, accordingly

i.e., $E = \{(x_i, y_i)|x_i \in X, z_i \in \mathcal{Z}, 1 \leq i \leq N\}$, where N is a number of examples and where the values of the target variables satisfy the hierarchical constraint " $\leq_p''$ i.e $\forall j : \exists i(\mathcal{Z}_i \leq_p \mathcal{Z}_j \Longrightarrow z_j = agg\{z_i|\mathcal{Z}_i \leq_p \mathcal{Z}_j\})$;

- A quality criterion q, which selects and chooses the models with the lowest predictive error and the highest accuracy.

Find:

- a function $f : X \rightarrow Y$ which maximizes the quality criterion q and all predictions $\hat{z} = f(x)$ are satisfying the hierarchical constraint.

The difference to the task of MTR is in the definition of the output space: for HMTR we have a set of numeric variables organized in a hierarchy instead of a flat tuple of numeric variables. The definition of the parent-child relationships (hierarchy constraint) states that the variable belonging to a given hierarchy node automatically contributes to all its parent nodes.

D. METHODS FOR HIERARCHICAL MULTI-TARGET REGRESSION

In this part, we present the existing (state-of-the-art) methods, related to the task of hierarchical multi-target regression. To begin with, *multilevel analysis* refers broadly to the methodology of research and data structures that deal with nested data, i.e., including more than one type of unit. This is directly related with involving several levels of aggregation. Consider an example from educational research, where students from different schools are considered, and their performance (e.g., grades) is being predicted.

Then, a separate regression model can be fitted within each school, and the model parameters from these schools can be modeled as depending on each school properties (such as the socioeconomic status of the schoolâĂŹs neighbourhood, whether the school is public or private, and so on). The student-level regression and the school-level regression here are the two levels of a multilevel model. The lowest level is the student-level and each student belonging to this level can be linked with appropriate class, and then each class to appropriate school and so on. With this, a kind of dependency levels (i.e., a hierarchy) is created. Moreover, in the higher levels in the multilevel model, regression parameters (hyperparameters) can be fitted for the regression model. That is the reason why in most of the research, the term "multilevel analysis" is mostly used interchangeably with "hierarchical linear modeling", although strictly speaking they are distinct.

Another application of the hierarchical linear modeling approach can be found in [45], where a two-level hierarchical linear model with multiple outputs was employed to analyze an information obtained from two different groups of informants (child and parents participants) in order to assess the demographic risk factors on children's exposure to violence (ETV) and how these effects vary by informants.

The main advantage of multilevel modeling is spreading of a residual components through each level of a hierarchy, thus the overall variance is partitioned and moreover, the predictors are included at each level. Hence, with application of multi-target regression at each level, the model can deal with between-level relations in the hierarchy. Latter makes multilevel modelling superior than regression modeling with respect to the model performance [46]. An extensive review for multilevel modeling is given by [47] and [48].

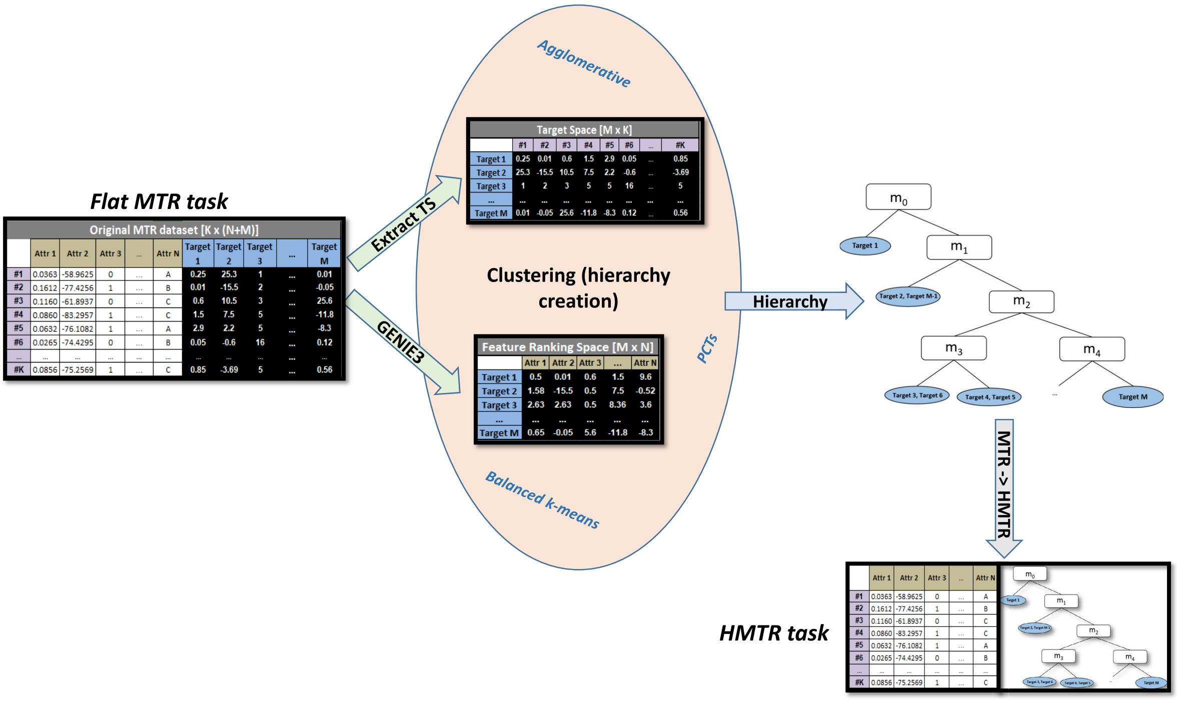
Next, online analytical processing (OLAP) is a method which allows to extract and analyze data from multiple sources at the same time. The data is multidimensional, hence the extracted information can be compared in different ways. For example, a book store might compare their book sales in September with sales in August, then compare those results with the sales from another location, which might be stored in a different database. The OLAP data is stored in multidimensional databases and all attributes are considered as a separate dimension. Considering the multi-dimensionality, the OLAP data is structured in a hierarchical form by using some of the OLAP tools: consolidation (roll-up), drill-down, and slicing and dicing [49]. This structuring and hierarchical representation enables a complex calculations and manipulation of the data (trend analysis, data modeling) [50]. The natural relationships in the data by using OLAP method are also researched by [51] by using a partially ordered set of levels (dimension schema).

Predictive clustering trees (PCTs) for HMTR task is proposed recently by [13]. The original PCTs for MTR are extended to HMTR task with defining prototype function and variance function. All operations for aggregation can be used as a prototype functions, but keeping in mind that with some of them (for example, *minimum* or *maximum*) after averaging, the hierarchical constraint (parent-child relation within the hierarchy) can be violated. For the variance function, the weighted Euclidean distance is used where the weights are defined such that they decrease exponentially with the depth of the node in the hierarchy.

E. METHODS FOR STRUCTURING THE OUTPUT SPACE

The main goal in this article is structuring the output space in MTR. To the best of our knowledge, structuring of the target space for MTR has not been explored yet. Hence, we overview the methods for structuring the output space for the related multi-label classification (MLC) task where learning hierarchies in the output space has been studied to a wider extent [16], [17], [52]–[55].

Joly et al. (2014) [52] propose a method for dimensionaltiy reduction of the output space by random projections of it, mainly focused on MLC task. The projections are made in such a way that preserve distances in projected space. The reduction of the variance function is made on the projected space, while the predictions are made directly in the original output space using a decoding procedure. Similarly, Joly et al. (2017) [56], proposes a gradient boosting method for MTR which automatically adapt the target correlations by random projection of the output space.



Flat MTR task

Original MTR dataset [K x (N+M)]										
	Attr 1	Attr 2	Attr 3	...	Attr N	Target 1	Target 2	Target 3	...	Target M
#1	0.0363	-58.9625	0	...	A	0.25	25.3	1	...	0.01
#2	0.1612	-77.4256	1	...	B	0.01	-15.5	2	...	-0.05
#3	0.1160	-61.8937	0	...	C	0.6	10.5	3	...	25.6
#4	0.0860	-83.2957	1	...	C	1.5	7.5	5	...	-11.8
#5	0.0632	-76.1082	1	...	A	2.9	2.2	5	...	-8.3
#6	0.0265	-74.4295	0	...	B	0.05	-0.6	16	...	0.12
...	...	...	...	...	...	...	...	...	...	...
#K	0.0856	-75.2569	1	...	C	0.85	-3.69	5	...	0.56

Target Space [M x K]								
	#1	#2	#3	#4	#5	#6	...	#K
Target 1	0.25	0.01	0.6	1.5	2.9	0.05	...	0.85
Target 2	25.3	-15.5	10.5	7.5	2.2	-0.6	...	-3.69
Target 3	1	2	3	5	5	16	...	5
...	...	...	...	...	...	...	...	...
Target M	0.01	-0.05	25.6	-11.8	-8.3	0.12	...	0.56

Clustering (hierarchy creation)

Feature Ranking Space [M x N]					
	Attr 1	Attr 2	Attr 3	...	Attr N
Target 1	0.5	0.01	0.6	1.5	9.6
Target 2	1.58	-15.5	0.5	7.5	-0.52
Target 3	2.63	2.63	0.5	8.36	3.6
...	...	...	...	...	...
Target M	0.65	-0.05	5.6	-11.8	-8.3

HMTR task

	Attr 1	Attr 2	Attr 3	...	Attr N
#1	0.0363	-58.9625	0	...	A
#2	0.1612	-77.4256	1	...	B
#3	0.1160	-61.8937	0	...	C
#4	0.0860	-83.2957	1	...	C
#5	0.0632	-76.1082	1	...	A
#6	0.0265	-74.4295	0	...	B
...	...	...	...	...	...
#K	0.0856	-75.2569	1	...	C

FIGURE 1. An illustration of the proposed framework for structuring the output space. We consider two spaces i.e., representations to cluster the targets: the original target space (TS) i.e., the values of a given target for each example and feature ranking space (FR) i.e., the importance scores for each feature with respect to a given target, for transforming the original MTR task to a HMTR task.

Madjarov et al. (2016) [16] present a comprehensive study of different data-derived methods for structuring the label space in the form of hierarchies for MLC. Namely, they use the label co-occurrence matrix to obtain a hierarchy of labels by using several clustering algorithms such as: agglomerative clustering with single and complete linkage, balanced k-means and PCTs. Their results say that divisive clustering methods (balanced k-means and PCTs) perform the best.

Tsoumakas et al. (2007) [55] propose a transformation-based ensemble method for random k-labelsets (RAkEL) for MLC by using existing algorithms for MLC. The RAkEL algorithm creates an ensemble by random sampling a small subset with k labels for each base model. The sampled subsets are structured as a label powerset and multi-class classifier is then used.

Next, Szymanski et al. (2016) [54] present a study which addressed to the question, whether data-driven methods on a graph consisting of label co-occurrences is significantly better than random generated graph of labels for MLC. This method is actually data-driven version of RAkEL method. Their results show that in general data-driven approach is superior to random created graphs of labels.

Nikoloski et al. (2017) [17] propose an algorithm for structuring the output space using feature ranking in MLC. They create a hierarchy from a space constructed by feature rankings for each of the classes. Furthermore, they present a comparative analysis with the approach from [16], where hierarchy is created by clustering the space consisting of label co-occurrences. In both cases, it is shown that some improvements in predictive performance can be achieved if data-driven approach for output space structuring is used, compared to using a flat multi-label classification task, despite the higher complexity added by additional procedure for calculating the feature importance and the clustering procedures.

III. STRUCTURING THE OUTPUT SPACE FOR MTR

The idea for structuring the output space in MLC proposed by [17] and [16] motivates the exploitation of methods for structuring the output space in MTR. In this study, we propose to transform a flat MTR task into a task of hierarchical multi-target regression (HMTR) [13]. Namely, we use the hierarchies created with data-driven clustering approaches to investigate whether the predictive models obtained with the HMTR task yield better predictive performance than predictive models obtained with the flat MTR task.

A. STRUCTURING THE TARGET SPACE

In our paper, we propose a framework that transforms the original multi-target regression (MTR) task into a hierarchical multi-target regression (HMTR) task, by clustering the output space. The flowchart of the framework is given in Figure 1.

The method for structuring the target space is outlined in the procedure *StructuringTargetSpace* from Algorithm 1. First, we take the original training dataset F^{train} and extract the target space W^{train} from the complete dataset. To obtain a hierarchy, we cluster the space W^{train} by using the procedure *Clustering* (it can use any arbitrary algorithm for clustering). With the function *TransformData*, we transform the original datasets F^{train} and F^{test} to new datasets F_H^{train} and F_H^{test} by including the obtained hierarchy and then, we learn a predictive model and generate the predictions. Next, we calculate the predictions for each node in the hierarchy and extract only the predictions related to the targets, which are in the hierarchy leafs. Finally, using those predictions, we evaluate the predictive performance.

Algorithm 1 The Algorithm for Structuring the Target Space

```
procedure StructuringTargetSpace($F^{train}$, $F^{test}$)
Input: $F^{train}$ - training dataset
Input: $F^{test}$ - test dataset
Output: $Performance$
  1: $W^{train}$ = ExtractTargetSpace($F^{train}$);
  2: $hierarchy = Clustering(W^{train})$;
  3: $F_H^{train} = TransformData(F^{train}, hierarchy)$;
  4: $F_H^{test} = TransformData(F^{test}, hierarchy)$;
  5: $HMTR\_Model$ = HMTRMethod($F_H^{train}$);
  6: $predictions$ = CalculatePredictions($HMTR\_Model$, $F_H^{test}$);
  7: $P$ = ExtractLeafsPredictions($predictions$);
  8: $Performance$ = Evaluate($P$);
  9: return $Performance$
```

B. STRUCTURING THE SPACE OF FEATURE RANKS OF THE TARGETS

The method for structuring the feature importance scores of the targets is outlined in procedure *StructuringFRSpace* from Algorithm 2. First, we take the original training dataset F^{train} and by using an arbitrary feature ranking approach (function *CreateFimp*), we create feature importance scores for each target separately. Then, the F^{ranks} dataset is constructed from the feature importance scores. Next, we obtain a hierarchy with clustering the F^{ranks} space, using an arbitrary clustering algorithm. Same as the previous Algorithm 1, we transform the original datasets F^{train} and F^{test} to new datasets F_H^{train} and F_H^{test} by including the obtained hierarchy and then, we learn a predictive model, generate the predictions and evaluate the predictive performance.

From the abovementioned procedures for structuring the output space, we can notice that in the procedure *StructuringFRSpace* (Algorithm 2), there is an additional step, compared to the procedure *StructuringTargetSpace* (Algorithm 1). The additional step is the function *CreateFimp* at line 1 (Algorithm 2), which increases the theoretical complexity of the algorithm *StructuringFRSpace*.

Algorithm 2 The Algorithm for Structuring the Target Space Using Feature Importance Scores per Target

```
procedure StructuringFRSpace($F^{train}$, $F^{test}$)
Input: $F^{train}$ - training dataset
Input: $F^{test}$ - test dataset
Output: $Performance$
  1: $FimpPath$ = CreateFimp($F^{train}$);
  2: $F^{ranks}$ = CreateArffFromFimp($FimpPath$);
  3: $hierarchy$ = Clustering($F^{ranks}$);
  4: $F_H^{train} = TransformData(F^{train}, hierarchy)$;
  5: $F_H^{test} = TransformData(F^{test}, hierarchy)$;
  6: $HMTR\_Model$ = HMTRMethod($F_H^{train}$);
  7: $predictions$ = CalculatePredictions($HMTR\_Model$, $F_H^{test}$);
  8: P = ExtractLeafsPredictions($predictions$);
  9: $Performance$ = Evaluate($P$);
 10: return $Performance$
```

Next, we describe the feature ranking approach for calculating the importance of the descriptive variables. Random forests are constructed by using the algorithm for learning PCTs in CLUS, modified according to the original random forest method proposed by [57]. Their use as feature ranking methods has been well studies in the literature (cf. [58]). First, random forests perform bootstrap sampling on the data and then build a decision tree for each bootstrap sample. Next, at each node of the tree, the best test is taken from a randomly selected feature subset.

Huynh-Thu et al. (2010) [59], proposed the GENIE3 algorithm for feature ranking. It uses reduction of the variance (of the target variables) at each node in the tree. The algorithm is checking which of the input variables reduce the variance more, and then, those which reduce more, are more important. Consequently, the ones which reduce the variance less, are less important. For each selected descriptive variable as a splitting variable, the produced reduction of the variance is being measured. The importance will be 0 if the descriptive variable is never been selected as a splitting variable (for any tree in the ensemble), meaning that it was not deemed important enough. The GENIE3 algorithm has been vastly evaluated for single-target regression tasks, for instance, in the domains of gene reconstruction. The random forest algorithm used for feature ranking is adapted with the idea proposed in the GENIE3 algorithm. For building the ensemble, the random forests of PCTs are used. The outcome is a feature ranking algorithm which is adapted to be used for various types of tasks for structure output prediction [60].

C. HIERARCHY CREATION (CLUSTERING) ALGORITHMS

In this part, we overview the clustering methods used to create the hierarchies of the target space. For achieving a good performance of the HMTR methods, it is necessary to construct target hierarchies that are capturing the relations (dependencies) among the target attributes. The main constraint in hierarchy creation is that the original MTR task should be defined by the leafs of the hierarchy. Specifically, each leaf in the hierarchy represents a set of targets from the original MTR problem. At the end, the number of targets in the hierarchy leafs must be the same as the number of targets from the original MTR problem. Furthermore, the internal nodes of the hierarchy (so called meta-labels) represent the potential relations among the original targets.

For creating the hierarchies, we use four different clustering methods (two divisive and two agglomerative):

- balanced k-means clustering (*divisive*);
- predictive clustering trees (*divisive*);
- agglomerative clustering with complete linkage and
- agglomerative clustering with single linkage.

Agglomerative clustering algorithms are bottom-up algorithms for clustering, where in the first iteration, each example is consider as a separate cluster. In the next iterations, the pairs of clusters are merged based on their linkage (distance metric). There are several possibilities for linkage of the examples. Namely, if the maximal distance of two examples from the clusters C_1 and C_2 is used, then this type of linkage is called *complete* linkage, i.e., $\max\{dist(c_1, c_2) : c_1 \in C_1, c_2 \in C_2\}$. Then, if the minimal distance between two examples for two different clusters is used, then we have an agglomerative clustering with *single* linkage i.e., $\min\{dist(c_1, c_2) : c_1 \in C_1, c_2 \in C_2\}$.

Balanced k-means is divisive top-down approach for clustering. First, root node of the hierarchy represents the one common cluster, consisting of all targets from the target space $\mathcal{T}$. Then, consecutively, this cluster is divided into k disjoint sub-clusters (meta-labels) ($k < |\mathcal{T}|$) using the k-means clustering algorithm. The number of cluster divisions k is an input to this algorithm, hence the algorithm output clusters with approximately equal size [61]. The procedure recursively is repeated on each sub-cluster until the number t of targets in each sub-cluster is smaller than $k - 1$. In other words, our target space $\mathcal{T}$ is covered by leafs of the hierarchy obtained by the balanced k-means clustering approach.

We also use predictive clustering trees (PCTs), which can be used as another divisive hierarchical clustering method, to build up the target hierarchies. More specifically, we treat the target space as descriptive space. Descriptive and target variables, all together, are used to provide descriptions for the obtained clusters. To calculate the heuristic score, a variance function is used during the learning process until some stopping criterion is met. This means that there is no need for using predefined number of clusters, as required by traditional clustering methods. The focus of using PCTs for clustering is on using predictive clustering framework in unsupervised manner i.e., on the task of clustering instead of predictive modelling [62], [63].

D. LEARNING METHODOLOGIES

1) PREDICTIVE CLUSTERING TREES (PCTS)

The PCT framework views a decision tree as a hierarchy of clusters, where the top-node corresponds to one cluster containing all the data. While moving downwards the tree, this top-cluster is sub-divided into smaller clusters recursively. The PCT framework is implemented in the CLUS software package (https://sourceforge.net/projects/clus/) [6], [9].

PCTs are obtained with a standard top-down induction of decision trees (TDIDT) algorithm [64]. As an input, TDIDT takes a set of examples to produce a tree as an output. By using a heuristic function, computed on the training instances, the TDIDT procedure selects a test for the root node. The heuristic aims to select a test which maximizes the variance reduction caused by the partitioning of the examples into subsets according to the test outcome. Recursive procedure of partitioning the examples continues until a stopping criterion is satisfied. Further partitioning of examples yields a tree with a lower quality. In this case, we store the prediction (output value of a prototype function) in the corresponding leaf of the tree.

Blockeel (1998) [8], proposed the predictive clustering framework, while predictive clustering trees (PCTs) for multi- target regression (MTR) were proposed by [9]. In PCTs for MTR, the prototype function calculates the mean vector of all target variables Y for the training examples that belong to the leaf. In the prediction phase, for each new example, the algorithm identifies the leaf it belongs to and returns the value predicted by the prototype function associated to that leaf. The PCTs can be instantiated for a specific given learning task by considering specific variance (for split selection) and prototype function (for calculating the predictions in each leaf). Actually, that is the main difference with standard decision tree learning.

The PCTs are developed to work for the task of multi-target regression (MTR) [65], multi-label classification (MLC) [66], prediction of time series [12], hierarchical multi-label classification (HMLC) [11] and recently, for hierarchical multi-target regression (HMTR) [13]. We will now describe how PCTs from hierarchical multi-target regression are build. In order to extend the PCTs for the HMTR task, we need to define variance and prototype functions.

The variance is calculated by applying a distance function on the values of the variables in analogy of the distances for HMLC and the implementation of that task, i.e., the variance is calculated as the average squared distance between each node Π_i of the examples and the mean node vector $\overline{\Pi}$:

$$Var(E) = \frac{1}{|E|} \cdot \sum_{i=1}^{|E|} dist(\Pi_i, \overline{\Pi})^2 \quad (1)$$

where $\overline{\Pi} = \frac{1}{|E|} \cdot \sum_{i=1}^{|E|} \Pi_i$.

Any distance d can be used as a distance function in Eq (1). [13] proposes for the task of HMTR to use a weighted Euclidean distance:

$$dist(\Pi_1, \Pi_2) = \sqrt{\sum_{s=1}^{|\Pi|} \theta(c_s) \cdot (\Pi_{1,s} - \Pi_{2,s})^2} \quad (2)$$

where $\Pi_{i,s}$ is the s'th component of the class vector Π_i of the instance E_i, $|\Pi|$ is the size of the class vector, and the class weights $\theta(c) = \theta_0^{depth(c)}$. The class weights $\theta(c)$ decrease exponentially with the depth of the node in the hierarchy thus making the differences in the lower parts of the hierarchy less influential to the overall score.

The prototype function used is averaging the values of the examples belong to a given leaf.

2) RANDOM FORESTS OF PCTS

Random forests of PCTs are implemented in the CLUS system [6] following the same method as for the simpler tasks of classification and regreesion [57]. A random forest represents an ensemble of trees where the diversity among the trees is achieved by bootstrap replicates and for each tree node in the learning phase, a randomly selected subset of descriptive attributes is considered for split selection. Bootstrap replicates are generated by random sampling of instances from the training set, with replacements, until the same number of instances as in the original training set is reached.

The difference between the PCT procedure for tree construction in random forest algorithm and the standard PCT procedure is in the selection of descriptive attributes. In the former, selection of the descriptive attributes is randomized. Namely, at each node in the decision tree, a random subset of attributes is taken from the descriptive space and the best attribute is chosen from this subset. There are different ways of retaining the number of attributes from descriptive space. The number of attributes that are chosen from descriptive space is given by function f of the total number of descriptive attributes D (e.g. $f(D) = 1, f(D) = [\sqrt{D} + 1], f(D) = [\log_2 D + 1]$ etc.). This randomness is chosen in order to avoid the correlation between the bootstrap samples. For example, if there are only few relevant descriptive attributes that are important for prediction of the target variables, these will be selected many times in the bootstrap replicates, hence providing more correlated trees.

Prediction of new instances in random forest algorithm for PCTs are made by combining the prediction of all base predictive models. For both MTR and HMTR, the prediction of each target is defined as an average of the predictions obtained from each predictive tree.

E. COMPUTATIONAL COMPLEXITY

1) SINGLE PCTS FOR MTR/HMTR

In this part, we analyze the computational complexity of PCTs for HMTR and compare it with the computational complexity of PCTs for MTR. We discuss the order of complexity for both single PCTs and ensembles of PCTs for HMTR. Let us assume that the size of the training set, i.e., the number of examples, is e, the number of descriptive attributes is d out of which c are continuous, the number of target attributes is t and the number of meta-labels is m.

The top-down induction algorithm of PCTs requires sorting of the the c numeric attributes, and it has a cost of $\mathcal{O}(c \cdot e \cdot \log e)$ and $c = \mathcal{O}(d)$. Calculating the best split for multiple variables has the complexity order of $\mathcal{O}(t \cdot d \cdot e)$ and applying the split to the examples has a linear complexity, i.e., $\mathcal{O}(e)$. We assume that the tree is balanced, which means that the depth of the tree is $\log e$. With these calculations, the computational cost of inducing a single MTR tree is:

$$\mathcal{O}(MTRtree) = \mathcal{O}(d \cdot e \log^2 e + t \cdot d \cdot e \cdot \log e + e \cdot \log e) \tag{3}$$

For the HMTR algorithm, we also have the meta-labes (intermediate nodes), which in this case act like targets. This affects the computational cost only when the best split is calculated. More specifically, this costs is given by $\mathcal{O}((t + m) \cdot d \cdot e \log e)$ compare to the $\mathcal{O}(t \cdot d \cdot e \log e)$ for PCTs for MTR. Given this, we can calculate the order of complexity for a HMTR tree, which is very similar to the one for a MTR tree:

$$\mathcal{O}(HMTRtree) = \mathcal{O}(d \cdot e \log^2 e + (t + m) \cdot d \cdot e \cdot \log e + e \cdot \log e) \tag{4}$$

2) RANDOM FOREST OF PCTS

The order of complexity of constructing ensembles depends on the complexity of the base predictive models and their number b. The random forest performs sampling of the instances and sampling of the features. This random sampling reduces the computational complexity of the ensemble and is lower than the intuitive $\mathcal{O}(b \cdot MTRtree)$. Let the number of examples used to train the base predictive model with sampling of the examples be e' and the number of descriptive attributes considered in random forests d', where $e' < e$ and $d' < d$. The computational complexity of the creation of the bootstrap replicates of the training set for random forests is $\mathcal{O}(e)$ and the complexity of the random sampling of the features at each node for random forests is $\mathcal{O}(d' \cdot \log e')$.

Hence, the computational costs random forest PCT ensembles for MTR is the following:

$$\mathcal{O}(Rforest\_MTR) = \mathcal{O}(b \cdot d' \cdot e' \log^2 e' + b \cdot t \cdot d' \cdot e' \cdot \log e' + b \cdot e' \cdot \log e' + b \cdot e + b \cdot d' \cdot \log e') \tag{5}$$

The computational complexity of the HMTR counterparts of the random forest PCT ensembles for HMTR is the following:

$$\mathcal{O}(Rforest\_HMTR) = \mathcal{O}(b \cdot d' \cdot e' \log^2 e' + b \cdot (t + m) \cdot d' \cdot e' \cdot \log e' + b \cdot e' \cdot \log e' + b \cdot e + b \cdot d' \cdot \log e') \tag{6}$$

In Eq.(6) we can see a linear increasing in complexity with respect to targets with introducing the meta-labels (intermediate nodes). The same translation we already considered for the single PCTs for HMTR (see Eq. (4)).

For all methods (PCTs and ensembles of PCTs for both MTR and HMTR), from their complexity cost, we can see that the dominant elements in the equations are the one containing the second logarithmic power of the number of examples, and the one that is multiplied with the number of targets. For single PCTs, the first element is $\mathcal{O}(d \cdot e \log^2 e)$, and the second is $\mathcal{O}(t \cdot d \cdot e \cdot \log e)$ or $\mathcal{O}((t+m) \cdot d \cdot e \cdot \log e)$ for MTR and HMTR, respectively. If we compare the two terms, we can see that the first term is greater than the second when $\log e > t$ for MTR and $\log e > t + m$ for HMTR. Let us explore the first case where the first term is smaller. This means that when comparing MTR and HMTR, HMTR will have higher computational cost, due to the addition of m. Let us now explore the second case where $\log e$ is higher. In this case, the computational cost is affected only with the first term, hence the linear increase in the second term (i.e., (i.e., the addition of s in $\mathcal{O}((t+m) \cdot d \cdot e \cdot \log e)$) will be insignificant, resulting in comparable performance between MTR and HMTR for all methods on a datasets with a sufficiently large number of examples.

3) ALGORITHMS FOR STRUCTURING THE OUTPUT SPACE

We discuss the computational complexity of the procedures for structuring the output space given in Algorithm 1 and Algorithm 2. In the procedure for structuring the feature ranking space, there is an additional function *CreateFimp* for calculating the feature importance for each target. Since it is done by random forest method with GENIE3, the order of complexity of this procedure is $\mathcal{O}(\mathit{CreateFimp}) \approx \mathcal{O}(\mathit{Rforest\_MTR})$.

The most important cost for the clustering procedure is the number of examples e – in the case of datasets with large number of examples, the clustering procedure will take more time to create the hierarchy. When balanced k-means is used as a clustering procedure, the time complexity will be $\mathcal{O}(e{\cdot}t^3)$. Moreover, if the agglomerative clustering methods are used, the time complexity will be $\mathcal{O}(e \cdot t^3)$ and memory consumption $\mathcal{O}(e \cdot t^2)$, which makes it too slow for even medium data sets. Time complexity of PCTs used as a clustering method is the same as in Eq (3).

Alternatively, the procedure for creating the hierarchy (*Clustering* at line 5 in procedure in Algorithm 2) using feature rankings has a dimension which depends of the cardinality of the feature space F^{ranks}, denoted as d. The feature space cardinality is typically much smaller than the number of examples (i.e., $|F^{ranks}| \ll |W^{train}|$, i.e., $d \ll e$), meaning that clustering of the rankings will finish faster than clustering of the original target space. Using balanced k-means, it will be $\mathcal{O}(d \cdot t^3)$, where $d \ll e$, then, by using agglomerative it will be $\mathcal{O}(d \cdot t^3)$, and memory consumption $\mathcal{O}(d \cdot t^2)$, where $d \ll e$. Finally, the time complexity of PCTs algorithm used for clustering will be the same as in Eq (3) when we cluster the feature rankings space, considering that $d \ll e$. All in all, the clustering procedure is much more efficient when feature ranking space is considered, since the number of features and number of targets, in most of our datasets are significantly smaller than number of instances.

IV. EXPERIMENTAL DESIGN

A. EXPERIMENTAL QUESTIONS

We set the experimental design focusing on the following research questions:

(1.) Does structuring the output space (using a hierarchies) improves the predictive performance compared to the original flat MTR task?

(2.) Which clustering method yields better hierarchy?

  (2.1.) Can we achieve better predictive models by using the hierarchies obtained by structuring the feature ranking or target space?

(3.) Are the data-driven hierarchies better than the hierarchies created by a domain expert?

(4.) How the structuring of the output space scales from small to large output spaces?

(5.) How the performance difference translates from single model to ensemble of models?

In order to answer the above questions, we perform an extensive evaluation on a diverse datasets from the environmental and socio-economic domain. In the following part, we will describe the data we use.

B. DATA DESCRIPTION

We use 16 datasets for multi-target regression benchmark problems from 2 different domains (8 from the domain of socio-economic sciences and 8 from the domain of environmental sciences, from which 14 with small and 2 with large number of targets). The number of targets in the datasets range from 6 to 492 and the number of descriptive attributes from 16 to 576. The datasets with large number of targets (> 100) are inspected separately. The number of instances is also diverse ranging from 42 to 16976. The basic information and statistics about these datasets are given in Table 1.

The *Andromeda (andro)* dataset is for prediction of 6 water quality variables in Thermaikos Gulf of Thessaloniki, Greece [67]. The *Airline Ticket Price* datasets are used to infer the minimal price of an airline ticket for the next day (*atp1d*) i.e., next 7 days (*atp7d*) [68]. *Metal* data (*mdv2*) is the data for meta-learning of an automated assistant system for choosing appropriate machine learning algorithms for a specific data mining process [69]. The *Occupational Employment Survey* datasets are from the US Bureau of Labor Statistics for the years 1997 (*oes97*) and 2010 (*oes10*) [15]. The *Online sales* (*osales*) dataset deals with the prediction of online sales of products described with various product features. The dataset is from the Kaggle's Online Product Sales competition in 2012 [70]. *Prespa Diatoms Lake* (*pd*) and *Prespa Diatoms Lake top 10* (*pdt*) datasets investigate the effect of the environmental conditions of Lake Prespa in the Republic of Macedonia on diatom communities [71]. The former (*pd*) is the complete data set with all 111 targets and examples, while the latter (*pdt*) consists of only top

TABLE 1. Properties of the used benchmark datasets in terms of number of instances (#inst), number of descriptive attributes (D), number of targets (T), percentage of missing values (MissVal) and sorted by number of instances. The datasets with * as superscript will be considered separately, since they have large number of targets.

Dataset name	Abbr.	#inst	D	T	MissVal
Water quality	**wq**	1060	16	14	/
Andromeda	**andro**	49	30	6	/
Online Sales	**osales**	639	413	12	3.79%
Occupational Employment Survey for 1997	**oes97**	334	263	26	/
Occupational Employment Survey for 2010	**oes10**	403	298	16	/
Metal data	**mdv2**	42	53	10	24%
Prespa Lake Diatoms*	**pd***	349	16	111	0.11%
Prespa Lake Diatoms Top 10	**pdt**	248	16	10	0.54%
Airline Ticket Price (1 day)	**atp1d**	337	411	6	/
Airline Ticket Price (7 days)	**atp7d**	296	411	6	/
Vegetation conditions	**vgc**	16967	40	7	/
River Flows 1	**rf1**	9125	64	8	0.5%
River Flows 2	**rf2**	9125	576	8	6.68%
Slovenian Rivers*	**SloRiv***	1060	16	492	/
Supply Chain Management tournament (1 day)	**scm1d**	9803	280	16	/
Supply Chain Management tournament (20 days)	**scm20d**	8966	61	16	/

10 the most abundant diatoms. *River Flows* (*rf1* and *rf2*) are datasets for prediction of the river network flows in the Mississippi river in the United States obtained from the US National Weather Service consists of 8 sites, with 8 attributes from each site [15]. The difference between *rf1* and *rf2* is that the latter includes the forecast information about the precipitation. The SCM datasets are from the 2010 Trading Agent Competition in the Supply Chain Management tournament (TAC SCM). It consists of 4-time delayed observations for traded prices of various computing equipment for the specific day (i.e., prices from 1, 2, 4 and 8 days ago vs. the price today) and trying to predict the forward trend of the next tournament day price (*scm1d*), i.e., the mean price of the next 20 tournament days (*scm20d*) [72]. The *Vegetation condition*(*vgc*) dataset concerns the prediction of the vegetation condition for the Victoria State in Australia and provided by the Arthur Rylah Institute for Environmental Research, Department of Sustainability and Environment (DSE) [3]. *Water quality* (*wq*) and *Slovenian Rivers* (*SloRiv*) are two datasets for predicting species abundance in water in Slovenian rivers using 16 chemical parameters as a descriptors. The *wq* data set consists of only 14 the most abundant species, while the *SloRiv* dataset consists of 492 different species which occur more than 5 times in the samples [65], [73].

C. EVALUATION MEASURES

We follow the literature recommendations regarding the evaluation measures [19]. We present the values of the average relative root mean squared error (*aRRMSE*) (Eq 7) for performance of the tested methods. To perform a fair comparison, we calculate these errors only for the target variables at the leafs of the hierarchy.

Let us assume that t is the number of target variables and N_{test} is the size of the test set. The actual value of a target variable of an example is Y, and $\hat{Y}$ is the predicted value using the model for that example. Similarly, $\bar{Y}$ is the average of the actual values for that target variable. The *aRRMSE* can be define as follows:

$$\begin{aligned} aRRMSE &= \frac{1}{t}\sum_{i=1}^{t} RRMSE_i \\ &= \frac{1}{t}\sum_{i=1}^{t}\sqrt{\frac{\sum_{k=1}^{N_{test}}(Y_i^{(k)} - \hat{Y}_i^{(k)})^2}{\sum_{k=1}^{N_{test}}(Y_i^{(k)} - \bar{Y}_i)^2}} \end{aligned} \qquad (7)$$

If $aRRMSE \approx 0$, then we have much better performance, but if $aRRMSE \approx 1$, we have a closer value to the default prediction that predicts the average value for each target.

D. PARAMETER INSTANTIATION

The majority of our experiments are performed using the CLUS software package (https://sourceforge.net/projects/clus/), where the predictive clustering framework for MTR and HMTR tasks, including PCTs for MTR/HMTR, random forests of PCTs for MTR/HMTR and feature ranking [6], [9] are implemented. The algorithms are developed to natively handle missing values.

A hierarchical tree defined by the used clustering methods in HMTR are defined as tree shaped hierarchies. For obtaining a hierarchy using the agglomerative clustering method, we use the non-commercial version of OCTAVE software package (functions *pdist()*, *linkage()* and *dendrogram()*). Furthermore, in OCTAVE, we used balanced k-means clustering for numerical type values, which is based on Hungarian (Munkres') assignment algorithm to assign the examples to the clusters [74]. Since most of the datasets have a relatively small number of targets (except the two with more than 100), we selected the value $k = 2$ for balanced k-means in order to obtain more branched hierarchies.

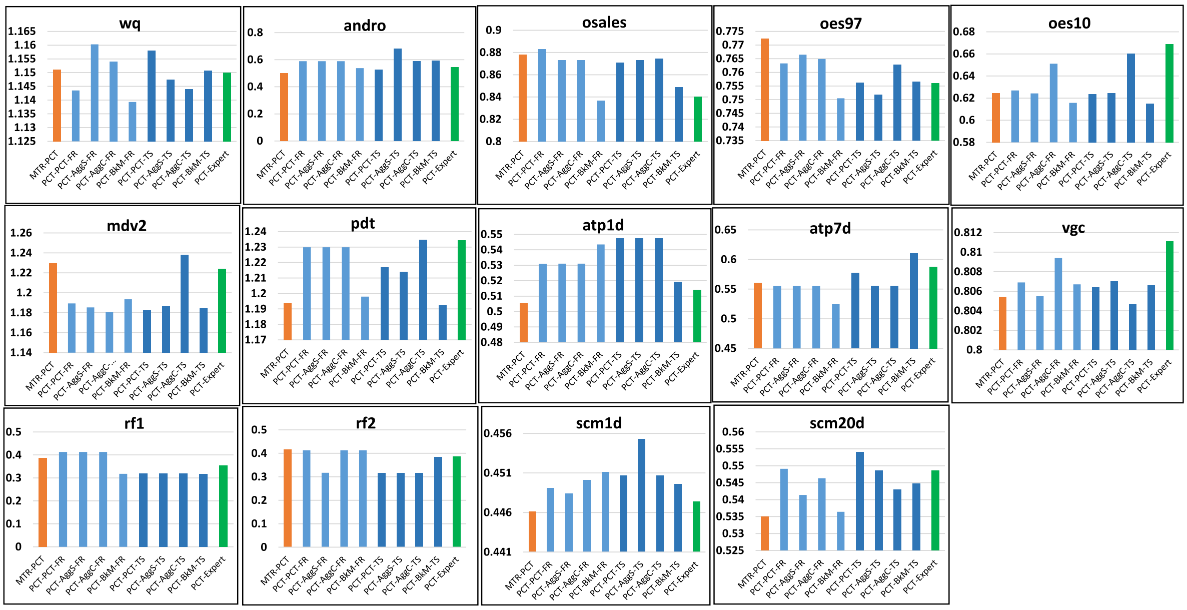


FIGURE 2. Results for the predictive performance of *single PCTs* from experiments per dataset represented by aRRMSE. Green bars represent hierarchies created by an expert and orange bars represent the flat MTR results.

We use Euclidean distance metric in all our algorithms that require distance. In HMTR, as defined in previous sections, we use weighted Euclidean distance. Moreover, for random forest for feature ranking, we use GENIE3 as a feature importance method based on variable selection with ensembles of PCTs [59], [60]. We use 100 base ppredictive models for the random forests in all tasks (MTR, HMTR and feature ranking). For PCTs for HMTR task, we use sum as an aggregation function with the weight set to 0.75 [13].

E. HIERARCHIES CREATED BY A DOMAIN EXPERT

In our analysis, we also use hierarchies created by the domain experts, defined as a class ontology or domain-specific class structure. In the following part, we explain the creation of the hierarchies for each dataset.

The hierarchy in *mdv2* (Metal data) dataset is created based on a type of machine learning algorithm in three hierarchy levels. For *andro* (Andromeda) dataset is created based on correlation matrix given in [67]. For *pdt* (Prespa Lake Top 10) data set, the top 10 most abundant diatoms are grouped into a hierarchy based on their taxonomic rank. For *atp1d* and *atp7d* (Airline ticket prices) datasets, the target classes are grouped based on the type of the flight, either non-stop flight or with any number of stops. For *oes97* and *oes10* (Occupational Employment Survey 1997 and 2010), the target classes are organized into a hierarchy based on the type of the occupation and specific job position. For *osales* (Online Sales) data set, the hierarchies are created based on sales products in first and the second half of the year. For *wq* (Water Quality) and *SloRiv* (Slovenian Rivers) datasets the hierarchies are created based on the taxonomic tanks of the species. The expert hierarchy for *rf1* and *rf2* (River Flows) datasets is constructed based on three different river network flows (Illionis, Iowa and Missouri). The hierarchy for *scm1d* and *scm20d* (Supply Chain Management) datasets is created based on the grouping the 16 PC configurations (targets) on 3 main market segments (low, medium and high) consisting of a combination of 10 different components, as it is given in Table 5 in the report [75]. Finally, the hierarchy for the *vgc* (Vegetation conditions) data set in created based on grouping of the target classes, either to tree related scores or other type of scores [3].

F. STATISTICAL EVALUATION

To validate our predictive models, we use 10-fold cross validation in all settings. More specifically, the whole dataset is first randomly split into 10 folds. Next, 9 folds are used for training, and the remaining one for testing. The procedure is repeated 10 times so that each fold is used exactly once as test set. The reported results represent an average of all 10 runs.

For statistical evaluation of the results, we adhered to the recommendations by [76]. For assessing the statistical significance of the differences, we used the non-parametric Friedman test [77] with the correction recommended by [78]. In order to compare the methods and to check the statistical significance among them, we used the Nemenyi post-hoc test [79]. The result from Nemenyi post-hoc test is presented with an average ranks diagram [76]. For statistical comparison between two algorithms, we used the Wilcoxon signed-rank non-parametric statistical hypothesis test [80].

V. RESULTS

In this section, we present the obtained results from the performed experiments using the procedures for structuring the output space. In our study, as output spaces, we consider the

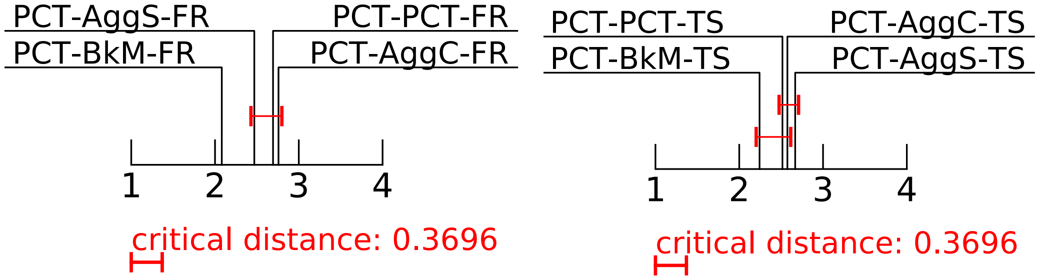


FIGURE 3. Average rank diagrams for algorithms that cluster the target space (left) and feature ranking space (right) using single PCTs.

space consisting of the target values or the space consisting of feature ranks for each target. We compare the following methods for hierarchy construction:

- flat MTR problem (no hierarchy) (*MTR*);
- agglomerative clustering with single linkage (*AggS*);
- agglomerative clustering with complete linkage (*AggC*)
- balanced k-means clustering (*BkM*)
- clustering using predictive clustering trees (*PCT*).
- hierarchy created by an expert (*Expert*)

Since we have two different models (single PCTs model and random forest of PCTs) and two different structured output spaces, we show separately the results for single PCTs (Fig 2) and random forest of PCTs (Fig 7). To clarify the notation, we need to distinguish between using either single tree or random forest of PCTs and different methods of structuring the output space (target space and feature ranking space). To achieve this, we use prefixes (*PCT-* and *RF-*) and suffixes (*-TS* and *-FR*) before and after the hierarchy construction method name, accordingly. For example, *RF-BkM-TS* refers to the balanced k-means method used on the original target space using random forest of PCTs for model creation. Then, *PCT-PCT-FR* refers to the clustering method with PCTs of the output space consisting of feature rankings using single PCTs for building the model, etc.

Fig 2 visually presents the results of the predictive performance of single PCTs for each dataset. Examining the figure, it is clear that data-driven hierarchies, generally, improve the predictive performance over the flat MTR task, except on five datasets (*andro*, *pdt*, *atp1d*, *scm1d* and *scm20d*). It is interesting to notice that, for most of the datasets with more than 12 targets (*oes97, oes10, osales, wq*), using hierarchies noticeably improve the performance over flat MTR (with no hierarchies). Those results give an insight that, for the datasets with large number of targets, there is an improvement of the performance if the hierarchies obtained by structuring the target space, are used.

In order to figure out which data-driven clustering method for hierarchy creation performed the best, we created an average rank diagrams for aRRMSE values per output space for $p-value = 0.05$. More specifically, Fig 3 (left) illustrates the average diagram for clustering methods over the target space and Fig 3 (right) gives the average rank diagram for clustering methods over the feature ranking space. We can see that the best method for hierarchy creation over target space is *PCT-BkM-TS*, and it is only significantly better than *PCT-AggS-TS*. From the other side, in the average rank diagrams for the clustering methods over the feature ranking space, we can see that *PCT-BkM-FR* is the best performing method and it is significantly better than all others. Therefore, for task of MTR with single PCTs, we can easily recommend using balanced k-means clustering method for creation of hierarchies from the output space (either target or feature rankings space).

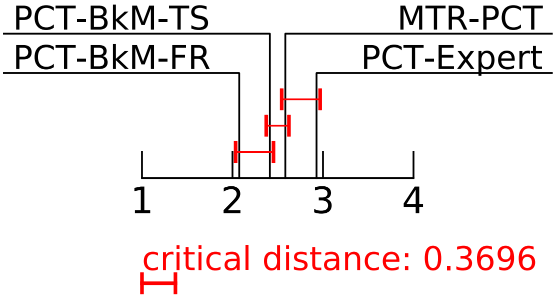


FIGURE 4. Average rank diagrams for the best algorithms from Fig. 3 compared to the flat MTR task and the use of an expert created hierarchy on the target space.

In order to check the significance of the performance between the two best approaches for hierarchy creation (considering the two target spaces), we perform non-parametric Wilcoxon hypothesis test for p-value = 0.05 for the *PCT-BkM-FR* and *PCT-BkM-TS* algorithms. The results show that *PCT-BkM-FR* > *PCT-BkM-TS*; p-value = 0.0325 < 0.05, which means that *PCT-BkM-FR* is statistically significantly better method than *PCT-BkM-TS*.

Considering this, we have that the hierarchies constructed over the space consisting of feature importances are superior to the hierarchies constructed over the target space, both using balanced k-means method for clustering.

For a clearer picture over the best clustering method performance and the performance of the flat MTR method and

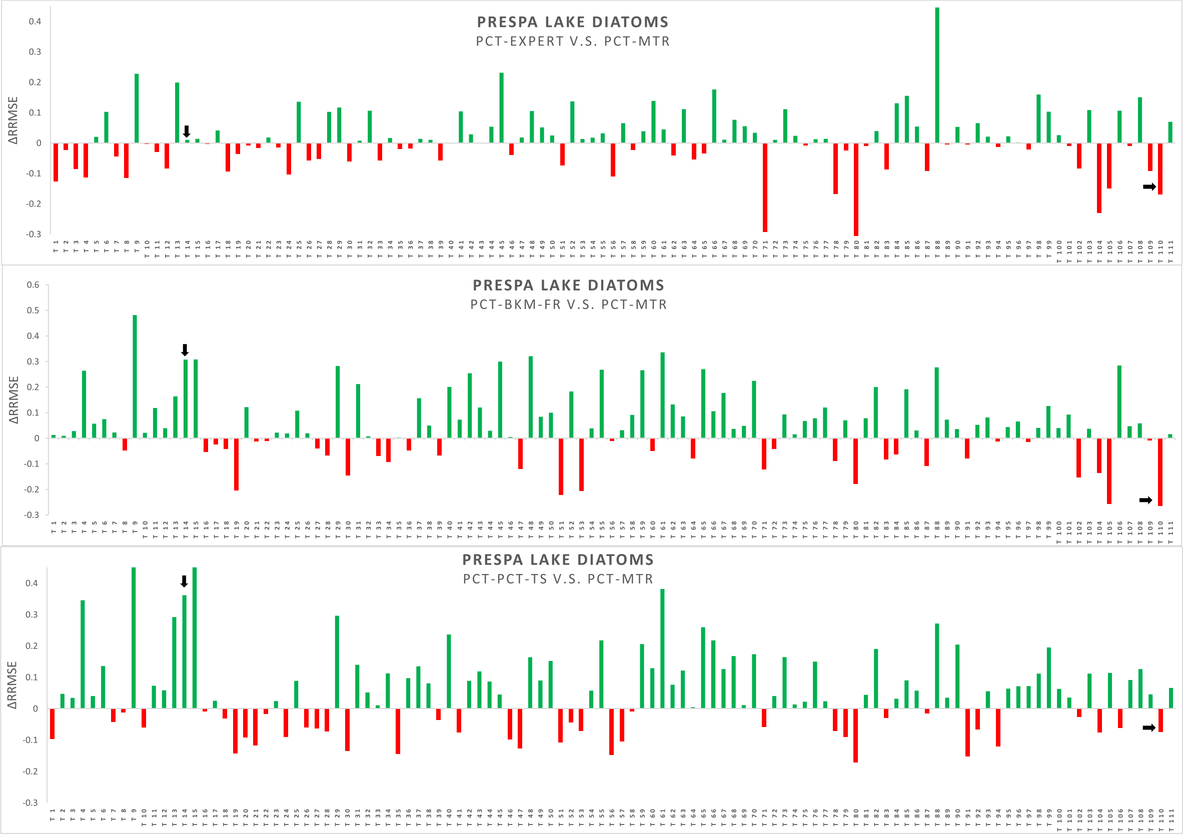


FIGURE 5. **Δ*RRMSE* values for Prespa Diatom Lake dataset *pd* using single PCTs and using expert created hierarchy (*PCT-Expert*), balanced k-means on a feature ranking space (*PCT-BkM-FR*) and predictive clustering trees for clustering the target space (*PCT-PCT-TS*). The arrows represent the chosen examples with good/bad performance.**

using the hierarchy created by an expert in HMTR task, we took the best performing methods for structuring the output space (*PCT-BkM-TS* and *PCT-BkM-FR*) and compare together with flat MTR task performance (*MTR-PCT*) and the performance of the hierarchy created by an expert (*PCT-Expert*). The average rank diagram from statistical evaluation is given in Fig 4. We can see that *PCT-BkM-FR* is the superior algorithm, and significantly better than *MTR-PCT* and *PCT-Expert*. All in all, data-driven hierarchies improve the predictive performance in multi-target regression problems.

If we consider the performances for aRRMSE using random forest of PCT algorithm, we can see that in all data sets, the aRRMSE is reduced, which is in accordance with the general rule-of-thumb for the random forest. In the Appendix part of the paper, the results for *aRRMSE* using random forest of PCTs per dataset are given in Fig 7.

To investigate the translation of predictive performance from single PCTs to ensemble of PCTs, we performed the same experimental analysis and statistical evaluation. Similar conclusions can be made as for the single PCTs. Generally, hierarchies improve the predictive performance over the flat MTR or expert created hierarchies (in eleven out of sixteen datasets). But, there is no statistically significant difference between the performances of used clustering algorithms and the flat MTR algorithm. The average rank diagrams for aRRMSE using random forest are given in Fig 8 and Fig 9 in the Appendix.

The detailed results of the predictive performance (*aRRMSE*) for each dataset that were used to draw the graphs in Figure 2 for single PCTs, i.e., in Figure 7 for random forests of PCTs, are given in Figure 12 in the Appendix.

We must note here that we exclude both large datasets (PD and SLORIV) from the statistical analysis, because the high number of targets will influence the overall per-target evaluation and will guide us towards the statistically incorrect conclusions. For that reason, we consider those two datasets separately in the next subsection.

A. STRUCTURING LARGE OUTPUT SPACES

In this subsection, we present the results from the experiments performed on the two datasets with large number of targets: *Prespa Diatoms Lake* (*pd*) with 111 targets and *Slovenian rivers* (*SloRiv*) with 492 targets. The main goal here is to make a more comprehensive and sustainable study which will take

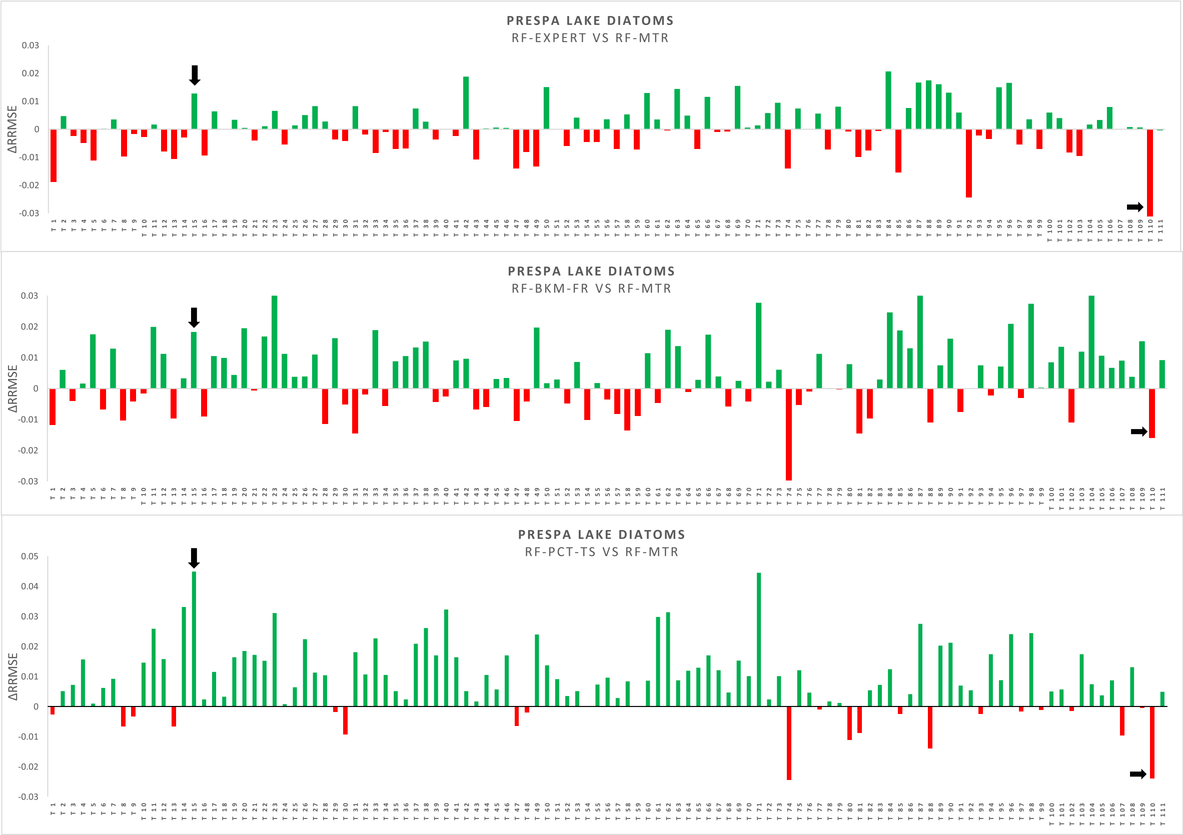


FIGURE 6. **Δ*RRMSE* values for Prespa Diatom Lake dataset *pd* using random forest of PCTs and using expert created hierarchy (*RF-Expert*), balanced k-means on a feature ranking space (*RF-BkM-FR*) and predictive clustering trees for clustering the target space (*RF-PCT-TS*). The arrows represent the chosen examples with good/bad performance.**

into consideration the size of the output space, i.e., the target space cardinality.

The balance k-means clustering algorithm for hierarchy creation, especially on the space consisting of feature rankings, is the best performing method based on above results. Furthermore, in the study of [66], they recommend to use the divisive methods for hierarchy creation and to some extend this relates with our results from the statistical evaluation. For that reason, we use the divisive methods (balanced k-means and predictive clustering trees) for clustering the output spaces for the two big datasets. More precisely, we show the results for clustering the target space using predictive clustering trees (PCT-PCT-TS and RF-PCT-TS) and for clustering the feature rankings space using balanced k-means (PCT-BkM-FR and RF-BkM-FR). The results are analysed as per target performance of the data-driven hierarchy creation methods and expert constructed hierarchy compared to the performance of the flat MTR task.

To better illustrate the results, we calculate the difference $\Delta RRMSE$, which is the difference between $RRMSE$ value of flat MTR and the $RRMSE$ from the appropriate method for hierarchy creation. The results for the *pd* dataset using single PCTs are shown in Fig 5. The green bars present the per target $RRMSE$ values that denote that HMTR models are better than flat MTR models (positive value for $\Delta RRMSE$), while the red bars present the per target $RRMSE$ values where MTR models are better than HMTR models (negative values for $\Delta RRMSE$). Examining the results, we can see that using the *PCT-BkM-FR* method, we obtain the best per-target performance. Specifically, by using *PCT-Expert* compared to *PCT-MTR* in the *pd* dataset, we have 60 out of 111 targets where *PCT-Expert*>*PCT-MTR*, then using *PCT-BkM-FR* compared to *PCT-MTR*, we have 76 out of 111 (68.5%) targets, where *PCT-BkM-FR* > *PCT-MTR* and finally, using *PCT-PCT-TS* v.s *PCT-MTR*, we have 72 out of 111 target, where *PCT-PCT-TS* > *PCT-MTR*.

The results from the *SloRiv* dataset are shown in Fig 10 (in Appendix). Here, by visual inspection of the results, we can see that using the hierarchy created by *PCT-PCT-TS* algorithm we obtain a better performance on the most of the targets compared to the *PCT-MTR* algorithm, i.e., *PCT-PCT-TS* > *PCT-MTR* in 325 out of 492 (66%) targets.

Furthermore, using random forest of PCTs yields quite similar situation. The difference here with single PCTs is

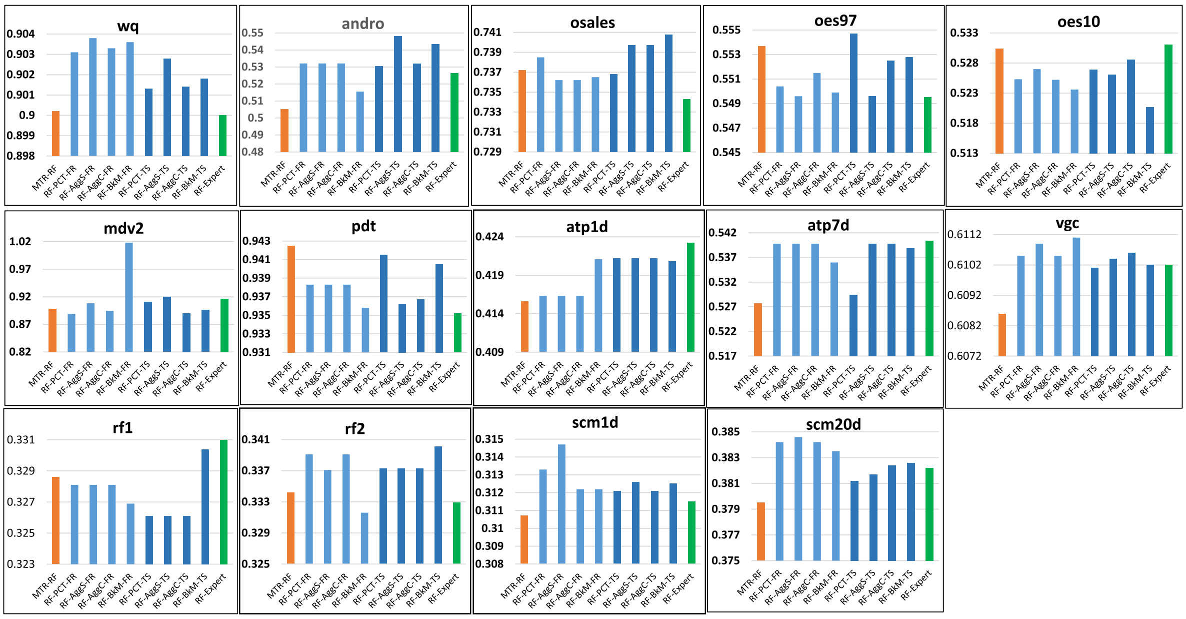


FIGURE 7. Results for the predictive performance of *Random forest of PCTs* from experiments per dataset represented by aRRMSE. Green bars represent hierarchies created by an expert and orange bars represent the flat MTR results.

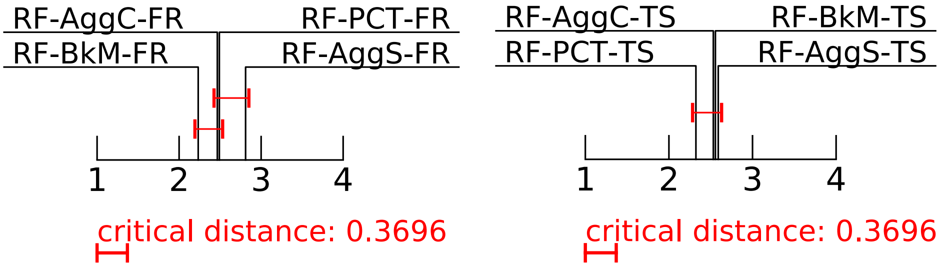


FIGURE 8. Average rank diagrams for algorithms cluster the target space (left) and feature ranking space (right) using random forest of PCTs.

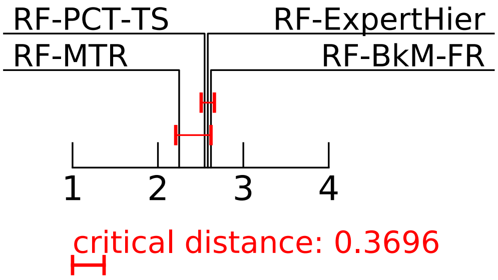


FIGURE 9. Average rank diagrams for the best algorithms from Fig.8 compared to the flat MTR task and the use of an expert created hierarchy on the target space.

that *RF-PCT-TS* clustering method gives the best results on the *pd* dataset. Specifically, we have *RF-PCT-TS* > *RF-MTR* in 89 out of 111 (80%) targets. This is a very good improvement compared to the other clustering methods for hierarchy creation. The results for the *pd* dataset are shown in Fig 6. Examining the results for the *SloRiv* dataset, again, same as single PCTs, we can see that by using *RF-PCT-TS* method we can obtain the best per-target performances i.e., *RF-PCT-TS* > *RF-MTR* in 287 out of 492 (59.5%) targets. The results for the *SloRiv* dataset are shown in Fig 11 from the Appendix.

Generally, on the larger datasets, there is an improvement of the performance, when the hierarchies are used. More precisely, divisive methods for clustering (hierarchy creation) are the best methods for structuring the output space, which is in accordance with the conclusions from the recent literature [17], [66]. Furthermore, data-driven hierarchies are generally better than the hierarchies created by an domain expert. It is confirmed by our results as well.

Examining the arrows in Fig 10 and Fig 11 (in Appendix) shown for Slovenian Rivers (*SloRiv*) dataset, we can see that for example, considering the target number 170 (which is taxa *Euglena viridis* from taxonomic group *EUGLENOPHYTA*),

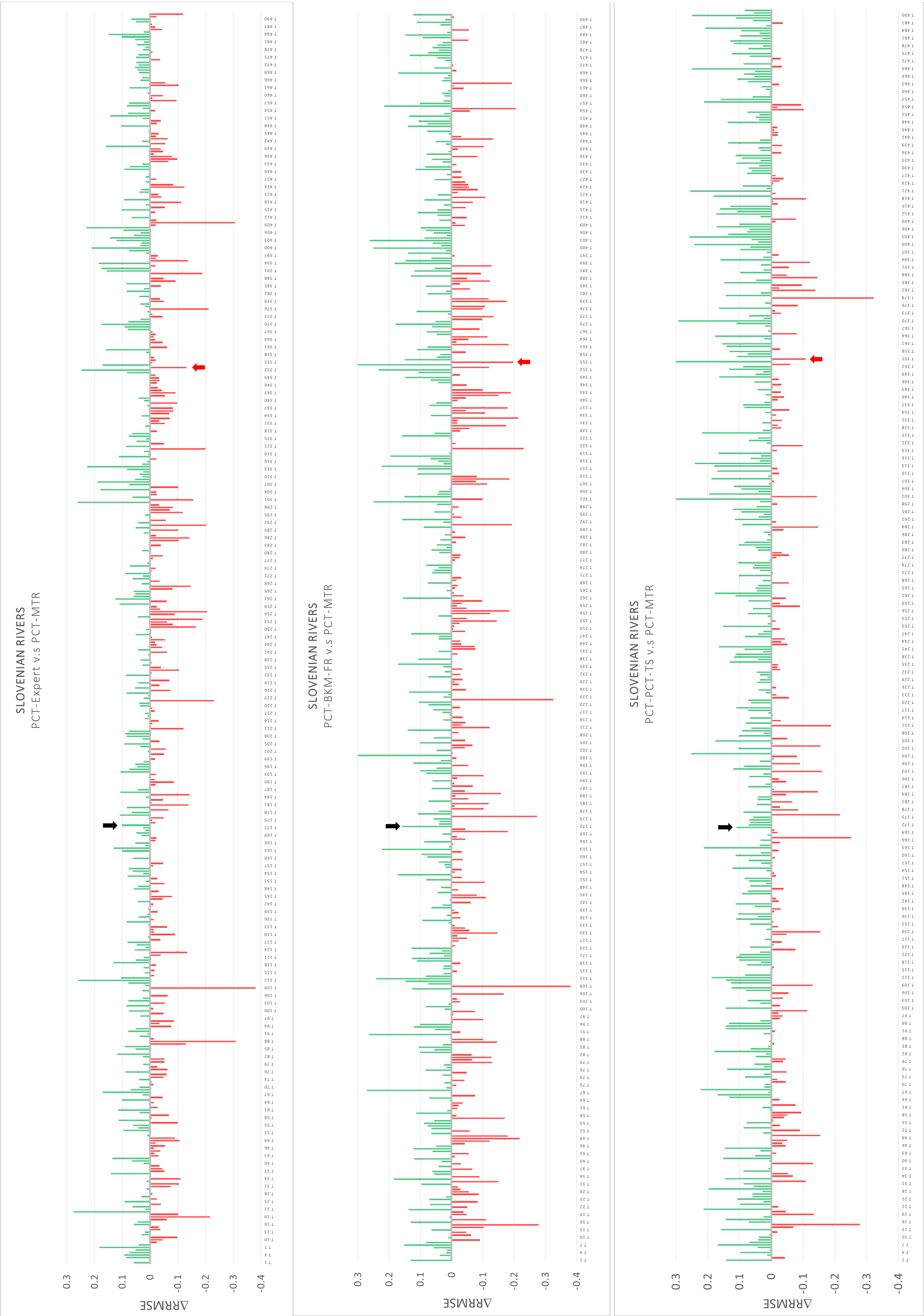


FIGURE 10. **Δ*RRMSE* values for Slovenian rivers *SloRiv* dataset using single PCTs and using expert created hierarchy (*PCT-Expert*), balanced k-means on a feature ranking space (*PCT-BkM-FR*) and predictive clustering trees for clustering the target space (*PCT-PCT-TS*). The arrows represent the chosen examples where we have good/bad performance.**

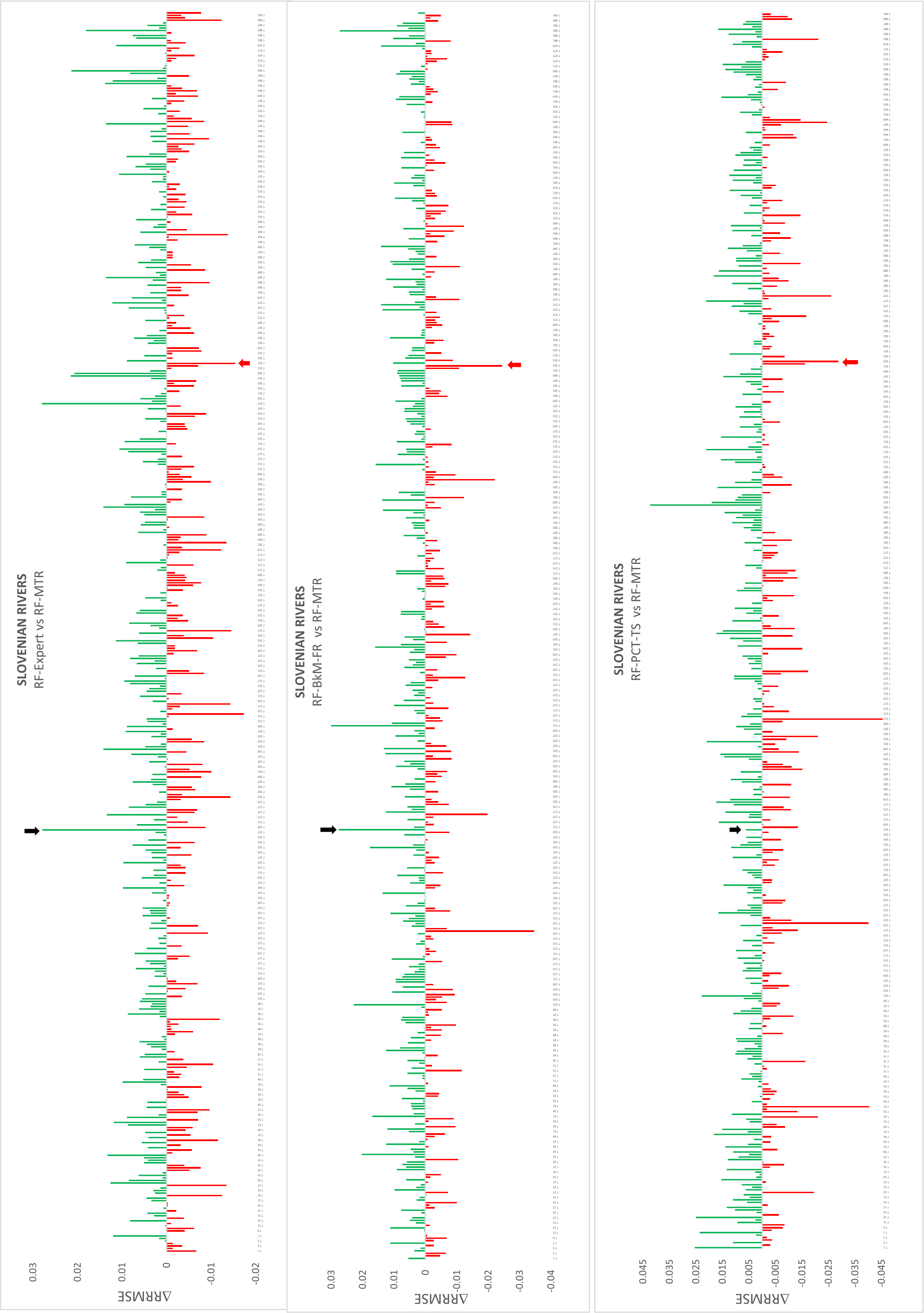
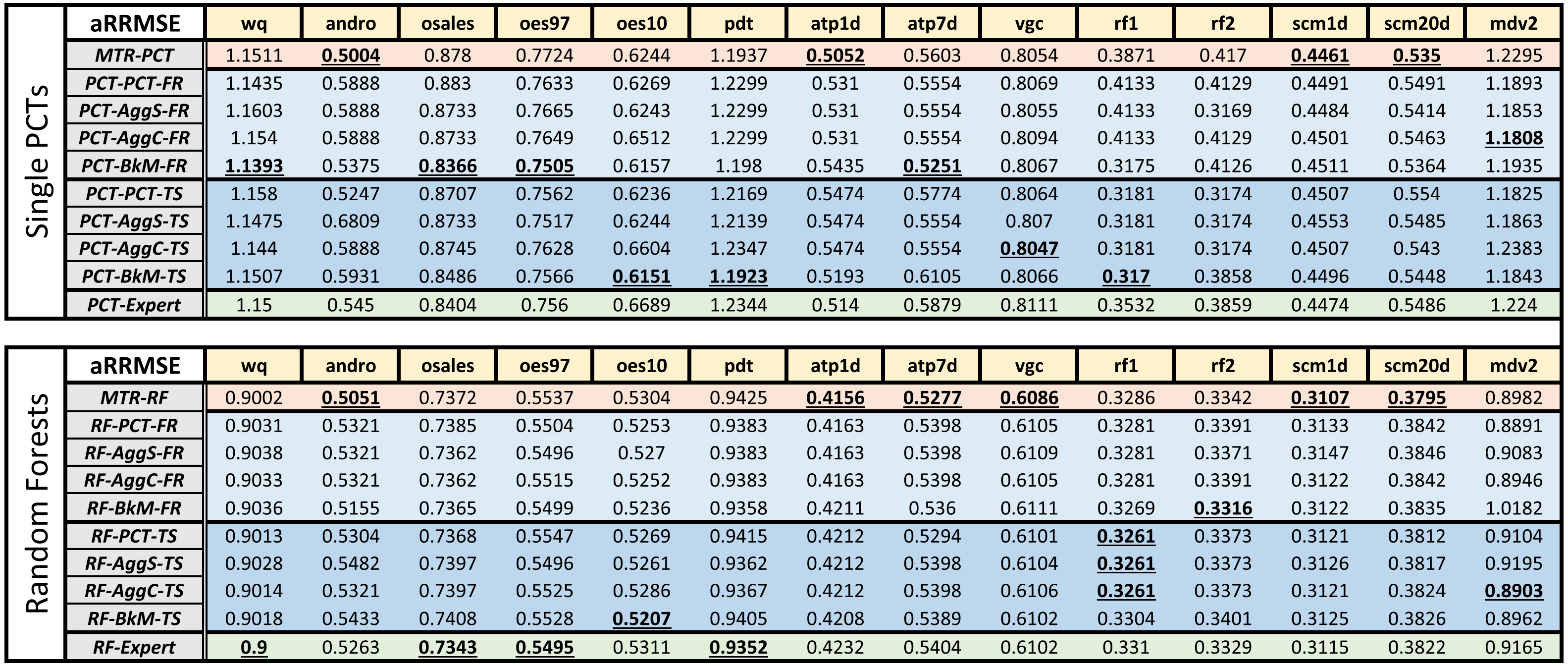


FIGURE 11. Δ*RRMSE* values for Slovenian rivers dataset *SloRiv* using random forest of PCTs and using expert created hierarchy (*PCT-Expert*), balanced k-means on a feature ranking space (*PCT-BkM-FR*) and predictive clustering trees for clustering the target space (*PCT-PCT-TS*). The arrows represent the chosen examples where we have good/bad performance.



Single PCTs	aRRMSE	wq	andro	osales	oes97	oes10	pdt	atp1d	atp7d	vgc	rf1	rf2	scm1d	scm20d	mdv2
	MTR-PCT	1.1511	**0.5004**	0.878	0.7724	0.6244	1.1937	**0.5052**	0.5603	0.8054	0.3871	0.417	**0.4461**	**0.535**	1.2295
	PCT-PCT-FR	1.1435	0.5888	0.883	0.7633	0.6269	1.2299	0.531	0.5554	0.8069	0.4133	0.4129	0.4491	0.5491	1.1893
	PCT-AggS-FR	1.1603	0.5888	0.8733	0.7665	0.6243	1.2299	0.531	0.5554	0.8055	0.4133	0.3169	0.4484	0.5414	1.1853
	PCT-AggC-FR	1.154	0.5888	0.8733	0.7649	0.6512	1.2299	0.531	0.5554	0.8094	0.4133	0.4129	0.4501	0.5463	**1.1808**
	PCT-BkM-FR	**1.1393**	0.5375	**0.8366**	**0.7505**	0.6157	1.198	0.5435	**0.5251**	0.8067	0.3175	0.4126	0.4511	0.5364	1.1935
	PCT-PCT-TS	1.158	0.5247	0.8707	0.7562	0.6236	1.2169	0.5474	0.5774	0.8064	0.3181	0.3174	0.4507	0.554	1.1825
	PCT-AggS-TS	1.1475	0.6809	0.8733	0.7517	0.6244	1.2139	0.5474	0.5554	0.807	0.3181	0.3174	0.4553	0.5485	1.1863
	PCT-AggC-TS	1.144	0.5888	0.8745	0.7628	0.6604	1.2347	0.5474	0.5554	**0.8047**	0.3181	0.3174	0.4507	0.543	1.2383
	PCT-BkM-TS	1.1507	0.5931	0.8486	0.7566	**0.6151**	**1.1923**	0.5193	0.6105	0.8066	**0.317**	0.3858	0.4496	0.5448	1.1843
	PCT-Expert	1.15	0.545	0.8404	0.756	0.6689	1.2344	0.514	0.5879	0.8111	0.3532	0.3859	0.4474	0.5486	1.224

Random Forests	aRRMSE	wq	andro	osales	oes97	oes10	pdt	atp1d	atp7d	vgc	rf1	rf2	scm1d	scm20d	mdv2
	MTR-RF	0.9002	**0.5051**	0.7372	0.5537	0.5304	0.9425	**0.4156**	**0.5277**	**0.6086**	0.3286	0.3342	**0.3107**	**0.3795**	0.8982
	RF-PCT-FR	0.9031	0.5321	0.7385	0.5504	0.5253	0.9383	0.4163	0.5398	0.6105	0.3281	0.3391	0.3133	0.3842	0.8891
	RF-AggS-FR	0.9038	0.5321	0.7362	0.5496	0.527	0.9383	0.4163	0.5398	0.6109	0.3281	0.3371	0.3147	0.3846	0.9083
	RF-AggC-FR	0.9033	0.5321	0.7362	0.5515	0.5252	0.9383	0.4163	0.5398	0.6105	0.3281	0.3391	0.3122	0.3842	0.8946
	RF-BkM-FR	0.9036	0.5155	0.7365	0.5499	0.5236	0.9358	0.4211	0.536	0.6111	0.3269	**0.3316**	0.3122	0.3835	1.0182
	RF-PCT-TS	0.9013	0.5304	0.7368	0.5547	0.5269	0.9415	0.4212	0.5294	0.6101	**0.3261**	0.3373	0.3121	0.3812	0.9104
	RF-AggS-TS	0.9028	0.5482	0.7397	0.5496	0.5261	0.9362	0.4212	0.5398	0.6104	**0.3261**	0.3373	0.3126	0.3817	0.9195
	RF-AggC-TS	0.9014	0.5321	0.7397	0.5525	0.5286	0.9367	0.4212	0.5398	0.6106	**0.3261**	0.3373	0.3121	0.3824	**0.8903**
	RF-BkM-TS	0.9018	0.5433	0.7408	0.5528	**0.5207**	0.9405	0.4208	0.5389	0.6102	0.3304	0.3401	0.3125	0.3826	0.8962
	RF-Expert	**0.9**	0.5263	**0.7343**	**0.5495**	0.5311	**0.9352**	0.4232	0.5404	0.6102	0.331	0.3329	0.3115	0.3822	0.9165

FIGURE 12. Detailed results for the predictive performance (*aRRMSE*) per dataset corresponding to the graphical results in Figure 2 for single PCTs i.e., in Figure 7 for random forests of PCTs.

there is a significant improvement in the performance, if the hierarchies are used rather than they are not used. The average abundance of all species in the examples is 71.8. The target 170 occurs 13 times in the examples, which is quite below the average. This confirms the fact that with small occurrence of the target in the examples, the model performance will be lower than considering a whole hierarchy (target dependence), where the target will be included. This is in accordance with the fact that, if we build a model with structuring of the output space (HMTR task), we can improve the predictive performance compared to the models built on a flat MTR task. Alternatively, if we want to check why the hierarchies do not help on some of the targets, as an example, we can select the target 353, which represents the taxa *Heptagenia sulphurea* from the *EPHEMEROPTERA* taxonomic group.

Similarly, examining the arrows in Fig 5 and Fig 6 for Prespa Diatoms dataset *pd* with 111 targets, we can make similar conclusions as for the previous dataset. For example, if we select the target number 14, on which we have the best performance by using hierarchies compared to the flat MTR task, the occurrence of this target in the examples is 5 times, but the average occurrence of the targets is 33.5. Therefore, as less the target occur in the examples, as much better performance can be achieved by using the structure of the output space (hierarchy) rather than using a flat MTR task, where no hierarchy is considered.

VI. CONCLUSION

In this paper, we present two data-driven approaches for structuring the output space. Namely, we present an algorithm for clustering the targets and the algorithm for clustering the targets according to the importance scores of each feature per target. Our research is focused on the question whether the two data-driven methods for structuring the output space can improve the predictive performance on the original flat multi-target regression task, and, moreover, whether data-driven hierarchies are better than expert created hierarchies.

For constructing the hierarchies, we investigate the use of agglomerative clustering method with single and complete linkage, balanced k-means clustering and clustering using PCTs. The resulting problem is then transformed into a HMTR problem, and finally addressed by using PCTs and random forests of PCTs for HMTR. We use 16 benchmark datasets to evaluate the performance of all methods. Two datasets have a large number of targets ($>$ 100 targets). After obtaining the results for the average RMMSE (*aRRMSE*), we perform a statistical evaluation by using Friedman non-parametric test with Nemenyi post-hoc testing and Wilcoxon statistical test for testing the two best methods for structuring the output space.

The results show that for single PCTs, the data-driven approach for structuring (clustering) the output space significantly increases the predictive performance over the original MTR task and over the performance obtained by using an expert created hierarchy. A recommendation that comes out from the statistical evaluation is that balanced k-means algorithm can be used for clustering the output space. Moreover, by using hierarchies created over the feature ranking space there is an improvement in the performance. The same, but to a lesser extent, conclusions can be made by using ensembles of PCTs, since they are not improving the predictive performance significantly.

For large output spaces, datasets with a large number of targets (greater than 100), the results show that hierarchies improve the performance compared to using the flat MTR task, where no hierarchy is considered. For structuring the large output spaces, the divisive methods for hierarchy creation are the best choice, since they are constructing good hierarchies that improve the predictive performance. Moreover, data-driven hierarchies are a better choice than expert created hierarchies, which implies that we could obtain good structure of the target space if we discover the knowledge from the data directly rather than using the structure based on some pre-defined relations defined by a domain expert.

For further work, we plan to make more extensive evaluation on more datasets with a larger number of targets and to investigate different feature ranking methods (for example, RReliefF and attention mechanism based feature ranking with NNs). There are some insights that there might be potential improvements of the performance that can be achieved with cutting the obtained hierarchies based on data density, distance between the nodes etc. and addressing the task of MTR as multiple smaller MTR tasks.

APPENDIX

See Figs. 7–12.

REFERENCES

[1] G. Tsoumakas and I. Katakis, "Multi label classification: An overview," *Int. J. Data Warehouse Min.*, vol. 3, no. 3, pp. 1–13, 2007.

[2] D. Demšar, S. Džeroski, T. Larsen, J. Struyf, J. Axelsen, M. B. Pedersen, and P. H. Krogh, "Using multi-objective classification to model communities of soil microarthropods," *Ecol. Model.*, vol. 191, pp. 131–143, Jan. 2006.

[3] D. Kocev, S. Džeroski, M. White, G. R. Newell, and P. Griffioen, "Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition," *Ecol. Model.*, vol. 220, no. 8, pp. 1159–1168, 2009.

[4] J. Levatić, D. Kocev, M. Debeljak, and S. Džeroski, "Community structure models are improved by exploiting taxonomic rank with predictive clustering trees," *Ecol. Model.*, vol. 306, pp. 294–304, Jun. 2015.

[5] C. N. Silla and A. A. Freitas, "A survey of hierarchical classification across different application domains," *Data Mining Knowl. Discovery*, vol. 22, nos. 1–2, pp. 31–72, 2011.

[6] D. Kocev, C. Vens, J. Struyf, and S. Džeroski, "Tree ensembles for predicting structured outputs," *Pattern Recognit.*, vol. 46, no. 3, pp. 817–833, Mar. 2013.

[7] G. H. Bakır, T. Hofmann, B. Schölkopf, A. J. Smola, B. Taskar, and S. V. N. Vishwanathan, *Predicting Structured Data* (Neural Information Processing). Cambridge, MA, USA: MIT Press, 2007.

[8] H. Blockeel, "Top-down induction of first-order logical decision trees," Ph.D. dissertation, Katholieke Univ. Leuven, Leuven, Belgium, 1998.

[9] J. Struyf and S. Džeroski, "Constraint based induction of multi-objective regression trees," in *Proc. 4th Int. Workshop Knowl. Discovery Inductive Databases (KDID)*, in Lecture Notes in Computer Science, vol. 3933. Berlin, Germany: Springer, 2006, pp. 222–233.

[10] D. Kocev, C. Vens, J. Struyf, and S. Džeroski, "Ensembles of multi-objective decision trees," in *Proc. 18th Eur. Conf. Mach. Learn.*, 2007, pp. 624–631.

[11] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, and H. Blockeel, "Decision trees for hierarchical multi-label classification," *Mach. Learn.*, vol. 73, no. 2, pp. 185–214, Nov. 2008.

[12] I. Slavkov, V. Gjorgjioski, J. Struyf, and S. Džeroski, "Finding explained groups of time-course gene expression profiles with predictive clustering trees," *Mol. BioSyst.*, vol. 6, no. 4, pp. 729–740, 2010.

[13] V. Mileski, S. Džeroski, and D. Kocev, "Predictive clustering trees for hierarchical multi-target regression," in *Advances in Intelligent Data Analysis XVI*, N. Adams, A. Tucker, and D. Weston, Eds. London, U.K.: Springer, 2017, pp. 223–234.

[14] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[15] E. Spyromitros-Xioufis, G. Tsoumakas, W. Groves, and I. Vlahavas, "Multi-target regression via input space expansion: Treating targets as inputs," *Mach. Learn.*, vol. 104, no. 1, pp. 55–98, 2016.

[16] G. Madjarov, D. Gjorgjevikj, I. Dimitrovski, and S. Džeroski, "The use of data-derived label hierarchies in multi-label classification," *J. Intell. Inf. Syst.*, vol. 47, no. 1, pp. 57–90, 2016.

[17] S. Nikoloski, D. Kocev, and S. Džeroski, "Structuring the output space in multi-label classification by using feature ranking," in *Proc. Int. Workshop NFMCP Conjunct ECML-PKDD*, Skopje, Macedonia, 2018, pp. 151–166.

[18] S. Džeroski, V. Gjorgjioski, I. Slavkov, and J. Struyf, "Analysis of time series data with predictive clustering trees," in *Proc. 5th Int. Workshop, Knowl. Discovery Inductive Databases (KDID)*, in Lecture Notes in Computer Science, vol. 4747. Berlin, Germany: Springer, 2007, pp. 63–80.

[19] H. Borchani, G. Varando, C. Bielza, and P. Larrañaga, "A survey on multi-output regression," *Data Mining Knowl. Discovery*, vol. 5, no. 5, pp. 216–233, 2015.

[20] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[21] J. H. Friedman, "Stochastic gradient boosting," *Comput. Statist. Data Anal.*, vol. 38, no. 4, pp. 367–378, 2002.

[22] E. Spyromitros-Xioufis, G. Tsoumakas, W. Groves, and I. Vlahavas, "Multi-label classification methods for multi-target regression," pp. 1159–1168, 2012, *arXiv:1211.6581*. [Online]. Available: https://arxiv.org/abs/1211.6581

[23] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," in *Proc. 20th Eur. Conf. Mach. Learn.*, 2009, pp. 254–269.

[24] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification," in *Advances in Knowledge Discovery and Data Mining*. Berlin, Germany: Springer, 2004, pp. 22–30.

[25] W. Zhang, X. Liu, Y. Ding, and D. Shi, "Multi-output LS-SVR machine in extended feature space," in *Proc. IEEE Int. Conf. Comput. Intell. Meas. Syst. Appl.*, Jul. 2012, pp. 130–134.

[26] J. Wang, Z. Chen, K. Sun, H. Li, and X. Deng, "Multi-target regression via target specific features," *Knowl.-Based Syst.*, vol. 170, pp. 70–78, Apr. 2019.

[27] A. J. Izenman, "Reduced-rank regression for the multivariate linear model," *J. Multivariate Anal.*, vol. 5, no. 2, pp. 248–264, 1975.

[28] A. van der Merwe and J. V. Zidek, "Multivariate regression analysis and canonical variates," *Can. J. Stat.*, vol. 8, no. 1, pp. 27–39, 1980.

[29] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B (Stat. Methodol.)*, vol. 58, no. 1, pp. 267–288, 1996.

[30] C. E. Rasmussen and C. K. Williams, *Gaussian Processes for Machine Learning*, vol. 38. Cambridge, MA, USA: MIT Press, 2006, pp. 715–719.

[31] L. Breiman and J. H. Friedman, "Predicting multivariate responses in multiple linear regression," *J. Roy. Stat. Soc. B (Stat. Methodol.)*, vol. 59, no. 1, pp. 3–54, 1997.

[32] A. Appice and S. Džeroski, "Stepwise induction of multi-target model trees," in *Proc. 18th ECML*, Warsaw, Poland, 2007, pp. 502–509.

[33] T. Hothorn, K. Hornik, and A. Zeileis, "Unbiased recursive partitioning: A conditional inference framework," *J. Comput. Graph. Statist.*, vol. 15, no. 3, pp. 651–674, Sep. 2006.

[34] T. Aho, B. Ženko, and S. Džeroski, "Rule ensembles for multi-target regression," in *Proc. 9th IEEE Int. Conf. Data Mining*, Dec. 2009, pp. 21–30.

[35] M. Breskvar, D. Kocev, and S. Džeroski, "Ensembles for multi-target regression with random output selections," *Mach. Learn.*, vol. 107, pp. 1673–1709, Nov. 2018.

[36] M. Pugelj and S. Džeroski, "Predicting structured outputs k-nearest neighbours method," in *Discovery Science* (Lecture Notes in Computer Science), vol. 6926. Berlin, Germany: Springer, 2011, pp. 262–276.

[37] W. J. Brouwer, J. D. Kubicki, J. O. Sofo, and C. L. Gilesd, "An investigation of machine learning methods applied to structure prediction in condensed matter," 2014, *arXiv:1405.3564*. [Online]. Available: https://arxiv.org/abs/1405.3564

[38] M. Sánchez-Fernández, M. de Prado-Cumplido, J. Arenas-García, and F. Pérez-Cruz, "SVM multiregression for nonlinear channel estimation in multiple-input multiple-output systems," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2298–2307, Jul. 2004.

[39] B.-H. Mevik and R. Wehrens, "The pls package: Principal component and partial least squares regression in R," *J. Stat. Softw.*, vol. 18, no. 2, pp. 1–24, 2007.

[40] T. Hastie, R. Tibshirani, and J. Friedman, "Additive models, trees, and related methods," in *The Elements of Statistical Learning*. New York, NY, USA: Springer, 2001, pp. 321–329.

[41] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The RPROP algorithm," in *Proc. IEEE Int. Conf. Neural Netw.*, Mar. 1993, pp. 586–591.

[42] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 855–868, May 2009.

[43] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," 2014, *arXiv:1402.1128*. [Online]. Available: https://arxiv.org/abs/1402.1128

[44] X. Zhen, M. Yu, X. He, and S. Li, "Multi-target regression via robust low-rank learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 497–504, Feb. 2018.

[45] M. Kuo, B. Mohler, S. L. Raudenbush, and F. J. Earls, "Assessing exposure to violence using multiple informants: Application of hierarchical linear model," *J. Child Psychol. Psychiatry*, vol. 41, no. 8, pp. 1049–1056, 2000.

[46] A. Gelman, "Multilevel (hierarchical) modeling: What it can and cannot do," *Technometrics*, vol. 48, no. 3, pp. 432–435, 2006.

[47] J. de Leeuw and E. Meijer, *Handbook of Multilevel Analysis*. New York, NY, USA: Springer, 2008.

[48] T. A. B. Snijders, *Multilevel Analysis*. Berlin, Germany: Springer, 2011, pp. 879–882.

[49] J. A. O'Brien and G. M. Marakas, *Management Information Systems*. New York, NY, USA: McGraw-Hill, 2010.

[50] R. Agrawal, A. Gupta, and S. Sarawagi, "Modeling multidimensional databases," in *Proc. 13th Int. Conf. Data Eng.*, 1997, pp. 232–243.

[51] T. B. Nguyen, A. M. Tjoa, and R. R. Wagner, "An object oriented multidimensional data model for OLAP," in *Proc. 1st Int. Conf. Web-Age Inf. Manage. (WAIM)*, in Lecture Notes in Computer Science, vol. 1846. New York, NY, USA: Springer-Verlag, 2000, p. 69.

[52] A. Joly, P. Geurts, and L. Wehenkel, "Random forests with random projections of the output space for high dimensional multi-label classification," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2014, pp. 607–622.

[53] J. Levatić, D. Kocev, and S. Džeroski, "The importance of the label hierarchy in hierarchical multi-label classification," *J. Intell. Inf. Syst.*, vol. 45, no. 2, pp. 247–271, 2015.

[54] P. Szymański, T. Kajdanowicz, and K. Kersting, "How is a data-driven approach better than random choice in label space division for multi-label classification?" *Entropy*, vol. 18, no. 8, p. 282, Jun. 2016.

[55] G. Tsoumakas and I. Vlahavas, "Random k-Labelsets: An ensemble method for multilabel classification," in *Proc. 18th Eur. Conf. Mach. Learn.*, 2007, pp. 406–417.

[56] A. Joly, "Exploiting random projections and sparsity with random forests and gradient boosting methods—Application to multi-label and multi-output learning, random forest model compression and leveraging input sparsity," 2017, *arXiv:1704.08067*. [Online]. Available: https://arxiv.org/abs/1704.08067

[57] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[58] A. Verikas, A. Gelzinis, and M. Bacauskiene, "Mining data with random forests: A survey and results of new tests," *Pattern Recognit.*, vol. 44, no. 2, pp. 330–349, 2011.

[59] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, "Inferring regulatory networks from expression data using tree-based methods," *PLoS ONE*, vol. 5, no. 9, 2010, Art. no. e12776.

[60] M. Petković, D. Kocev, and S. Džeroski, "Feature ranking for multi-target regression," *Mach. Learn.*, pp. 1–26, Aug. 2019. doi: 10.1007/s10994-019-05829-8.

[61] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Effective and efficient multilabel classification in domains with large number of labels," in *Proc. ECML/PKDD Workshop Mining Multidimensional Data*, 2008, pp. 30–44.

[62] D. Kocev, "Ensembles for predicting structured outputs," Ph.D. dissertation, IPS Jožef Stefan, Ljubljana, Slovenia, 2011.

[63] I. Dimitrovski, D. Kocev, S. Loskovska, and S. Džeroski, "Fast and scalable image retrieval using predictive clustering trees," in *Proc. Int. Conf. Discovery Sci.*, 2013, pp. 33–48.

[64] L. Breiman, J. Friedman, R. Olshen, and C. J. Stone, *Classification and Regression Trees*. London, U.K.: Chapman & Hall, 1984.

[65] H. Blockeeel, S. Džeroski, and J. Grbović, "Simultaneous prediction of multiple chemical parameters of river water quality with TILDE," in *Proc. 3rd Eur. Conf. Princ. Data Mining Knowl. Discovery*, in Lecture Notes in Computer Science, vol. 1704. Berlin, Germany: Springer, 1999, pp. 32–40.

[66] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning," *Pattern Recognit.*, vol. 45, no. 9, pp. 3084–3104, 2012.

[67] E. V. Hatzikos, G. Tsoumakas, G. Tzanis, N. Bassiliades, and I. Vlahavas, "An empirical study on sea water quality prediction," *Knowl.-Based Syst.*, vol. 21, no. 6, pp. 471–478, 2008.

[68] W. Groves and M. Gini, "On optimizing airline ticket purchase timing," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 3, pp. 1–28, 2015.

[69] L. Todorovski, H. Blockeel, and S. Džeroski, "Ranking with predictive clustering trees," in *Proc. 13th Eur. Conf. Mach. Learn. (ECML)*, Helsinki, Finland, Aug. 2002, pp. 444–455.

[70] Kaggle. (2012). *Kaggle: Online Product Sales*. Accessed: May 5, 2017. [Online]. Available: https://www.kaggle.com/c/online-sales

[71] D. Kocev, A. Naumoski, K. Mitreski, S. Krtić, and S. Džeroski, "Learning habitat models for the diatom community in Lake Prespa," *Ecol. Model.*, vol. 221, no. 2, pp. 330–337, 2010.

[72] W. Groves and M. Gini, "Improving prediction in TAC SCM by integrating multivariate and temporal aspects via PLS regression," in *Agent-Mediated Electronic Commerce. Designing Trading Strategies and Mechanisms for Electronic Markets*. Berlin, Germany: Springer-Verlag, 2013, pp. 28–43.

[73] S. Džeroski, D. Demšar, and J. Grbović, "Predicting chemical parameters of river water quality from bioindicator data," *Appl. Intell.*, vol. 13, no. 1, pp. 7–17, 2000.

[74] M. Malinen and P. Fränti, "Balanced k-means for clustering," in *Proc. Joint Int. Workshop Struct., Syntactic, Stat. Pattern Recognit. (S+SSPR)*, in Lecture Notes in Computer Science, vol. 8621. Berlin, Germany: Springer, 2014, pp. 32–41.

[75] J. Collins, R. Arunachalam, N. Sadeh, J. Eriksson, N. Finne, and S. Janson, "The supply chain management game for the 2007 trading agent competition," Tech. Rep., 2006. [Online]. Available: https://www.sics.se/tac/tac07scmspec.pdf

[76] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Jan. 2006.

[77] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *Ann. Math. Statist.*, vol. 11, no. 1, pp. 86–92, 1940.

[78] R. L. Iman and J. M. Davenport, "Approximations of the critical region of the Friedman statistic," *Commun. Statist.-Theory Methods*, vol. 9, no. 6, pp. 571–595, 1980.

[79] P. B. Nemenyi, "Distribution-free multiple comparisons," Ph.D. dissertation, Princeton Univ., Princeton, NJ, USA, 1963.

[80] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, 1945.

STEVANCHE NIKOLOSKI received the M.Sc. degree in numerical optimization and mathematical modeling from the Department of Applied Mathematics, Faculty of Natural Sciences and Mathematics, Skopje, Macedonia. He is currently pursuing the Ph.D. degree with the Jožef Stefan International Postgraduate School. He is also a Walsh Fellow at Teagasc, Ireland. He was a member of the EU H2020 LANDMARK project. His current research topic is based on structuring the output spaces in multilabel classification and multitarget regression tasks, and application of the methods for structured output prediction in environmental domain. He has presented his work at and attended several workshops and conferences, such as PEDOMETRICS 2017 and ECML PKDD 2017.

DRAGI KOCEV received the Ph.D. degree in learning ensemble models for predicting structured outputs from the Jožef Stefan International Postgraduate School, in 2011. He was a Visiting Research Fellow with the University of Bari, Italy, in 2014 and 2015. He is currently a Researcher with the Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia. He has participated in several national Slovenian projects and the EU funded projects IQ and PHAGOSYS and is involved in the Human Brain Project. He was a Co-Coordinator of the FP7 FET Open project MAESTRA. He has been a member of the PC of many conferences (e.g., DS, ECML PKDD, AAAI, and IJCAI) and a member of the Editorial Board of *Data Mining and Knowledge Discovery* and *Ecological Informatics*. He served as the PC Co-Chair for DS 2014 and the Journal Track Co-Chair for ECML PKDD 2017.

SAŠO DŽEROSKI received the Ph.D. degree in computer science from the University of Ljubljana, in 1995. He is currently a Senior Researcher with the Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia, and a Professor with the Joef Stefan International Postgraduate School, Ljubljana. His research interests include artificial intelligence (AI), with a focus on the development of data mining and machine learning methods for a variety of tasks, including the prediction of structured outputs and the automated modeling of dynamic systems, and their applications to practical problems from science and engineering, e.g., environmental sciences (ecology) and life sciences (biomedicine). In 2008, he was an elected fellow of the European AI Society for his Pioneering Work in the field of AI and Outstanding Service for the European AI Community. In 2015, he became a Foreign Member of the Macedonian Academy of Sciences and Arts. In 2016, he was elected as a member of Academia Europaea.

Chapter 6

Multi-Target Prediction for Modeling Soil Functions

In this chapter, we present two case studies of applying multi-target prediction (Chapter 2) approaches to two different tasks of modeling soil functions. The first is related to modeling *primary productivity*, while the other is related to modeling a combination of the functions *water purification and regulation* and *regulation and provision of nutrients* (Chapter 4). The multi-target prediction approaches applied are those of trees and tree ensembles for multi-target regression, introduced in Chapter 3. The first case study uses the standard, fully supervised version of trees and tree-ensembles for multi-target regression. The second case study uses the semi-supervised variant of trees and tree-ensembles for multi-target regression.

An important property of trees and tree-ensembles for MTR is the interpretability of the produced models. In the case of individual trees, the models themselves can be inspected by domain experts and compared to existing domain knowledge. In the case of tree ensembles, feature rankings, based on the ensembles can be produced and inspected to obtain insight into the relative importance of the features for predicting the targets.

These two types of methods follow the new/emerging paradigm of explainable artificial intelligence. *Explainable AI* (XAI) (Edwards & Veale, 2017; Gunning et al., 2019; Samek et al., 2019) refers to the artificial intelligence methods and techniques that are producing explainable and interpretable solutions that can be understood by human experts. The XAI concept is in contrast with the existing 'black-box' concepts in machine learning where even the designers of the methods have difficulties to understand how the algorithms arrived at the specific solution. Thus, the domain experts have doubts about using such a solution in their decision-making process. XAI programs aim to produce models that are explainable to human experts, without significantly sacrificing the performance of the solution.

In both case studies presented in this chapter, we use data provided by Environment, Soils & Land-Use Department at TEAGASC, Ireland. Our first case study is related to the estimation of total grass yield potential and nutrient uptake in Irish dairy farms using soil (S), environmental (E) and management (M) data. Here, we use supervised PCTs for multi-target regression since all data is complete, i.e., all the values of the target variables are known. The second case study is related to predicting water quality parameters such as: biological water quality (Q-value), phosphorus (P) and nitrogen (N) concentration from existing pressure-pathway descriptive attributes. The main characteristic of this study, except the explainable modeling approach, is that we learn from 'incomplete' (i.e., partially-labeled) data, i.e., not all of the values for target attributes were measured (known) for each data instance. Unlike in the first case study, here, we use semi-supervised PCTs for multi-target regression, adapted for the semi-supervised learning task that can handle partially-labeled data. Moreover, in the second case study, we used supervised and semi-supervised ensembles of PCTs for MTR in order to create very accurate maps and improve upon the predictive performance of the single tree MTR models.

Our results, represented by explainable PCT models, are consistent with current findings in the application domain. Moreover, they show some insights and demonstrate potential findings of new knowledge from data.

6.1 Estimation of Herbage Production and Nutrient Uptake on Irish Dairy Farms

Maintaining productivity levels in grasslands is very important, since grass is one of the most important and cheapest feedstuffs for ruminants. The latter support high quality meat and milk production. Therefore, modeling nutrient management and grass yield production, to better understand the most important controllable (and non-controllable) factors in grazing grasslands, is one of the most important questions for the farming community.

An attempt towards modeling primary productivity has recently been made by Trajanov et al. (2018) within the LANDMARK H2020 project (LANDMARK, 2019), using data from Austrian fields. In this study, decision rules in the decision support (DS) model of primary productivity used in the Soil Navigator (Debeljak et al., 2019) were replaced with predictive rules derived from single-target regression models learned using data-driven methods. After the support that the DS model obtained from the data-driven tree models, some improvements in the estimation accuracy of the DS model were noted. However, in the above study of Trajanov et al. (2018), single-target models for predicting primary productivity were presented, since only one target variable was predicted.

In our study, we move beyond the existing studies of modeling primary productivity with single-target models and propose the use of machine learning modeling techniques for simultaneous prediction of multiple outputs, representing different aspects and outcomes of the primary productivity soil function. In particular, our derived multi-target models estimate nutrient (N, P, K) uptake and total herbage production using existing data from 15 commercial Irish dairy farms. The data consist of soil, environmental and management factors and measured data for P, N and K uptake by grass herbage as well as total herbage production.

The predictive modeling experiments were performed on four datasets (un)stratified by soil drainage factor. We learn PCT models on all fields taken together as well as on well-drained, somewhat-poorly, and poorly drained fields. As learning techniques, we use single PCTs and ensembles (random forests) of PCTs for multi-target regression.

Our results show that PCTs are accurate and easily explainable, provide enough information about the interactions between descriptive factors and are found to embody the existing understanding of the task at hand. If we combine more PCTs into an ensemble of PCTs (random forest of PCTs), we can achieve improved accuracy of the predictions. However, the ensembles can be used only for accurate predictions and for creating accurate maps, rather than to understand the interconnections between the descriptive attributes, because ensembles are non-interpretable. Moreover, in practical terms, one of the most important moderating factors that drives the total herbage production and nutrient uptake is the number of grazing events, which is closely related to the soil drainage class. Furthermore, we found that in the fields with medium yield potential, the nutrient (N, P, and K) uptake and herbage nutrient concentration are conservative, but nutrient uptake was more variable and potentially limiting in fields that had higher and lower herbage production. Our models also show that phosphorus is the most limiting nutrient for herbage production across the fields on these Irish dairy farms, followed by nitrogen and potassium.

The paper included in this section is:

- NIKOLOSKI, Stevanche, MURPHY, Philip, KOCEV, Dragi, DŽEROSKI, Sašo, WALL, David, P. (2019), Using machine learning to estimate herbage production and nutrient uptake on Irish dairy farms. *Journal of Dairy Science*, 102 (11): 10639-10656, doi:10.3168/jds.2019-16575.

The contributions of Stevanche Nikoloski to this paper are as follows. SN contributed to designing the experimental setting based on the experimental scenarios defined by DW and PM. He carried out the experiments and evaluated their results. He drafted the paper and revised it according to the co-authors and reviewers feedback to the paper.

J. Dairy Sci. 102:10639–10656
https://doi.org/10.3168/jds.2019-16575

Using machine learning to estimate herbage production and nutrient uptake on Irish dairy farms

Stevanche Nikoloski,[1,2]* Philip Murphy,[2] Dragi Kocev,[1,3] Sašo Džeroski,[1,3] and David P. Wall[2]
[1]Jožef Stefan International Postgraduate School, Jamova cesta 39, Ljubljana 1000, Slovenia
[2]Teagasc, Environment, Soils and Land-Use Department, Johnstown Castle, Co. Wexford, Y35 Ireland
[3]Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, Ljubljana 1000, Slovenia

ABSTRACT

Nutrient management on grazed grasslands is of critical importance to maintain productivity levels, as grass is the cheapest feed for ruminants and underpins these meat and milk production systems. Many attempts have been made to model the relationships between controllable (crop and soil fertility management) and noncontrollable influencing factors (weather, soil drainage) and nutrient/productivity levels. However, to the best of our knowledge not much research has been performed on modeling the interconnections between the influencing factors on one hand and nutrient uptake/herbage production on the other hand, by using data-driven modeling techniques. Our paper proposes to use predictive clustering trees (PCT) learned for building models on data from dairy farms in the Republic of Ireland. The PCT models show good accuracy in estimating herbage production and nutrient uptake. They are also interpretable and are found to embody knowledge that is in accordance with existing theoretical understanding of the task at hand. Moreover, if we combine more PCT into an ensemble of PCT (random forest of PCT), we can achieve improved accuracy of the estimates. In practical terms, the number of grazings, which is related proportionally with soil drainage class, is one of the most important factors that moderates the herbage production potential and nutrient uptake. Furthermore, we found the nutrient (N, P, and K) uptake and herbage nutrient concentration to be conservative in fields that had medium yield potential (11 t of dry matter per hectare on average), whereas nutrient uptake was more variable and potentially limiting in fields that had higher and lower herbage production. Our models also show that phosphorus is the most limiting nutrient for herbage production across the fields on these Irish dairy farms, followed by nitrogen and potassium.

Key words: nutrient uptake, herbage production, predictive clustering trees, random forest

INTRODUCTION

Grasslands make a significant contribution to food security through providing part of the feed requirements of ruminants used for meat and milk production. There is a renewed interest in grazing systems in many temperate and subtropical regions of the world. In Ireland, more than 90% of the agricultural area consists of pasture, grass silage or hay, and rough grazing (O'Mara, 2008). The utilization of grass by grazing should provide a sustainable basis for livestock production systems, as grazed grass is the cheapest source of nutrients for ruminants (O'Donovan et al., 2011). With feed cost accounting for more than 75% of the total variable costs on these livestock farms (Connolly et al., 2010), the production of sufficient grass for the grazing herd has a significant effect on farm profitability (Shalloo et al., 2004; Finneran et al., 2010). From 2013 to 2015, average levels of grass DM production on intensive dairy farms measuring grass in Ireland ranged from 8.0 to 18.5 t/ha (O'Leary et al., 2016). Grass production between and within farms can vary widely depending on several soil-, climate-, and management-related factors.

Potential herbage production on a farm system is the result of management practices in a given environment. Management practices are controllable factors and include crop management, soil fertility management, and sward composition. Environmental conditions are noncontrollable factors and include soil type/drainage and weather.

Most agricultural soils are treated periodically with fertilizers or organic manures and lime to correct mineral element deficiencies or toxicities and subsequently promote growth of grass. The nutrient management strategy practiced on grassland farms in Ireland is

Received March 5, 2019.
Accepted July 1, 2019.
*Corresponding author: stevanche.nikoloski@ijs.si

usually based on soil nutrient leaves and the stocking rate of the farm (Wall and Plunkett, 2016). Managing soil fertility levels closely, especially soil pH, can ensure that potential herbage production is not being limited by nutrient availability in the soil. Nitrogen is often a main limiting nutrient in temperate soils. Intensively managed grazed grasslands generally receive multiple applications of fertilizer N during the growing season to increase the forage available to grazing animals. Losses of N and P from such intensively managed systems have also come under scrutiny due to their effect on water quality, air quality, acidification, and anthropogenic climate change (Dillon and Delaby, 2009). As a result of such concerns, restrictions on fertilizer use for grasslands were implemented in many parts of Europe under the European Union Nitrate Directive (Nitrate Directive, 1991). This combination of economic and environmental factors makes improved efficiency of fertilizer N use central to any strategy for sustainable grassland production systems.

Fertilizer input in grazed grasslands is usually linked to the grazing schedule, with fertilizer N in particular applied after each grazing. Recovery of N, P, and K in grass herbage can be highly variable depending on the date and rate of application (Vellinga et al., 2010). In Ireland, swards are typically grazed at intervals of between 21 and 28 d during most of the grazing season and there may be a carry-over effect of nutrient applications and deposition by the grazing animals to the following growth interval. Herbage production and nutrient recovery are also affected by site factors such as soil type. Soil N mineralization rates can vary considerably, both seasonally and between soils (Herlihy, 1979; Nunan et al., 2000), and contribute a significant proportion of N to grass growth on farms (Humphreys et al., 2008).

However, emphasis is currently being put on achieving a high number of grazings per field, as this is closely correlated with herbage production (Hanrahan et al., 2017). The rate of reseeding currently practiced in Ireland is low (Creighton et al., 2011). Sward composition is an important contributor to potential yield also. A perennial ryegrass (*Lolium perenne* L.) dominated sward is likely to produce higher yield than mixed grass species swards (Smith and Allcock, 1985; Ergon et al., 2016). A sward that incorporates white clover (*Trifolium repens* L.) can be managed to offset N fertilizer inputs due to its N fixing capacity.

Soil type and soil drainage class refers to the physical, chemical, and biological soil characteristics. The sum of these properties is a very important factor that contributes to the potential herbage production, but is commonly overlooked. Weather, specifically rainfall, is a major limiting factor in the implementation of any agronomic strategy.

Modeling herbage production, as well as nutrient uptake, is difficult because of the number of environmental and management factors that affect the final result. The farm system is made up of multiple moving parts. It can be very difficult to implement a practice change and expect to achieve an isolated and easily measurable difference. The model can consider multiple factors acting together (i.e., multiple moving parts). Potential yield depends on factors associated with the site in question. It is better to consider all the factors together, or as many of them as possible, although this could be impractical because of economic and time constraints.

This complexity of the farm system is the reason why research often uses component type studies. Schils et al. (2007) developed a whole-farm dairy simulation model, called DairyWise, which simulates some environmental, technical, and economic processes on a dairy farm. The DairyWise model is evaluated using 2 data sets consisting of 29 dairy farms. As output, this model provides a farm plan describing all nutrient flows, as well as the consequences to the environment and economy. The outputs of DairyWise model components are further used as inputs in other environmental, economic, or technical sub-models.

Plot-, field-, or farmlet-scale studies provide the insight needed to address a problem, but they does not give the real picture of possible trade-offs and synergies between the controlling factors at hand. For that reason, in this study, we rely on a data-driven approach to modeling. We use predictive clustering trees (**PCT**; Blockeel et al., 1998), which are a generalization of decision trees, adapted for structured output prediction tasks. So far, PCT have been applied in many different environmental domains, for instance, for predicting the abundance of different species occupying the same habitat (Demšar et al., 2006), estimating different vegetation quality indices for the same site (Kocev et al., 2009), or predicting the composition of a community of organisms (Levatić et al., 2015a).

Predictive clustering trees can consider multiple factors acting together (i.e., multiple moving parts) and can also deal with multiple targets (responses), where the task at hand is called multi-target regression (**MTR**). The MTR task is to predict/estimate the values of multiple targets simultaneously and PCT solve this task by building one predictive model for all of the targets. Recent research shows that PCT are superior to most of the state-of-the-art machine learning algorithms for MTR (Kocev et al., 2009). Furthermore, trying to improve the predictive performance of a single

PCT, Kocev et al. (2013) proposed to combine a set of single (base) predictive models into an ensemble of tree models. For basic regression tasks, it is widely accepted that ensemble model learners improve over the predictive performance of single-tree learners (Kocev et al., 2013).

In this study, we use PCT for MTR, as well as ensembles (random forests) of PCT to estimate herbage production potential and N, P, and K uptake by using soil, environmental, and management attributes. Individual PCT are interpretable and can be used for the visualization of input variable interactions and the dependency of the target thereof. The lack of interpretability is the main drawback of ensemble learners because overall predictions of ensemble are the average of the predictions from each tree in the ensemble. We use the PCT to obtain insight into the domain of study and ensemble of PCT to obtain estimates of herbage production and nutrient uptake. The latter can be used, for example, to create accurate maps for the response(s) of interest.

This study aims to address the question of why herbage production potential can differ greatly between regions and even farms. We have a data set for 15 dairy farms in Ireland, where a range of soil (**S**), environmental and weather (**E**), and management (**M**) variables has been measured (see Appendix Table A1).

The goal of this study was to address the following research questions related to herbage production and nutrient uptake on grazed grassland-based dairy farms in temperate regions:

- What are the main drivers of herbage production on grassland-based dairy farming systems?
- How do nutrient supply and nutrient (N, P, and K) uptake affect herbage production?
- How do S, E, and M variables interact within a grazed grassland farming system to affect herbage production and herbage nutrient uptake?

The remainder of this paper is organized into the following sections. The Materials and Methods section describes the data we used in our experiments, as well as the way in which it was collected. In this section, we also present the machine learning methodology we used for building the models and specify the design of the machine learning experiments. The Results and Discussion section presents and discusses the obtained models (trees and ensembles) in terms of their predictive performance and interpretability. Finally, we present the conclusions of this work and outline its implications and potential outcomes for advisory services and grassland management on dairy farms.

MATERIALS AND METHODS

Data Description and Collection

A scoping process was carried out with several advisors across 3 counties (Wexford, Cork, and Tipperary), which led to the selection of 15 commercial Irish dairy farms.

All of these farms are specialized dairy farms and were selected based on farmer willingness to adopt new practices and have good record-keeping skills. The final selection included production intensity and soil drainage differences so that a range in each category would be captured. It must be noted that this approach to selection may bias the results toward more progressive farmers who farm in the south and south-east of Ireland.

In terms of milk delivered and concentrate per cow, this cohort was slightly above the national average of 861 kg per cow in 2015 (Hennessy and Moran, 2015). However, this selected group of farms is representative of main intensive dairy regions of Ireland. It was expected that many dairy farms would expand or intensify post milk quota abolition in March 2015. In 2015, 11 out of the 15 farms had a derogation to farm more intensively (i.e., stocking rates between 170 and up to a maximum of 250 kg of organic N/ha) and in 2016 all 15 farms were in this more intensive stocking rate category.

General Farm System and Soil Data Collection

Management (controllable factors) and environmental (noncontrollable factors) data were collected for 804 fields on the 15 farms for 2 yr (2015 and 2016). A detailed description for each factor is given in Apppendix Table A1. Information on how the data were collected is given below.

General biophysical, farm system, and management activity data were collected by visiting each farm 3 times per year. During these visits, information such as the number of fields and paddocks, area of individual fields, area used for grazing the dairy herd, duration of periods that the livestock are grazing versus indoors, slurry production system and quantity, and grazing infrastructure and grassland management (i.e., areas used for grazing vs. silage) were recorded and further verified by repeating pertinent questions during subsequent visits. A survey of the soils (general soil classification using the Irish Soil Information System (Simo et al., 2008) and ground-truthing using soil auguring, field orography (aspect, topography), and sward composition on each farm was conducted during 2015.

Nutrient and Management Activity Data Collection

The farms in the study recorded nutrient use and grassland management at the field or paddock scale using an online software package PastureBase Ireland (Hanrahan et al., 2017; Teagasc, 2019). Some farmers choose to keep written records of fertilizer and manure applications and other field management information, such as reseeding and grazing events. The accuracy of record keeping was improved by sending monthly text message reminders to each farmer participant over the 2-yr period to visit the farms quarterly to record any missing information. For all farms, at least the following details were collected: field name and area, fertilizer type (chemical fertilizer type, organic manure type, soiled water, lime, or other), quantity applied (kg or t/ha) and date of application, and number of grazings per field or paddock area. Total concentrates imported and organic N stocking rate data were collected from the fertilizer plans developed by the advisor and farmer for each individual farm. These records were collected annually on site or downloaded from the online software package used by the farmer. To maintain consistency, all records were downloaded or transcribed individually and structured before analysis. Total milk sold from the farm and cow herd size data were collected online from the Irish Cattle Breeding Federation website (ICBF, 2018).

Herbage Production and Accumulation

Total herbage production and annual herbage accumulation was recorded by the farmer on a per field basis throughout the growing season using a sward cut and weigh technique or a calibrated falling plate meter (Li et al., 1998; Smit et al., 2005). Farmers were asked to carry out weekly pasture measurements on each field or paddock on the main grazing area used for the dairy herd. These measurements were entered into the PastureBase Ireland software, which calculated a grass feed budget and the total quantity of grass grown and accumulated annually. At the end of each year, the annual herbage accumulation (kg of DM/ha) corresponding to each field or paddock on each farm was downloaded from the PastureBase Ireland software.

Herbage Nutrient Concentration

On each farm, herbage samples were taken from all fields/paddocks at 3 times over the growing season, corresponding to spring, summer, and autumn, to determine macro- and micro-nutrient concentrations in the herbage DM. On each sampling occasion, a 0.5 m × 0.5 m area was randomly selected at 3 locations moving down the long axis of each field, an adaptation of the approach of Sheridan et al., (2008). The herbage was sampled from all 3 areas, using electronic grass shears to a height of 4 cm, as would be typical of grazing conditions. The samples from the 3 areas were bulked and a subsample was taken for nutrient testing in the laboratory. The subsample was oven-dried for 48 h at 40°C and following this was ground to pass through a 1-mm mesh in preparation for chemical analysis.

Herbage nutrient concentrations (g/kg of DM) were determined in the laboratory as follows: herbage N concentration was determined by C & N analyzer (Leco Corporation, St. Joseph, MI). Major nutrients (g/kg of DM) such as P, K, and Mg were determined by inductively coupled plasma atomic emission spectroscopy following hot acid (HNO_3) digestion and following the method by Byrne (1979). Pasture nutrient uptake was calculated for each field during the spring, summer, and autumn periods of 2015 and 2016 by multiplying the herbage DM produced (kg of DM/ha) during each period by the measured herbage nutrient concentration (g/kg of DM) for the period. Total annual herbage nutrient uptake was expressed as kg N, P, or K per hectare by summing the nutrient uptake values for each of the 3 periods in 2015 and 2016.

Machine Learning Methods

To estimate herbage production and nutrient uptake from soil, environmental, and management variables, we applied machine learning methods to the data described above. In particular, we used PCT to capture and visually represent the dependencies between the input variables and the response variables, where the latter are considered both individually (single-target PCT) or jointly (multi-target PCT). Moreover, to get more accurate estimates, we used ensemble of PCT (i.e., random forests of PCT). In this subsection, we will present in detail the methodology used for building the PCT models.

Predictive Clustering Trees. Predictive clustering trees are obtained by using the well-known top-down induction of decision trees (**TDIDT**) algorithm (Blockeel et al., 1998). The TDIDT takes a set of examples as input and produces a tree model as output. At the beginning, the TDITD procedure selects test on an attribute (independent variable) for the top node, by using a heuristic function computed on the training examples. The heuristic function favors tests that partition the data so that the examples that go to one branch/cluster (tree node) are as similar as possible. To increase cluster homogeneity, heuristic function chooses the partition that maximally reduces the inhomogeneity, as measured by the variance function. The partitioning procedure continues to recursively split the examples in each subset of resulting partition until a stopping criterion is satisfied. The stopping criterion prevents the tree from overfitting to the training data at hand. When the stopping criterion is met, examples are not split further. A representative value (i.e., prototype) is calculated for the response variables is are stored in the corresponding leaf of the tree (as a prediction).

Two main functions define the algorithm for learning PCT, the variance function, and the prototype function, which computes a representative prediction value for each leaf.

In PCT, both functions can handle multiple response variables, as is required in MTR. This is the main difference between PCT and standard decision trees. The PCT are implemented in the CLUS system (https://sourceforge.net/projects/clus/). For further information on PCT, we refer the reader to Kocev et al. (2013).

Several known tree pruning (stopping) criteria are known that can be used to prevent overfitting, such as specifying the minimal number of examples that must be present in each leaf of a tree and performing *F*-test pruning, which checks whether a given split yields a significant reduction of its variance. The use of pruning methods typically increases a tree's interpretability and improves its predictive performance (accuracy).

The advantageous properties of PCT are inherited from standard decision trees. In PCT, no assumptions are made on the probability distributions of the independent and response variables. The PCT can handle missing values in both the independent and the response variables and are tolerant to noisy and redundant variables as well. Moreover, PCT work with different type of both input and response variables, such as discrete or continuous. Furthermore, they are computationally inexpensive to learn and very easy to interpret. While constructing clusters, PCT also produce cluster descriptions. Hence, PCT are readily interpretable, efficient and robust, and have satisfactory predictive performance.

Random Forests of PCT. Random forests of PCT (Kocev et al., 2013) is an ensemble learning method also implemented in the CLUS system. They are constructed by using the PCT learning algorithm in CLUS, modified to follow the random forest method proposed by Breiman (2001). The forest of trees is built by using different bootstrap replicates of the training data and by using a randomized version of the PCT learning algorithm that changes the space of input variables dynamically during the learning process. Bootstrap replicates are generated by random sampling of examples from the training set with replacement, until the same number of examples as in the original training set is sampled. In the random forest algorithm, there is a random selection of input variables (attributes); that is, at each node, a random subset of attributes is taken from the descriptive space D and the best split selected among those is used at the given node. There are different ways of setting the number of randomly selected descriptive attributes $\{\mathrm{f}(D) = 1$, $\mathrm{f}(D) = [\mathrm{sqrt}(D)] + 1$, $\mathrm{f}(D) = [\log_2(D)] + 1$, and so on$\}$. The response value predictions for a new instance in a random forest of PCT are calculated by combining the predictions from all base predictive models. In the MTR task, the prediction for each target is defined as the average of the predictions obtained from each PCT.

Design of Machine Learning Experiments

To obtain insights into the most influential factors driving the herbage production potential of grasslands on dairy farms and finding some potentially new knowledge from data collected from such farms, we investigate 4 different scenarios. Over time the interaction of intrinsic soil factors and environment factors creates stable soil environments that can be categorized by soil drainage class. To further explore the influence of field management on herbage production, we split the original data set into 3 data sets based on a different drainage class to further isolate these management effects in the model analysis. Hence, we used 4 data sets for analysis:

- complete data set (**CD**): consists of all 804 examples, CD = WD ∪ SPD ∪ PD;
- well-drained data set (**WD**): consists of 606 examples that belong to well-drained soil samples;
- somewhat poorly drained data set (**SPD**): consists of 122 examples that belong to somewhat poorly drained soil samples; and
- poorly drained data set (**PD**): consists of 76 examples that belong to poorly drained soil samples.

In our machine learning experiments, we learned PCT and random forests of PCT from the above data sets. In particular, we learned single-target PCT to estimate herbage production from S, E, and M variables and from nutrient uptake. We also learned multi-target trees to estimate the 4 response variables (herbage production and N/P/K uptake) for each of the 4 data sets.

When learning single PCT, we used the *F*-test pruning algorithm with 8 significance levels: 0.001, 0.005, 0.01, 0.05, 0.1, 0.125, 0.25, and 1.0. By using internal 3-fold cross validation, the optimal significance level was chosen that minimizes the evaluation measure. Besides *F*-test pruning, we also used different values for the parameter minimum number of instances per leaf in all different scenarios. Namely, in CD and WD, we specified 32 as the minimum number of instances in a leaf, since we have a larger number of instances. Then, in SPD and PD data sets we specified 8 as the minimum number of instances in a leaf, since SPD and PD are data sets with a smaller number of examples. To obtain the tree where N, P, and K uptake are used as descriptors to estimate herbage production potential, we used single PCT with a minimum of 64 examples in a leaf with *F*-test pruning with a 1.0 significance level (default setting).

In the ensemble setting, that is, when using random forests of PCT, we set $f(D) = [\text{sqrt}(D)] + 1$ as the number of randomly chosen attributes from the descriptive space *D*. Moreover, we set 100 as the number of base-level models (PCT) in the ensemble.

We evaluated the models by using 10-fold cross validation. More specifically, the whole data set was first randomly split into 10 folds. Next, 9 folds were used for training, and the remaining one for testing. The second step in the procedure was repeated 10 times, so that each fold is used exactly once as a test set. The reported results represent the average of all 10 runs.

For assessing the performance of machine learning algorithms, several empirical evaluation measures can be used. In our experiments, we employ 2 well-known measures: the Pearson correlation coefficient ($\mathbf{r^2}$) and relative root mean square error (**RRMSE**). The values of r^2 range between 0 and 1. Perfect correlation is obtained when $r^2 = 1$. Therefore, the closer to 1 the value of r^2 is, the better performance is achieved (higher r^2 is better). The RRMSE relates the average magnitude of the error (differences between predictions and actual observations) to the error made by the default predictive model, predicting the mean of the observed values. The best performance in terms of RRMSE is obtained if the value of RRMSE is 0 (lower RRMSE values are better).

Our experiments were performed in the PCT framework. The PCT framework is implemented in the CLUS system (Blockeel and Struyf, 2002), which is written in Java and is open source software licensed under the GNU General Public License. The CLUS system is available for download at http://clus.sourceforge.net/projects/clus/.

RESULTS AND DISCUSSION

In this section, we provide interpretations of the obtained trees and discuss them in the context of the research questions defined in the introductory section.

What Are the Main Drivers of Herbage Production on Grassland-Based Dairy Farming Systems?

We used single-target regression PCT to estimate herbage production using the available S, E, and M variables to investigate the main drivers of herbage production on Irish dairy farms. We started with an interpretation of the tree that estimates the herbage production potential given the S, E, and M attributes, [i.e., herbage production = f (S, E, M), which means that the created trees construct the f function that outputs the predictions for herbage production and uses the S, E, and M attributes as an input] learned from the entire data set (Figure 1).

The model in Figure 1 selects the number of grazings (NoGrazings) as the top descriptor related to the total annual herbage accumulation (grazing and silage). It

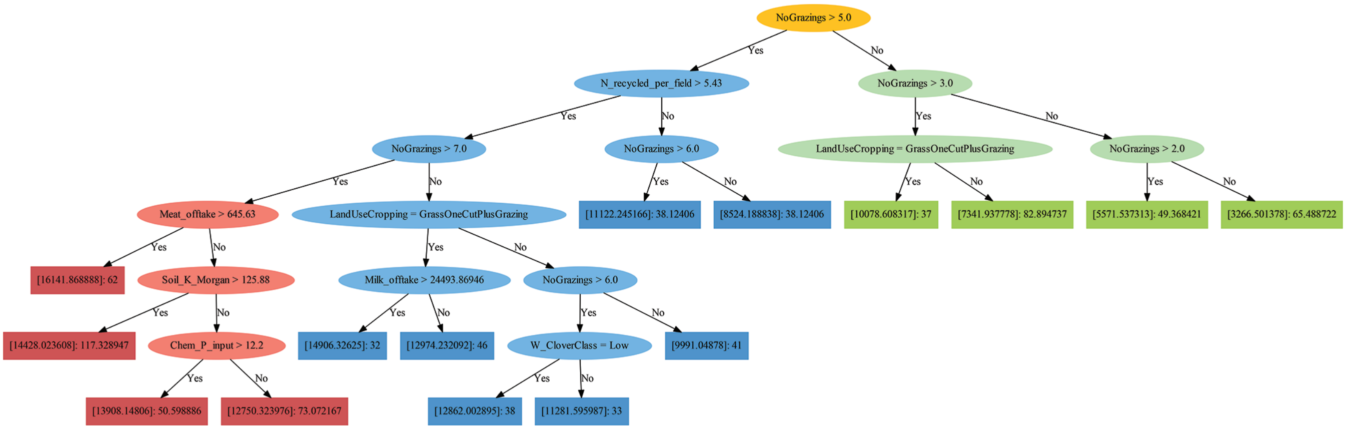


Figure 1. Single-target regression tree for estimating herbage production [i.e., herbage production = f (S, E, M)]. S = soil; E = environment and weather; and M = management. Colored nodes are related to 3 different categories: category 1 (red): high herbage production potential; category 2 (blue): medium herbage production potential; and category 3 (green): low herbage production potential.

is logical that nutrient uptake and herbage accumulation will increase with the number of grazing events (NoGrazings). Where herbage production is higher, there is more biomass for the grazing animals to eat and to support grazing events more frequently over the growing season. The majority of grazing events were preceded by an application of chemical fertilizer, mainly N, which is a farm management factor promoting increased herbage production. Additionally, during every grazing, some of the nutrients in the herbage consumed were recycled back to the soil in the form of nutrient excretion and deposition (dung and urine) by the grazing livestock, which is not a farm management factor, but a natural process. Overall, the organic and inorganic nutrient inputs, coupled with plant growth stimulation (tillering) through grazing events, led to increased herbage production and nutrient uptake.

Table 1 shows the proportion of well-drained fields in each NoGrazings category identified by the model. Table 1 indicates that the number of grazing events per field is also related to the soil drainage class. The number of fields in the well-drained class decreased as the number of grazings decreased (see Table 1 and Figure 1). The proportion of well-drained fields (WD/Σ) decreases with NoGrazings, as we examine the model (PCT) from the left to the right. We have 82.4% of well-drained sites for NoGrazings > 7 and only 36.8% for NoGrazings < 3.

Therefore, the model also captures to some degree the biophysical constraints on grazing events and separates the explanation of total annual herbage accumulation into 3 main categories: high (14,300 kg of DM/ha on average), medium (11,600 kg of DM/ha on average), and low (6,300 kg of DM/ha on average), strongly related to drainage class differences. In general, the model indicates high annual herbage accumulation potential on well-drained soils, medium accumulation potential on somewhat poorly drained soils, and low accumulation potential on poorly drained soils.

Next we interpret the parts of the model (PCT in Figure 1) that correspond to each of the 3 categories (i.e., to fields with different herbage accumulation potential).

Category 1: Fields with High Herbage Accumulation Potential. At the main node of this component, we find the test Meat_offtake >645.63 kg/ha. Meat offtake is directly related to the stocking rate per hectare, as each cow produces a calf and will gain BW as they mature. In addition, the greater the stocking rate, the greater the nutrient recycling under grazing management as the excess N and K, in particular, are excreted in dung and urine. However, increased nutrient offtake in meat may also affect the nutrient balance and can typically lead to deficits in P where fertilizer P inputs are low (Buckley et al., 2016). For these fields with high herbage accumulation potential, the model shows that larger Meat_offtake is positively related to higher herbage accumulation. Fields with high stocking rates typically receive high inputs of inorganic and organic fertilizer inputs seasonally to boost herbage production rates. This Meat_offtake >645.63 kg/ha threshold is very high and indicates fields with very high stocking rate, much higher than the average for this study group of farms.

Next, the model splits the fields samples based on the attribute Soil_K_Morgan >125.88 mg/L. In this case, the model chooses soil K fertility to discriminate between the herbage accumulation potential of different fields. We found that, in general, Soil_K_Morgan was positively related to soil P and Mg fertility and can be viewed as a proxy for soil fertility levels. When Soil_K_Morgan >125.88 mg/L, then, on average, P and Mg were higher too (i.e., average soil fertility was higher). When Soil_K_Morgan ≤125.88 mg/L, the average P and Mg were lower (i.e., the average soil fertility was lower). Lower soil fertility results in lower herbage production (Wall and Plunkett, 2016), which is also indicated by the model. The Mg, K, and P are all essential nutrients for optimum soil fertility for grass production. Each of them is determined by taking a soil sample from the field and chemically testing them for their nutrient concentration. According to the soil

Table 1. The percentage of samples in different drainage classes calculated according to the intervals of number of grazings that appear in the predictive clustering tree model in Figure 1 (from left to right)

	Drainage class						
Item[1]	Well-drained (WD)	Somewhat poorly drained (SPD)	Poorly drained (PD)	Σ (%)	WD/Σ (%)	SPD/Σ (%)	PD/Σ (%)
X > 7	266	41	16	323	82.4	12.7	5.0
5 < X < 7	216	24	2	242	89.3	9.9	0.8
3 < X < 5	76	22	21	119	63.9	18.5	17.6
X < 3	42	35	37	114	36.8	30.7	32.5

[1] X = no. of grazings.

index system for K in mineral soils in Ireland, a value above 100 mg/L is considered agronomically optimal for grass production (Wall and Plunkett, 2016). The model identified a Soil_K_Morgan ≤125.88 mg/L threshold, slightly above the level deemed to be optimal by Wall and Plunkett (2016). However, on very high yielding pastures, the K requirement is also high, and under such circumstances, the short-term K supply capacity for plant uptake may be limited in soils with high levels of K fertility.

Next, the model splits fields based on the test Chem_P_input >12.2 kg/ha. Fields with higher chemical P inputs had higher herbage accumulation over the growing season. Chemical fertilizer P is readily available for plant uptake and is typically applied in several fertilizer applications during the growing season to meet the seasonal growth requirements of the grass. Chemical fertilizer P inputs were proportionally the largest of total P inputs on these dairy farms. The majority of the remaining P inputs came from organic manure P inputs, with some concentrate feed-derived P inputs entering the grazing system through parlor feeding at milking time. The model selected a Chem_P_input threshold of 12.2 kg/ha to discriminate between fields of different herbage production potential. This threshold P input value is slightly less than the chemical P input requirements for soil P fertility maintenance for grazing only fields on dairy farms (14–19 kg of P/ha; Wall and Plunkett, 2016) and indicates that fields with Chem_P_inputs lower that the threshold of 12.2 kg/ha were likely to be in P deficit (i.e., mining P from the soil over time). This situation is likely to affect soil P fertility and negatively affect herbage production as indicated by the model.

Category 2: Fields with Medium Herbage Accumulation Potential. For this component of the model tree, N_recycled_per_field >5.43 kg of organic N/ha was selected as the test for selecting fields, with higher N recycling, having a higher average herbage accumulation as compared with to those with less N recycled. Nitrogen recycled refers to the quantities of N excretion (dung and urine) the cow recycles back to the field during grazing throughout the year. The nutrients available for recycling are left over after the cow metabolizes nutrients for milk and meat production first. In a grazed grassland farming system, dung and urine patches contain very high concentrations of N. While the patches are not distributed evenly across the field and lead to heterogeneous soil mineral N levels, if the fields are grazed more often or the stocking rate is relatively high, it is likely that the density of urine and dung patches per unit area grazed will be higher, thus contributing more nutrients across the area to drive grass production. On closer investigation, the fields identified by the model with N_recycled_per_field >5.43 kg of organic N/ha corresponded to farms with an average dairy grazing platform stocking rate greater than 170 kg of organic N/ha (i.e., 2 livestock units/ha) and were predominantly grazed. This indicates that fields with N_recycled_per_field >5.43 kg of organic N/ha are associated with higher stocking rates, are predominantly grazed, and have more N excretion compared with fields with very low N recycled (<5.43 kg of organic N/ha).

At the next level of the tree, the test selected was LandUseCropping = GrassOneCutPlusGrazing. While taking a cut of silage removes nutrients in the harvested grass, increased fertilizer applications are used on fields selected for grass silage production and may drive increased herbage production. Typically, 2 fertilizing scenarios arise: (1) Fertilizer added pre-harvest: the field was managed to provide sufficient fertilizer inputs to produce enough herbage biomass for a silage cut. During this period of the year, typically early summer, higher qualities of fertilizer were applied to these fields compared with fields that are used for grazing only. (2) Fertilizer added post-silage harvest: after the herbage biomass has been harvested for silage, fertilizer was added to the system to ensure the grassland was adequately supplied with nutrients to recover after the cutting and harvesting event. In addition, grassland that is managed for silage production has higher yield potential as the plants grow to a more mature stage up to harvesting time, where they can intercept more light for photosynthesis and the total herbage biomass is greater compared with typical cumulative grazing biomass yield for the same period. As a result, the herbage biomass used for silage had higher nutrient concentration and total nutrient removal from the soil.

Next the Milk_offtake >24,493 L/ha, which was the total amount of produced milk for each field, was used to split the herbage accumulation of the fields into 2 groups. Those with higher milk offtake had a mean herbage accumulation of 14,906 kg of DM/ha and those with lower milk offtake had a mean of 12,974 kg of DM/ha. This splitting condition Milk_offtake distinguishes between higher and lower stocking rates. The average milk yield per cow in Ireland was 5,036 kg of milk per cow between 2013 and 2015 (Teagasc, 2016) and this threshold of Milk_offtake >24,493 L/ha indicates that, on average, the stocking rate on the dairy grazing platform fields above this milk offtake threshold was 4.75 cows per ha.

At the bottom level of the tree, the test W_CloverClass = Low best discriminated herbage production across the remaining fields that had the lowest herbage production in the medium herbage production category. The presence of white clover is expected to contribute a source of N for grass uptake. However, in this case, the model identifies high levels of white clover as an herbage production limiting factor (i.e., fields with white clover had lower overall annual herbage accumulation). According to Murphy et al. (2018), fields with high levels of white clover present also received high levels of fertilizer N inputs. Under such practice these fields with white clover present will not efficiently use the fixing capacity of the white clover for total herbage production. Fields with a high sward composition for clover received the highest average chemical N input (236 kg of N/ha). However, these fields had lower annual herbage accumulation (11,200 kg of DM/ha) compared with fields with low sward clover composition (12,800 kg of DM/ha) with slightly lower average chemical N input (228 kg of N/ha). This may be a result of too much clover in the grass sward having a negative effect on perennial ryegrass growth because of shading and competition for nutrients and water, thus reducing the total annual herbage accumulation.

Category 3: Fields with Low Herbage Accumulation Potential. Examining the fields with lower overall herbage production potential in the tree (on the right side of Figure 1), we can see that only 2 attributes are used for splitting: NoGrazings and LandUseCropping. The number of grazings on this side of the tree is very low (<3) and the fields within this node had a very high proportion of low drained sites. This indicates that soil drainage was an overriding factor limiting grassland management and herbage production, which was on average 6,264 kg of DM/ha for all the fields in this category.

Next, we discuss the descriptive and predictive performance of the models. We consider the descriptive (training) and predictive (testing) performance of single PCT and ensemble of PCT (random forest) for single-target regression and MTR tasks. All results considering the model performances for all possible scenarios, based on the CD, are shown in Table 2. The results consist of per target values and averaged values for r^2 and RRMSE.

We can see that the (pruned) model (i.e., single PCT), shown in Figure 1 and obtained using *F*-test pruning, has a descriptive performance of $r^2 = 0.766$ and RRMSE = 0.484. We show the descriptive performance of the original nonpruned model as well ($r^2 = 0.777$ and RRMSE = 0.472) to compare how good the *F*-test pruning method is. We can see that the descriptive performances are very similar (i.e., the difference is only 1%). This is an advantage of this pruning method because we do not use the original model and avoid the possibility of overfitting.

The predictive performance estimate obtained using by 10-fold cross validation is $r^2 = 0.715$ and RRMSE

Table 2. Performance results for the complete data set achieved by predictive clustering trees (PCT) for single- and multi-target regression[1]

		Regression trees (PCT) built with *F*-test pruning								Random forest of PCT	
		Training performance				Testing performance				Testing performance	
	Optimal *F*-test significance level	Unpruned tree		Pruned tree		Unpruned tree		Pruned tree		Forest with 100 trees	
Response variable		r^2	RRMSE	r^2	RRMSE	r^2	RRMSE	r^2	RRMSE	r^2	RRMSE
Single-target regression											
Total_herbage_production (herbage production)	0.05	0.777	0.472	0.766	0.484	0.715	0.535	0.715	0.535	0.798	0.477
Total_Herbage_N_uptake (N)	0.1	0.751	0.499	0.738	0.512	0.684	0.563	0.678	0.569	0.779	0.494
Total_Herbage_P_uptake (P)	0.001	0.749	0.501	0.744	0.506	0.691	0.556	0.693	0.554	0.768	0.508
Total_Herbage_K_uptake (K)	0.005	0.743	0.507	0.741	0.509	0.696	0.551	0.692	0.555	0.753	0.525
Average (herbage production, N, P, K)		0.755	0.495	0.747	0.503	0.697	0.551	0.694	0.553	0.775	0.501
Multi-target regression											
Herbage production + N + P + K	0.005										
Herbage production		0.772	0.478	0.767	0.483	0.720	0.529	0.719	0.530	0.795	0.478
N		0.740	0.510	0.733	0.516	0.691	0.557	0.691	0.557	0.789	0.483
P		0.736	0.514	0.728	0.521	0.686	0.560	0.684	0.562	0.761	0.514
K		0.712	0.536	0.703	0.545	0.647	0.595	0.643	0.598	0.761	0.515
Average (herbage production, N, P, K)		0.740	0.510	0.733	0.516	0.686	0.560	0.684	0.562	0.777	0.498

[1] RRMSE = relative root mean square error; r^2 = Pearson correlation coefficient.

= 0.5346. This is a quite good predictive performance, considering the problem complexity and domain diverseness. Observing the results obtained by using ensembles of PCT (i.e., random forests of PCT), as expected, we can see an improvement (approximately 8%) for the predictive performance (r^2 = 0.798 and RRMSE = 0.477). The problem with ensembles from a domain expert point of view is their interpretability. The random forests cannot be interpreted, but can be easily used if the domain expert is only interested in the accurate predictions (for example, for drawing an accurate map). In our case, we are also interested in elucidating the interactions and interconnections among the variables (attributes).

How Do Nutrient Supply and Nutrient (N, P, and K) Uptake Affect Herbage Production?

To investigate this question, we model herbage accumulation as a response variable, using nutrient uptake for each field as an independent variable, to investigate how the herbage production is driven by soil fertility and nutrient supply. The nutrient uptake value represents the nutrients that the grassland herbage has removed from the soil and will be consumed and utilized by the cow. The cow requires these nutrients to produce milk and meat. After this, the cow excretes the residual nutrients, not used for production, in dung and urine. While we expect nutrient uptake to be closely related to the annual herbage production, we propose this modeling analysis to identify the most limiting nutrient influence on herbage production. In Figure 2, we examine how the herbage production levels across the grassland dairy fields were categorized based on their nutrient (N, P, and K) uptake. The purpose of Figure 2 is to present how herbage production is driven by nutrient uptake rather to make predictions. Nutrient uptake variables are considered to be response variables in the remaining analyses.

The tree clearly indicates that Total_Herbage_P_uptake is the most limiting nutrient driving herbage production, since it appears in the top levels of the tree. The tree is generated by using the PCT algorithm, where the descriptors (independent variables) are total herbage N, P, and K uptake and annual herbage accumulation is the response. The next most important driver was Total_Herbage_N_uptake that appears in the third level of the tree and is followed by Total_Herbage_K_uptake in the fourth level. This model provides new insight into the obvious interconnection between these nutrients and how they relate at different herbage production levels, where the supply and uptake for these nutrients vary.

We further investigated the nutrient (N, P, and K) uptake levels in the herbage produced as per the groupings (leaf nodes) of fields identified by the PCT in Figure 1 and their relation to the N:P ratio, as shown in Figure 3. Each panel depicts on the y-axis one of the 4 response variables, whereas the N:P ratio is shown on the x-axis. Each point corresponds to one of the leaves of the PCT in Figure 1.

Figure 3A shows a weak positive relationship (r^2 = 0.197) between the herbage accumulation and the N:P ratio for these field groupings. For the majority of these field groupings, N, P, and K appear to be sufficiently supplied. However, variability exists between them and when the average nutrient concentrations in the herbage measure in the herbage form fields within these

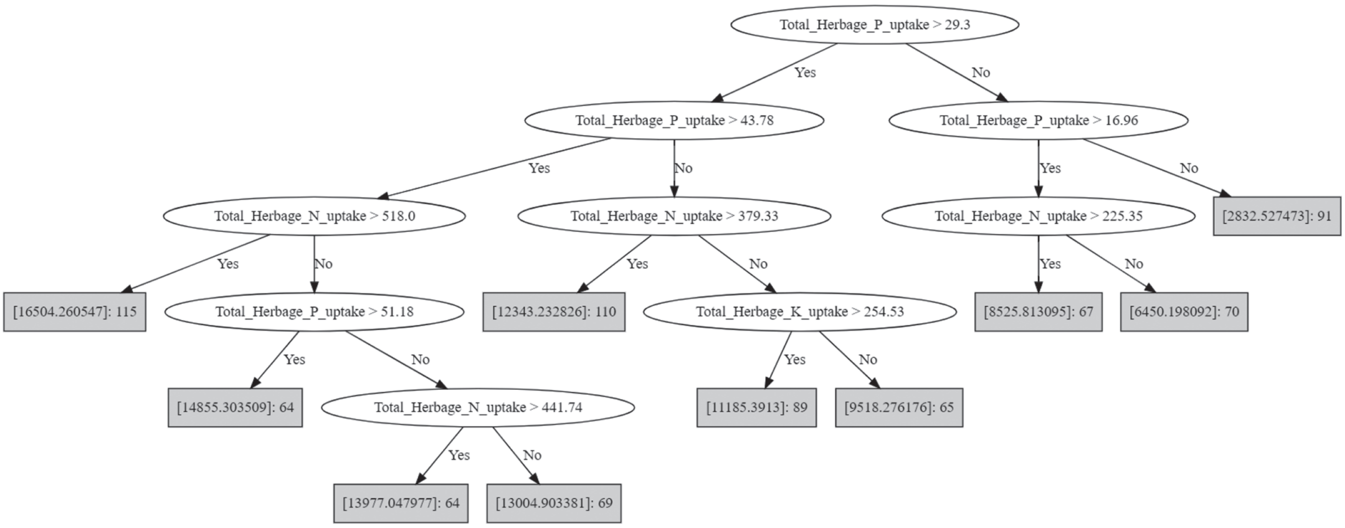

Total_Herbage_P_uptake > 29.3
Yes
No
Total_Herbage_P_uptake > 43.78
Total_Herbage_P_uptake > 16.96
Yes
No
Yes
No
Total_Herbage_N_uptake > 518.0
Total_Herbage_N_uptake > 379.33
Total_Herbage_N_uptake > 225.35
[2832.527473]: 91
Yes
No
Yes
No
Yes
No
[16504.260547]: 115
Total_Herbage_P_uptake > 51.18
[12343.232826]: 110
Total_Herbage_K_uptake > 254.53
[8525.813095]: 67
[6450.198092]: 70
Yes
No
Yes
No
[14855.303509]: 64
Total_Herbage_N_uptake > 441.74
[11185.3913]: 89
[9518.276176]: 65
Yes
No
[13977.047977]: 64
[13004.903381]: 69

Figure 2. Estimating herbage production by using N, P, and K uptake as descriptors [i.e., herbage production = f (N, P, K)].

groupings is evaluated, indications of nutrient limitations for achieving maximum herbage production arise. Figure 3B indicates that N may be somewhat limiting in field groups with N:P ratio <9 as the N concentrations dropped below 30 g/kg of DM, and similarly Figure 3C indicates that P may be somewhat limiting on field groups with a N:P ratio > 11, as the herbage P concentration drops to ~3 g/kg of DM. Figure 3D shows that other field groups had low K concentrations (<25 g/kg) even when the N and P appeared to be optimal. This multi-target modeling approach shows utility for assessing and identifying what nutrients may be limiting herbage production across a range of field sites with different S, E, and M conditions. It could also be used to assess the robustness of nutrient management programs where multiple nutrient input practices interact with varying soil and soil fertility levels both within and between farms.

How Do S, E, and M Variables Interact Within a Grazed Grassland Farming System to Affect Herbage Production and Herbage Nutrient Uptake?

The above discussion provides insight into the effects of nutrient uptake on herbage production. However, many factors affect both soil nutrient supply and plant uptake of nutrients beyond fertilization and general nutrient management practices. To further elucidate which factors may be most important in this respect, we introduce a third research question: How do S, E, and M variables interact within a grazed grassland farming system to affect herbage production and herbage nutrient uptake? To answer this question, we evaluated if our models could simultaneously predict N, P, and K uptake and herbage production using various S, E, and M attributes collected at each field site [i.e., (N, P, K, herbage production) = f (S, E, M)]. The tree shown in Figure 4 is pruned by using the *F*-test pruning procedure, which selected the optimal significance level of 0.005.

First, we examine the MTR model performance (descriptive and predictive), shown in Table 2. The descriptive (training) performance of the pruned tree, averaged across the 4 targets, is $r^2 = 0.733$ and RRMSE = 0.516, which does not differ significantly from the descriptive performance of the original unpruned tree. This fact confirms that the pruning method performed well. The predictive (testing) performance of the pruned tree is $r^2 = 0.684$ and RRMSE = 0.562, which is quite good for this specific domain, considering the problem complexity. Furthermore, if we compare the performances of the single-target tree that predicts herbage production potential and the multi-target tree that predicts herbage production potential plus N, P, and K uptake, simultaneously, using the same descriptive variables, we can see that the difference is only 1% (the average performance of single-target trees is 1% better than that of the MTR tree). This is an insignificant dif-

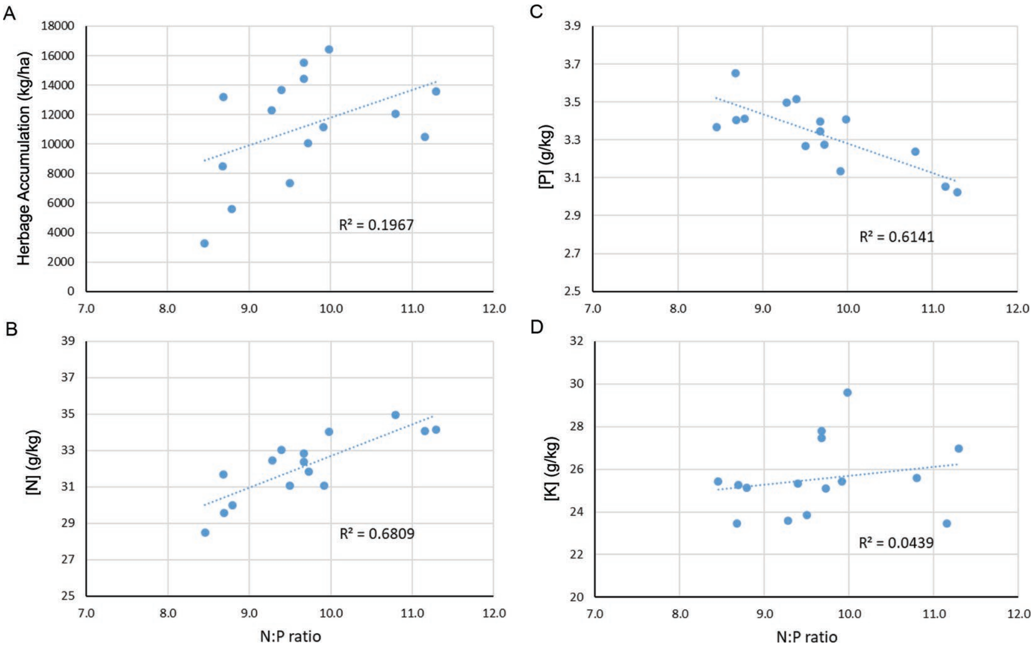


Figure 3. Relationships between (A) herbage accumulation, (B) nitrogen, (C) phosphorus, and (D) potassium concentrations and mass N:P ratio of herbage produced within the groups of different field sites identified by the model (Figure 2).

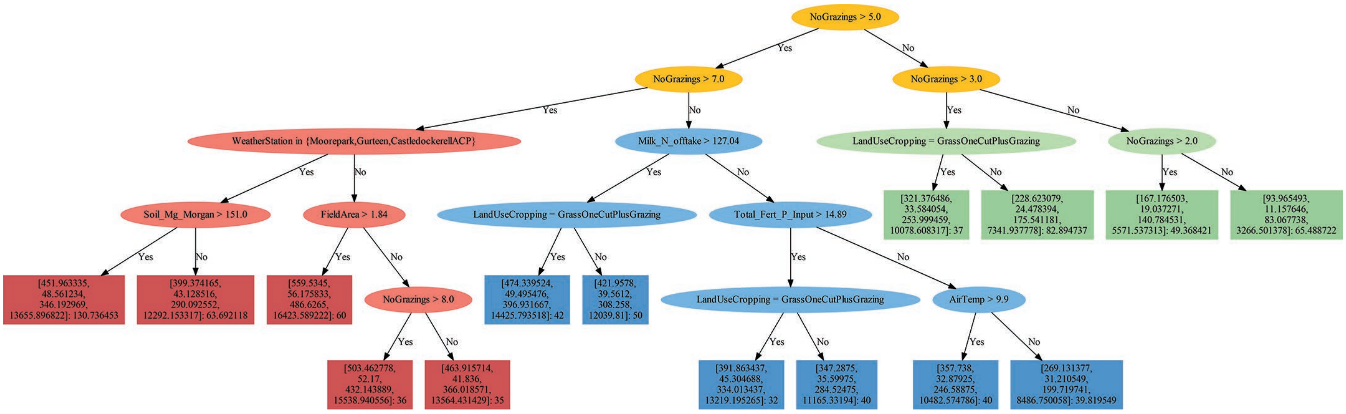


Figure 4. Multi-target regression tree learned on the complete data set. Colored nodes are related on 3 different categories: category 1 (red): high herbage production potential; category 2 (blue): medium herbage production potential; and category 3 (green): low herbage production potential.

ference in performance, but a strong advantage of using MTR modeling because instead of looking separately at 4 different single-target models for N, P, and K uptake and herbage production, we look at and use only one tree that predicts all 4 values at the same time. Hence, using MTR has practical advantages. As expected, the predictive performance is improved by approximately 10%, if we use an ensemble of PCT for MTR (in our case a random forest). We have $r^2 = 0.777$ and RRMSE = 0.498.

Next, we continue with an interpretation of the MTR tree model given in Figure 4. The top descriptive attribute was the number of grazings (NoGrazings), where sites with higher numbers of grazings had higher annual herbage accumulation as compared with those which were grazed less frequently over the year. Similar to the discussion of Figure 1 previously, the NoGrazings variable was closely linked with herbage production and utilization, with nutrient inputs and recycling and also with the drainage class of the soils within each grouping of fields (Table 1). Hence, the NoGrazings variable divides the sites into the same herbage production potential categories (high, medium, and low annual herbage production potential). Examining the differences between the trees produced by the single-target and MTR tree models (Figure 1 versus Figure 4), we can see that there are differences in the nutrients used to split the groups. For example, in Figure 1, the attribute used to split the fields is Soil_K_Morgan, but in Figure 4, the Soil_Mg_Morgan attribute appears. In many parts, the discussion of the attributes used for splitting the tree of Figure 1 are the same for Figure 4. In the following part, we only discuss the new tests that appear in this MTR tree, but not in the single-target tree in Figure 1. Specifically, we interpret the tests: WeatherStation = [Moorepark, Gurteen, Castledockerell], FieldArea >1.84 ha (in high herbage production potential fields) and AirTemp >9.9 (in medium herbage production potential fields).

The weather stations Moorepark, Gurteen, and Castledockerell have similar low rainfall and air temperatures in contrast to the other weather stations (Johnstown, Ballycanew, and Timoleague). In Ireland, the amount of rainfall is often a limiting factor on the times when grazing animals can enter a field for grazing events, because when the soils are too wet severe poaching of the soil can occur, which has negative consequences for subsequent herbage production and utilization. An area with low rainfall would suggest that there are more opportunities (days available) to graze a field, compared with an area with high rainfall (less days available). Note that rainfall will affect the trafficability of the soil, where poor trafficability due to high rainfall means animals will not be able to graze because the soil is too soft and the animals would only damage it. Good trafficability during dry spells means animals can graze without damaging the soil.

Next, we move to the test FieldArea >1.84 ha. This side of the branch was not distinguishable specifically by the field area, drainage class, or slope, but was related back to the weather station difference. This side of the branch represents the weather stations that are nearer the coast (Johnstown, Ballycanew, and Timoleague). Farms on the coast are generally slightly warmer and are not as severely affected by frost, which can affect herbage production levels. Based on the evidence, we found for the sites that belong to this branch of the tree, we see that these farms (and associated weather stations) have a lower number of degree days (days below 15°C), which gives them a longer growing/grazing season. It is likely that field area and number of grazings are linked. A high stocking rate and small field area could result in more grazings. A high stocking rate on a large field would take more days to graze out as compared with a small field. With a limited amount of time during the grazing season, a smaller field may be grazed out more times than a larger field. For this data set, the average field size was 1.37 ha, with field sizes ranging from 1.74 to 0.82 ha, excluding outliers.

The next new condition in the tree given in Figure 4 is AirTemp >9.9. Air temperature may be a proxy for weather differences per region. This proxy is also connected to rainfall, solar radiation, and degree days. Air temperature may indicate a longer growing/grazing season. A longer growing season allows more time to apply fertilizers and increase the pasture production (i.e., more grazings and higher herbage production). Moreover, air temperature is an indicator of growing season length and can be connected to the number of growing days (i.e., degree days below 15°C). A higher number of degree days with temperatures below 15°C (i.e., a colder growing season) versus a lower number of degree days, which means a lower number of days below 15°C (i.e., a warmer growing season). Air temperature and soil temperature (to 10 cm depth) are very closely linked. Nitrogen uptake begins around 5°C and grass growth around 6°C. An average air temperature above 9.9°C would indicate the soil conditions were suitable for pasture production for a longer time period. This enables fields or farms with air temperatures greater than 9.9°C to experience a longer growing season.

Additionally, we split the CD into 3 drainage classes: well-drained, somewhat poorly drained, and poorly drained sites. We perform the same analysis on those 3 different data sets to see if some new insights and knowledge can be extracted considering the tree models learned from each data set. Although the distinguishing based on drainage class is made by the number of grazings on the CD, we tried to find additional information in the models for each drainage class. For example, considering the low herbage production potential fields (green subtree) in the tree (Figure 4), we can see that for this subtree, there are no other splitting attributes but NoGrazing and LandUseCropping that we already discussed before. Following the fact that most of the examples belonging to this subtree are poorly drained, the idea of considering the tree obtained on the PD is justified.

If we look in the tree obtained from the PD (see Figure 5), we can see that there is additional important information that complements the green subtree given in Figure 4. Namely, if NoGrazings > 2 or ≤ 2, the next attribute used is N_recycled_per_field, which indicates the fact that the reason of higher or lower annual herbage accumulation is the amount of nutrients recycled (dung and urine) by the cows at grazing.

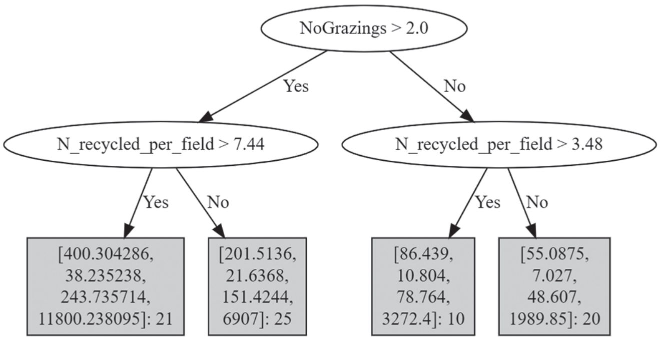


Figure 5. Multi-target regression tree learned on the poorly drained data.

The predictive performance of the tree built on the PD is the highest ($r^2 = 0.761$ and RRMSE $= 0.489$). The ensembles, again, as expected, improve the predictive performance ($r^2 = 0.830$ and RRMSE $= 0.438$).

The predictive and descriptive performances for all 3 scenarios (well-drained, somewhat poorly drained, and poorly drained) are given in Appendix Tables A2, A3, and A4, respectively.

Summary and Discussion

The technical results show that performance of a MTR tree is not significantly different from that of the 4 single-target models. The predictive performance of $r^2 = 0.684$ shows that we have a quite good predictive model, despite the complexity of the task at hand. Practically, the MTR tree is more efficient, since we have to interpret only one tree instead of looking at 4 different trees, one for each response variable. As we discussed, a MTR tree can be easily interpreted by a domain expert (e.g., agronomist). As expected, we get improved predictive performance if we use ensembles of PCT (random forest of PCT) in all scenarios in both single- and MTR tasks, but we lose interpretability.

Using this modeling approach, we found that the N, P, and K uptake is not always proportional relative to herbage accumulation levels. We could explore this variability by interpreting the learned models to better understand which nutrient may be limiting herbage production. Variability in herbage nutrient concentrations across dairy farms could lead to significant variability in nutrient use efficiency for a given level of production. Across these dairy grassland fields used for dairy production, we found the strength of limitation to herbage production based on nutrient uptake to follow this order P > N > K. This finding indicates that grassland swards are undersupplied with P from either soil reserves or fertilizer P input to maximize herbage production potential. Uptake of N and K is less limiting; however, given the P limitations, their uptake would be less efficiently used by the grassland.

In this study, we identified several important S, E, and M variables driving grass production in Ireland. Out of these factors, the number of grazing events was the most significant factor related to annual herbage accumulation. Across the large number of fields used in this study, the number of grazing events was also linked with soil drainage class, which indicates that soil type moderates/controls herbage production potential and also herbage utilization under grazing management. On high herbage production potential grassland, the regional weather has the biggest effect, whereas on medium and low herbage production potential grassland, the factors grassland land use (grazing vs. silage) and fertilizer P input have the largest effects.

Overall, the MTR tree provided the most useful information in terms of explaining herbage production potential and nutrient uptake across these grassland sites. Although this modeling approach could be used to identify herbage production potential with high accuracy, it can also inform the most influential dynamic factors that could be managed to increase herbage production in the future. These models could be also used as a basis for integrated soil fertility management, where other factors, such as soil type and environment factors, constrain optimum N, P, and K recommendations.

CONCLUSIONS

Our research study combined machine learning (i.e., data-driven modeling techniques) with practices and knowledge based on various soil, environmental, and management indicators, which describe interactions between nutrient uptake and herbage production. This approach has several technical advantages and implication for future nutrient management and advice for farmers to increase herbage production on dairy farms. The implications of this work for Irish grass-based dairy farms are as follows: (1) the models we have learned from data can be used to identify fields with poorer herbage production performance and to direct on-site investigation to ascertain the problem or the constraints. (2) This data-driven modeling approach suggests that (1) guiding a more balanced approach to fertilizer inputs, including P and also K, is required, in addition to high quantities of *N* fertilizer input; (2) to improve environmental sustainability, explicit geo and climatic recommendations for fertilization are required; and (3) to monitor and assess grassland productivity, only a few variables are required, including soil drainage class (grazing events), grassland management, soil nutrient status, production intensity, as well as region and local weather. Further work could be conducted to evaluate other farm production and environmental sustainability targets, as well as trade-offs and synergies between the underlying factors by using different modeling approaches for solving the MTR task.

ACKNOWLEDGMENTS

We thank the TEAGASC, Environment Soils and Land-Use Department, for providing the data. Furthermore, we acknowledge the support of the European Commission through the project LANDMARK - Land Management, Assessment, Research, Knowledge Base (H2020 grant number 635201). As their fellow, SN thanks the Teagasc Walsh Fellowship Programme (Johnstown Castle, Co. Wexford, Ireland).

REFERENCES

Blockeel, H., L. Raedt, and J. Ramon. 1998. Top-down induction of clustering trees. Pages 55–63 in Proceedings of the 15th International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA.

Blockeel, H., and J. Struyf. 2002. Efficient algorithms for decision tree cross-validation. J. Mach. Learn. Res. 3:621–650.

Breiman, L. 2001. Random forests. Mach. Learn. 45:5–32.

Buckley, C., D. P. Wall, B. Moran, S. O'Neill, and P. N. C. Murphy. 2016. Phosphorus management on Irish dairy farms post controls introduced under the EU Nitrates Directive. Agric. Syst. 142:1–8.

Byrne, E., 1979. Chemical Analysis of Agricultural Materials. An Foras Taluntais, Dublin, Ireland.

Connolly, L., A. Kinsella, G. Quinlan, and B. Moran. 2010. National Farm Survey 2008. Teagasc, Athenry, Ireland.

Creighton, P., E. Kennedy, L. Shalloo, T. M. Boland, and M. O'Donnovan. 2011. A survey analysis of grassland dairy farming in Ireland, investigating grassland management, technology adoption and sward renewal. Grass Forage Sci. 66:251–264.

Demšar, D., S. Džeroski, T. Larsen, J. Struyf, J. Axelsen, M. B. Pedersen, and P. H. Krogh. 2006. Using multi-objective classification to model communities of soil. Ecol. Modell. 191:131–143.

Dillon, P., and L. Delaby. 2009. Challenges from EU and International Environmental policy and legislation to animal production from temperate grassland. Tearmann (Dublin) 7:51–68.

Ergon, Å., L. Kirwan, M. A. Bleken, A. O. Skjelvåg, R. P. Collins, and O. A. Rognli. 2016. Species interactions in a grassland mixture under low nitrogen fertilization and two cutting frequencies: 1. Dry-matter yield and dynamics of species composition. Grass Forage Sci. 71:667–682.

Finneran, E., P. Crosson, P. O'Kiely, L. Shalloo, D. Forristal, and M. Wallace. 2010. Simulation modelling of the cost of producing and utilising feeds for ruminants on Irish farms. J. Farm Manag. 14:95–116.

Hanrahan, L., A. Geoghegan, M. O'Donnovan, V. Griffith, E. Ruelle, M. Wallace, and L. Shalloo. 2017. PastureBase Ireland: A grassland decision support system and national database. Comput. Electron. Agric. 193:193–201.

Hennessy, T., and B. Moran. 2015. National Farm Survey 2015, Athenry, Co. Galway, Ireland. Agricultural Economics and Farm Surveys Department, Rural Economy Development Programme.

Herlihy, M. 1979. Nitrogen mineralisation in soils of varying texture, moisture and organic matter, I. Potential and experimental values in fallow soils. Plant Soil 53:269–275.

Humphreys, J., K. O'Connell, and I. A. Casey. 2008. Nitrogen flows and balances in four grassland based systems of dairy production on a clay-loam soil in a moist temperate climate. Grass Forage Sci. 63:467–480.

ICBF. 2018. Irish Cattle Breeding Federation website. Accessed Feb. 14, 2019. https://www.icbf.com/wp/.

Kocev, D., S. Džeroski, M. D. White, G. R. Newell, and P. Griffioen. 2009. Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition. Ecol. Modell. 220:1159–1168.

Kocev, D., C. Vens, J. Struyf, and S. Džeroski. 2013. Tree ensembles for predicting structured outputs. Pattern Recognit. 46:817–833.

Levatić, J., D. Kocev, M. Debeljak, and S. Džeroski. 2015a. Community structure models are improved by exploiting taxonomic rank with predictive clustering trees. Ecol. Modell. 306:294–304.

Li, G. D., K. R. Heylar, L. J. Castleman, G. Norton, and R. P. Fisher. 1998. The implementation and limitations of using a falling plate meter to estimate pasture yield. In Agronomy—Growing a Greener Future, Proceedings of the 9th Australian Agronomy Conference, July 1998, 20–23, Sydney, (Australia).

Murphy, P. M., P. N. C. Murphy, and D. P. Wall. 2018. An evaluation of nutrient balances at the whole-farm and field scale on 21 Irish dairy farms. Proceedings of the 27th European Grassland Federation General Meeting, Cork 17th-21st June 2018, 23 Grassland Science in Europe.

Nitrate Directive. 1991. Council Directive of December 1991 concerning the protection of waters against pollution caused by nitrates from agricultural sources. Urban Waste Water Treatment Directive. Council Directive 91/271/EEC of May 1991 concerning urban waste-water treatment, European Commission, Brussels, Belgium.

Nunan, N., M. A. Morgan, J. Scott, and M. Herlihy. 2000. Temporal changes in nitrogen mineralisation, microbial biomass, respiration and protease activity in a clay loam soil under ambient temperature. Biol. Environ. 100B:107–114.

O'Donovan, M., E. Lewis, and P. O'Keily. 2011. Requirements of future grass-based ruminant production systems in Ireland. Ir. J. Agric. Food Res. 50:1–21.

O'Leary, M., A. Geoghegan, and M. O'Donovan. 2016. PastureBase Ireland – Capturing grassland data on commercial Irish farms. Accessed Jun. 20, 2019. https://www.teagasc.ie/media/website/crops/grassland/IGA-Student-Conference.pdf.

O'Mara, F. 2008. Country Pasture/Forage Resource Profile/Ireland. Accessed Mar. 11, 2019. http://www.fao.org/ag/AGP/AGPC/doc/pasture/forage.htm.

Schils, R. L. M., M. H. de Haan, J. G. Hemmer, A. van den Pol-van Dasselaar, J. A. de Boer, A. G. Evers, G. Holshof, J. C. van Middelkoop, and R. L. Zom. 2007. DairyWise, A whole-farm dairy model. J. Dairy Sci. 90:5334–5346.

Shalloo, L., P. Dillon, M. Rath, and M. Wallace. 2004. Description and validation of the Moorepark Dairy System Model. J. Dairy Sci. 87:1945–1959.

Sheridan, H., J. Finn, N. Culleton, and G. O'Donovan. 2008. Plant and invertebrate diversity in grassland field margins. Agric. Ecosyst. Environ. 123:225–232.

Simo, I., R. Creamer, B. Reidy, G. Jahns, P. Massey, B. Hamilton, J. Hannam, E. McDonald, and P. Sills. 2008. Irish Soil Information System. Accessed Feb. 13, 2019. http://gis.teagasc.ie/isis/.

SIS. 2017. Soil Index System. Accessed Jan. 27, 2019. https://www.teagasc.ie/crops/soil–soil-fertility/soil-analysis/soil-index-system/.

Smit, H. J., H. Z. Taweel, B. M. Tas, S. Tamminga, and A. Elgersma. 2005. Comparison of techniques for estimating herbage intake of grazing dairy cows. J. Dairy Sci. 88:1827–1836.

Smith, A., and P. J. Allcock. 1985. The influence of species diversity on sward yield and quality. J. Appl. Ecol. 22:185–198.

Teagasc. 2016. Sectoral road map: Dairying. Accessed Nov. 14, 2018. https://www.teagasc.ie/publications/2016/road-map-2025-dairy.php.

Teagasc. 2019, PastureBase Ireland. Accessed May 2019. https://www.teagasc.ie/crops/grassland/pasturebase-ireland/.

Vellinga, T. V., G. Andre, R. L. M. Schils, T. Kraak, and O. Oenema. 2010. Accounting for residual effects of previously applied nitrogen fertilizer on intensively managed grasslands. Grass Forage Sci. 65:58–75.

Wall, D. P., and M. Plunkett. 2016. Major and micro nutrient advice for productive agricultural crops. Teagasc Johnstown Castle, October 2016. https://www.teagasc.ie/media/website/publications/2016/soil-fertility-green.pdf.

APPENDIX

Table A1. Descriptions of the soil (S), environment/weather (E), and management (M) factors (variables) and response variables (annual herbage accumulation and nutrient herbage uptake – N, P, and K uptake)

Model category	Data type/category	Data heading	Data description
Metadata	Identification	Field code	Individual fields (paddocks), experimental units
M	Field details	Land use-cropping	2 types (grazing only and grazing + 1 cut of silage)
		Field area (ha)	Area of the experimental unit
M	Production level	Total no. of cows	Total number within the herd on each farm
		Average stocking rate (LU/ha)	Average live units per milking platform (1 mature cow/ha = 1 LU/ha)
M	Milk offtakes	Milk offtake (L/ha)	Milk produced/ha for each field
		Milk N offtake (kg/ha)	N in milk removed/ha for each field
		Milk P offtake (kg/ha)	P in milk removed/ha for each field
		Milk K offtake (kg/ha)	K in milk removed/ha for each field
M	Meat offtakes	Meat offtake (kg/ha)	Meat produced/ha for each field
		Meat N offtake (kg/ha)	N in meat removed/ha for each field
		Meat P offtake (kg/ha)	P in meat removed/ha for each field
		Meat K offtake (kg/ha)	K in meat removed/ha for each field
E	Weather (annual)	Weather station	Name of the weather station
		Rainfall (mm)	
		Radiation (J/cm^2)	
		Air temperature (°C)	
		Degree days below 15°C	Cumulative degrees below a base temperature of 15°C (linked with reduced grass growth)
S	Soil characteristics	SIS[1] class	Soil classification
		Drainage class	4 classes
		Slope Y/N	Yes or no
		Slope class	3 classes
		Soil pH	Measured soil pH (acidity) in top 10 cm of soil
		Soil LR (t/ha)	Calculated lime required to neutralize soil acidity and correct pH to target of 6.3
		Soil P_Morgan (mg/L)	Measured soil P concentration in top 10 cm of soil
		Soil K_Morgan (mg/L)	Measured soil K concentration in top 10 cm of soil
		Soil Mg_Morgan (mg/L)	Measured soil Mg concentration in top 10 cm of soil
M	Pasture management	PRGrass class	Perennial ryegrass (*Lolium perenne*) class
		W. clover class	White clover class
		No. of harvests	No. of silage harvesting events per year
		No. of grazings	No. of grazing events by cows per year
M	Lime management	Lime type	Type of lime (2 types: granulated and ground limestone)
		Lime (kg/ha)	Lime input level
M	Nutrient management	Org N input (kg/ha)	N input level in the form of organic manure
		Org P input (kg/ha)	P input level in the form of organic manure
		Org K input (kg/ha)	K input level in the form of organic manure
		Chem. N input (kg/ha)	N input level in the form of chemical fertilizer
		Chem. P input (kg/ha)	P input level in the form of chemical fertilizer
		Chem. K input (kg/ha)	K input level in the form of chemical fertilizer
		Conc. N input (kg/ha)	N input level in the form of concentrated feed
		Conc. P input (kg/ha)	P input level in the form of concentrated feed
		Total Fert N input (kg/ha)	Total N input level in the form of fertilizer (organic + chemical)
		Total Fert P input (kg/ha)	Total P input level in the form of fertilizer (organic + chemical)
		Total Fert K input (kg/ha)	Total N input level in the form of fertilizer (organic + chemical)
M	Nutrients recycled	Average N recycled (kg of Org N/ha)	Average N recycled by grazing animals (N excretion rate per cows) over the milking platform
		Average P recycled (kg of Org N/ha)	Average P recycled by grazing animals (P excretion rate per cows) over the milking platform
		N recycled/field (kg of Org N/ha)	N recycled by grazing animals (N excretion rate per cows) per field
		P recycled/field (kg of Org N/ha)	P recycled by grazing animals (N excretion rate per cows) per field
Response	Herbage uptake	Total herbage N uptake (kg/ha)	Total N uptake by herbage (grazed grass and silage)
		Total herbage P uptake (kg/ha)	Total P uptake by herbage (grazed grass and silage)
		Total herbage K uptake (kg/ha)	Total K uptake by herbage (grazed grass and silage)
Response	Herbage production	Total herbage accumulation (kg/ha)	Total annual herbage biomass production

[1] SIS (2017).

[2] Org = organic.

Table A2. Performance results for the well-drained data set achieved by predictive clustering tree (PCT) for single- and multi-target regression[1]

		Regression trees (PCT) built with *F*-test pruning								Random forest of PCT	
		Training performance				Testing performance				Testing performance	
		Unpruned tree		Pruned tree		Unpruned tree		Pruned tree		Forest with 100 trees	
Response variable	Optimal *F*-test significance level	r^2	RRMSE	r^2	RRMSE	r^2	RRMSE	r^2	RRMSE	r^2	RRMSE
Single-target regression											
Total_pasture_yield (herbage production)	0.005	0.680	0.566	0.674	0.571	0.601	0.634	0.600	0.634	0.763	0.528
Total_herbage_N_uptake (N)	0.05	0.664	0.579	0.657	0.586	0.553	0.671	0.546	0.677	0.734	0.552
Total_herbage_P_uptake (P)	0.05	0.667	0.577	0.667	0.577	0.565	0.662	0.550	0.673	0.729	0.562
Total_herbage_K_uptake (K)	0.05	0.673	0.572	0.666	0.578	0.575	0.654	0.571	0.657	0.715	0.569
Average (herbage production, N, P, K)		0.671	0.574	0.666	0.578	0.574	0.655	0.567	0.661	0.735	0.552
Multi-target regression											
Herbage production + N + P + K	0.05										
Herbage production		0.673	0.572	0.659	0.584	0.589	0.642	0.579	0.650	0.764	0.555
N		0.630	0.608	0.623	0.614	0.550	0.673	0.540	0.680	0.730	0.566
P		0.628	0.610	0.624	0.613	0.564	0.662	0.559	0.665	0.725	0.570
K		0.620	0.616	0.604	0.630	0.514	0.700	0.505	0.707	0.714	0.531
Average (herbage production, N, P, K)		0.638	0.602	0.627	0.610	0.554	0.669	0.546	0.676	0.733	0.556

[1] RRMSE = relative root mean square error; r^2 = Pearson correlation coefficient.

Table A3. Performance results for the somewhat poorly drained data set achieved by predictive clustering tree (PCT) for single- and multi-target regression[1]

		Regression trees (PCT) built with *F*-test pruning								Random forest of PCT	
		Training performance				Testing performance				Testing performance	
		Unpruned tree		Pruned tree		Unpruned tree		Pruned tree		Forest with 100 trees	
Response variable	Optimal *F*-test significance level	r^2	RRMSE	r^2	RRMSE	r^2	RRMSE	r^2	RRMSE	r^2	RRMSE
Single-target regression											
Total_pasture_yield (herbage production)	0.005	0.838	0.403	0.775	0.474	0.613	0.628	0.554	0.674	0.720	0.557
Total_herbage_N_uptake (N)	0.1	0.838	0.402	0.791	0.457	0.679	0.569	0.666	0.579	0.722	0.561
Total_herbage_P_uptake (P)	1	0.825	0.419	0.752	0.498	0.588	0.647	0.565	0.666	0.708	0.567
Total_herbage_K_uptake (K)	0.01	0.811	0.435	0.737	0.513	0.592	0.644	0.601	0.635	0.724	0.557
Average (herbage production, N, P, K)		0.828	0.415	0.764	0.486	0.618	0.622	0.597	0.639	0.718	0.561
Multi-target regression											
Herbage production + N + P + K	0.005										
Herbage production		0.801	0.446	0.769	0.481	0.638	0.606	0.639	0.605	0.730	0.552
N		0.780	0.469	0.760	0.490	0.645	0.599	0.650	0.594	0.703	0.575
P		0.776	0.474	0.752	0.498	0.632	0.611	0.631	0.611	0.716	0.559
K		0.755	0.495	0.737	0.513	0.611	0.628	0.618	0.622	0.713	0.567
Average (herbage production, N, P, K)		0.778	0.471	0.754	0.495	0.631	0.611	0.634	0.608	0.715	0.563

[1] RRMSE = relative root mean square error; r^2 = Pearson correlation coefficient.

Table A4. Performance results for the poorly drained data set achieved by predictive clustering trees (PCT) for single- and multi-target regression[1]

		Regression trees (PCT) built with *F*-test pruning								Random forest of PCT	
		Training performance				Testing performance				Testing performance	
		Unpruned tree		Pruned tree		Unpruned tree		Pruned tree		Forest with 100 trees	
Response variable	Optimal *F*-test significance level	r^2	RRMSE	r^2	RRMSE	r^2	RRMSE	r^2	RRMSE	r^2	RRMSE
Single-target regression											
Total_pasture_yield (herbage production)	0.125	0.930	0.265	0.908	0.304	0.804	0.450	0.805	0.449	0.859	0.405
Total_herbage_N_uptake (N)	0.001	0.902	0.313	0.902	0.313	0.899	0.319	0.892	0.329	0.884	0.366
Total_herbage_P_uptake (P)	0.05	0.893	0.327	0.862	0.372	0.733	0.523	0.754	0.498	0.811	0.464
Total_herbage_K_uptake (K)	0.1	0.874	0.355	0.839	0.401	0.751	0.502	0.752	0.499	0.746	0.531
Average (herbage production, N, P, K)		0.900	0.315	0.878	0.348	0.797	0.448	0.801	0.444	0.825	0.441
Multi-target regression											
Herbage production + N + P + K	0.001										
Herbage production		0.899	0.319	0.899	0.319	0.803	0.445	0.806	0.441	0.862	0.403
N		0.875	0.354	0.875	0.354	0.788	0.462	0.787	0.463	0.882	0.378
P		0.862	0.372	0.862	0.372	0.745	0.507	0.745	0.507	0.818	0.454
K		0.812	0.433	0.812	0.433	0.716	0.535	0.705	0.545	0.756	0.515
Average (herbage production, N, P, K)		0.862	0.369	0.862	0.369	0.763	0.487	0.761	0.489	0.830	0.438

[1] RRMSE = relative root mean square error; r^2 = Pearson correlation coefficient.

6.2 Exploiting Partially-Labeled Data for Learning Water Quality Models in Irish Agricultural Catchments

Environmental scientists, during monitoring programs, generate vast amounts of data by measuring the main environmental indicators related to the monitoring purpose. In this vein, within the national water quality monitoring program in the Republic of Ireland, thousands of water and soil samples are collected all across the country, recording the source- and transport-related indicators, as well as water quality indicators, such as biological water quality (Q-value), nitrogen (N) and phosphorus (P) concentration, the main factors driving the process of eutrophication. Despite the fact that thousands of samples are collected, there is a large number of samples where the measurements are not complete (Schulte et al., 2006).

The data from national monitoring programs typically consists of database tables for each water quality indicator separately. In order to construct representative and structured learning data, data analysts pre-process these thousands of examples and join databases to take only those which are complete in the descriptive (i.e., input) space. In particular, there are different descriptive (i.e., source- and transport-related) indicators for each water quality indicator. However, the final harmonized data consist of only those examples with known values for the descriptive indicators. The 'data incompleteness' we consider in this chapter is related to the target space i.e., the water quality indicators (Q-value, N and P concentrations).

In classical, i.e., supervised machine learning algorithms, these 'incomplete' samples are excluded from the learning process, which is performed only on complete, i.e., fully-labeled examples. Semi-supervised algorithms for SOP can handle (fully-) unlabeled in addition to (fully-) labeled examples. Moreover, in the case of SOP, if an example is partially-labeled, it is considered as unlabeled, i.e., the known values for some of the targets are entirely removed and the experiments, as usual, are performed on labeled and unlabeled examples (K. Chen et al., 2020; Chou et al., 2018; Giri et al., 2019). However, with purposely removing the known labels, i.e., delabeling incompletely labeled examples, potential information that can affect the quality of the obtained models can be overlooked. Semi-supervised predictive clustering trees (PCTs) have been shown to yield both more accurate and more succinct (and hence interpretable) models as compared to their fully supervised counterparts that use only the labeled part of the same SSL datasets. Therefore, they are an ideal choice as a methodology that can be improved in order to handle partially-labeled examples rather than to discard or delabel them.

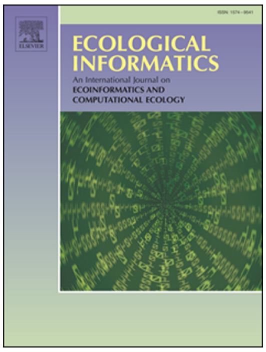
In this chapter, we propose such adapted semi-supervised PCTs for multi-target regression, which can handle partially-labeled examples. To the best of our knowledge, this semi-supervised modeling approach is a novel approach in the domain of environmental (soil) sciences. There are only several studies applying semi-supervised learning in environmental sciences, but they focus on single-target predictive modeling, i.e., SSL for classification and regression (cf. (Abraham & Tan, 2009; Herrera et al., 2010)). The adapted semi-supervised PCTs for MTR learning from partially-labeled examples are deliberately chosen as the most proper learning method, since our dataset consisting of 708 pre-processed data samples, collected within the national water quality monitoring program (2001-2003), has almost 50% of 'incomplete' (i.e., partially-labeled) examples. Namely, we use predictive clustering trees (PCTs) and ensembles (random forests) of PCTs for multi-target regression and we build single-target (i.e., local) PCT models for each target separately and a multi-target (i.e., global) PCT model which predicts all targets simultaneously. Our results have shown that the best models in terms of predictive performance (i.e., RRMSE) are semi-supervised models which can handle incompletely- (i.e., partially-) labeled data and the best models in terms of complexity (i.e., model size) are multi-target regression (i.e., global) models. Moreover, the predictive clustering paradigm in itself unifies the approaches of predictive modeling and clustering and with that ideally shows the interactions between input variables and targets which are consistent with existing findings.

The paper included in this section is:

- NIKOLOSKI, Stevanche, KOCEV, Dragi, LEVATIĆ, Jurica, WALL, David, P. and DŽEROSKI, Sašo. (2020), Exploiting partially-labeled data in learning predictive clustering trees for multi-target regression: A case study of water quality assessment in Ireland, *Ecological Informatics*, 2020, doi:10.1016/j.ecoinf.2020.101161.

The contributions of Stevanche Nikoloski to this paper are as follows. SN adapted the existing computer code in order to handle partially-labeled examples. He contributed to the design and execution of the machine learning experiments. He evaluated their results and presented the model predictions in the form of maps. He drafted the paper and revised it according to the co-author's and reviewer's feedback.

Journal Pre-proof



Exploiting partially-labeled data in learning predictive clustering trees for multi-target regression: A case study of water quality assessment in Ireland

Stevanche Nikoloski, Dragi Kocev, Jurica Levatić, David P. Wall, Sašo Džeroski

PII: S1574-9541(20)30111-4

DOI: https://doi.org/10.1016/j.ecoinf.2020.101161

Reference: ECOINF 101161

To appear in: *Ecological Informatics*

Received date: 21 April 2020

Revised date: 3 August 2020

Accepted date: 30 August 2020

Please cite this article as: S. Nikoloski, D. Kocev, J. Levatić, et al., Exploiting partially-labeled data in learning predictive clustering trees for multi-target regression: A case study of water quality assessment in Ireland, *Ecological Informatics* (2020), https://doi.org/10.1016/j.ecoinf.2020.101161

Exploiting partially-labeled data in learning predictive clustering trees for multi-target regression: A case study of water quality assessment in Ireland

Stevanche Nikoloski[a,c,1], Dragi Kocev[b,a,2], Jurica Levatić[b,3], David P. Wall[c,4], Sašo Džeroski[b,a,5]

[a] *Jožef Stefan International Postgraduate School, Jamova c. 39, 1000, Ljubljana, Slovenia*
[b] *Department of Knowledge Technologies, Jožef Stefan Institute, Jamova c. 39, 1000, Ljubljana, Slovenia*
[c] *Teagasc, Crops, Environment & Land-use Programme, Johnstown Castle, Co. Wexford, Y35, Ireland*

Abstract

Many environmental problems give rise to predictive modeling tasks where several dependent variables need to be predicted simultaneousy from a given set of independent variables. When the target variables are numeric, the task at hand is called multi-target regression (MTR). An example task of this type is the assessment of quality of agricultural waters in Ireland according to three indicators: biological water quality, nitrogen concentration and phosphorus concentration.

Multi-target regression models are typically learnt from labeled training examples, where the values of both the dependent variables (labels) and the independent variables are provided, in a setting known as supervised learning. Many different approaches to supervised multi-target regression have been developed, among which predictive clustering trees and ensembles thereof stand out due to their effectiveness and efficiency. Recently, these approaches have been extended to exploit not only labeled examples, but also unlabeled examples, where only the values of the independent variables are provided, a setting known as semi-supervised learning.

In practice, training data can also contain partially labeled examples, where the values of some of the dependent variables are provided and others are missing (in addition

[1] *stevanche.nikoloski@ijs.si*
[2] *dragi.kocev@ijs.si*
[3] *jurica.levatic@ijs.si*
[4] *David.Wall@teagasc.ie*
[5] *saso.dzeroski@ijs.si*

Accepted for publication in Ecological Informatics *30 August 2020*

to fully labeled examples where all target values are provided and completely unlabeled examples where no target values are provided). For the task of water quality assessment in Ireland, we encounter this kind of partially labeled data. Existing supervised and semi-supervised MTR approaches typically ignore partially labelled data.

In this paper, we propose the use of semi-supervised predictive clustering trees for MTR that can handle partially labeled examples. We apply these to the task of assessment of water quality in Ireland, showing that better performance can be achieved if partially labeled examples are exploited, rather than discarded. We build both local models (collections of single-target models predicting each target separately) and global models (multi-target models simultaneously predicting all targets), showing that global models are both smaller and easier to interpret, and also overfit less (and have better performance) as compared to local models.

Keywords: Multi-target regression, random forests, predictive clustering trees, partially-labeled data, semi-supervised learning, water quality

1. Introduction

This paper is situated at the intersection between machine learning, on one hand, and environmental modeling, on the other hand. On the machine learning side, it deals with the task of multi-target regression (MTR), where models need to be learned to predict several dependent variables simultaneously. These are learned from training data that consists of pairs of input/ output vectors, in the setting of semi-supervised learning, where some of the values of the output/ target variables may be missing in some of the training data. Here, it considers cutting edge methods for MTR, which can deal with data points that are fully labeled, unlabeled, or partially labeled with values of the target variables. On the environmental modeling side, the paper considers the task of relating pressure-pathway data to three different indicators of water quality in agricultural fields in Ireland. This case study nicely fits the machine learning methods considered, as multiple (three) targets need to be predicted and MTR models need to be learned. In addition, the data from which MTR models need to be learned has many missing values for the target variables.

1.1. Multi-target regression

In the last decades, machine learning [1] has gained significant prominence and is now used in many domains, including the domains of agriculture, ecology and soil science. Data mining [2] typically uses tools from machine learning or statistics to find patterns in and extract knowledge from the data. The predominant paradigm in machine learning, called supervised learning, is concerned with learning predictive models learning from data. In this paradigm, a dataset consists of pairs of values of input/ descriptive/ independent variables and an output/ target/ dependent variable. In the classical machine learning tasks of classification and regression, the output variables is discrete, resp. continuous.

Recent research has studied predictive modeling tasks where the output can be a complex data structure, such as a tuple/ vector of values, sequence (such as a time series) and a hierarchy [3, 4, 5]. The task of learning models that predict structured outputs is called structured output prediction.

In this work, we are concerned with the task of multi-target regression (MTR)—a type of structured output prediction task where the goal is to simultaneously predict multiple continuous target variables. Structured outputs are encountered in many real world problems. Prominent examples can be found in ecology, for example, predicting the abundance of different species occupying the same habitat [6] or estimating different vegetation quality indices for the same site [7], predicting the composition of a community of organisms [8] or predicting the nutrient uptake and herbage production in grassland soils [9].

Table 1: *A multi-target regression (MTR) dataset.*

ID	**Description Space**				**Target Space**			
#	**Hi drain q1**	**drainage factor**	**Total N input**	**...**	**Q**	**P**	**N**	**Example type**
#1	3	0.35	121.26	...	**3.86**	**0.04**	**0.47**	Labeled examples
#2	2	0.25	119.86	...	**4.28**	**0.34**	**20.71**	
#3	2	0	120.4749	...	**4.35**	**0.024**	**0.43**	
#4	2	0.24	120.95	...	**3.47**	**0.155**	**0.15**	

In our case study in this paper, we address the MTR task of three different indicators of water quality in agricultural fields in Ireland from pressure-pathway data. The descriptive variables include, for example, a drainage factor and total nitrogen input. The three target variables are biological water quality, nitrogen concentration and phosphorus concentration. The dataset we use comes from a study by Schulte et al. [10]. An excerpt from this MTR dataset is given in Table 1.

Models for MTR can be learned locally, i.e., as collections of single-target models, learning one model for each target variable at a time. MTR models can also be learned globally to simultaneously predict all targets. An example of a global model is a decision tree for multi-target regression, such as a predictive clustering tree (PCT) for MTR [11]: Such a tree has the same structure as a regression tree, but predicts vectors of real values (one for each target) in the leaves, rather than a single real value (for a single target). An example PCT for MTR can be seen in Figure 10. Unlike local models, global models (such as PCTs for MTR) exploit the potential dependencies that might exist among the target variables to learn models with better predictive performance (as compared to local models) and provide a global and comprehensive overview on the modelled system.

Among the best approaches for MTR are predictive clustering trees (PCTs) for MTR and ensembles thereof [11, 12, 13]. PCTs are readily interpretable, while ensembles of PCTs provide excellent predictive performance. In addition, PCTs inherit several desirable properties of regular decision trees: (1) they do not make assumptions on the probability distributions of the descriptive and target variables, (2) they can handle both discrete/nominal and real/numeric descriptive variables and missing values, (3) they have low computational cost for learning. Finally, PCTs are very general and can handle a number of structured output prediction tasks: So far, PCTs have been employed for the tasks of multi-target prediction (which includes multi-label classification and MTR) [11], hierarchical multi-target prediction [14, 15, 16], and prediction of time-series [17, 18].

1.2. Semi-supervised multi-target regression

In the classical supervised machine learning tasks with structured outputs, the examples (i.e., instances) are labeled, i.e., the values for the target/class attributes are known. However, the labels, i.e., values of the target variable(s) are typically not easy to come by, as significant time and effort are needed to acquire their values. This is the case even for a single target variable, and even more so when several target variables are considered. On the other hand, unlabeled examples, without the value(s) of the target variable(s) are much easier to come by. This has been a major motivation for the development of the paradigm of semi-supervised learning (SSL)[19].

Table 2: *A multi-target regression (MTR) dataset in the semi-supervised learning (SSL) setting. '?' denotes a missing value.*

ID	**Description Space**				**Target Space**			
#	**Hi drain q1**	**drainage factor**	**Total N input**	**...**	**Q**	**P**	**N**	**Example type**
#1	3	0.35	121.26	...	**3.86**	**0.04**	**0.47**	Labeled examples
#2	2	0.25	119.86	...	**4.28**	**0.34**	**20.71**	
#3	2	0	120.4749	...	**4.35**	**0.024**	**0.43**	
#4	2	0.24	120.95	...	**3.47**	**0.155**	**0.15**	
#5	2	0.15	119.2568	...	**?**	**?**	**?**	Unlabeled examples
#6	3	0	121.0236	...	**?**	**?**	**?**	
#7	2	0.86	119.3698	...	**?**	**?**	**?**	
#8	2	0	115.3987	...	**?**	**?**	?	

Semi-supervised learning can exploit both labeled examples, for which the value(s) of the target variable(s) are known, and unlabeled examples, where the target values are missing. As such, it is situated on the spectrum between fully supervised learning (where all the target values are present and the task is to learn a predictive model) and fully unsupervised learning (where no target values are present and the task is to find a clustering). An example excerpt from a dataset for SSL in the context of multi-target regression is given in Table 2, where the two types of examples are clearly indicated.

Just as the majority of methods for fully supervised learning deal with the tasks of classification and regression, the majority of methods for SSL deal with semi-supervised classification and semi-supervised regression. Very few methods exist for SSL in the context of structured-output prediction (SOP), especially for multi-target regression. Example methods for semi-supervised learning in SOP include Gaussian process models for MTR [20] and multi-task classification by adapting kernel parameters in Gaussian processes for regression to handle unlabeled data [21, 22]. They also include SVMs using the co-training paradigm accompanied by using the joint input-output spaces and an arbitrary loss [23]. These approaches do not produce interpretable models.

Recently, the approach of learning predictive clustering trees for MTR has been extended to the semi-supervised learning setting [13]. The predictive clustering paradigm in itself unifies the approaches of predictive modeling and clustering and is an ideal match for the task of SSL. PCTs for SSL have been shown to yield both more accurate and more succint (and hence interpretable) models as compared to their fully supervised counterparts using only the labeled part of the same SSL datasets. Ensembles of PCTs have also been adapted to the SSL paradigm. PCTs and PCT ensembles for SSL are described in detail in Section 2.2.1.

1.3. Learning MTR models from partially labeled data

In SSL for (single-target) classification and regression, label values can either be present or absent, giving rise to labeled and unlabeled examples. In multi-target regression (and SOP in general), we can also have incomplete labels. For example, if we have three target variables, the values for one or two of them can be present/ missing, in contrast to the case where all three are present/ absent.

An example excerpt from a SSL dataset with partially labelled examples is given in Table 3. It contains all three possible kinds of examples: unlabeled, partially labeled and (fully) labelled examples. The partially labeled examples include all six possible combinations of known and missing values of the three targets: three examples have only one target value known and three examples have only one target value unknown.

On the methodological side, the task of SSL from fully labeled and fully unlabeled data in structured output prediction has been studied to some extent, as described above. However, the task of learning from partially labeled data has not received much interest

Table 3: *A MTR dataset with partially labeled examples. '?' denotes a missing value.*

ID	**Description Space**				**Target Space**			
#	**Hi drain q1**	**drainage factor**	**Total N input**	**...**	**Q**	**P**	**N**	**Example type**
#1	3	0.35	121.26	...	**3.86**	**0.04**	**0.47**	Labeled examples
#2	2	0.25	119.86	...	**4.28**	**0.34**	**20.71**	
#3	2	0	120.4749	...	**4.35**	**0.024**	**0.43**	
#4	2	0.24	120.95	...	**3.47**	**0.155**	**0.15**	
#5	2	0	120.26	...	**3.7**	**?**	**?**	Partially labeled examples
#6	3	0.75	121.26	...	**2.95**	**?**	5.71	
#7	2	0.565	120.47	...	**4.4**	**0.026**	**?**	
#8	3	0	116.86	...	**?**	**0.68**	**?**	
#9	3	0.547	116.56	...	**?**	**?**	**0.21**	
#10	2	0.65	118.36	...	**?**	**0.13**	**0.11**	
#11	2	0.15	119.2568	...	**?**	**?**	**?**	Unlabeled examples
#12	3	0	121.0236	...	**?**	**?**	**?**	
#13	2	0.86	119.3698	...	**?**	**?**	**?**	
#14	2	0	115.3987	...	**?**	**?**	?	

by the research community, even though it can be regarded as a special case of the semi-supervised learning task, where, besides fully labeled examples and completely unlabeled examples, there are also incompletely labeled examples. The majority of the approaches of SSL for SOP mentioned in the above subsection cannot handle partially labelled data. Fortunately, the approach of semi-supervised learning with predictive clustering is capable of also handling incompletely labeled examples, as described in Section 2.2.1.

On the applied side, in the domain of environmental sciences, there is not much research on using ML for predictive modeling in the context of partially labeled data. The need for using such methods, however, is strong as in most of the national monitoring programs, a vast amount of incomplete data is collected [10]. The few examples of applying semi-supervised learning in environmental sciences focus on single-target predictive modeling, i.e., SSL for classification and regression (cf. [24, 25]).

1.4. An application of MTR with partially labelled data in water quality assessment

To illustrate the use of SSL methods for MTR with partially labelled data, we consider an application in the assessment of water quality. This application domain is of high importance, as one of the most basic and important human needs is quality drinking water. The risk of water pollution is increasing as population growth and urbanization increase waste production and intensification of agricultural land management increases to meet food production demand.

Due to anthropogenic uses of land and management of nutrient resources and constant changes in weather and climate, aquatic ecosystems are affected continuously by eutrophication - a process caused by nutrient enrichment, which is the most significant environmental issue for surface waters [26, 27, 28, 29, 30, 31]. In the Republic of Ireland, because of the high rainfall environment, nutrient losses from agriculture fields can flow directly to ground and surface water bodies if correct management is not implemented and can cause a significant environmental risks for water quality [32, 33]. The studies that predict biological water quality, as well as nutrient loss to surface waters, use the available (mostly incomplete) pressure-pathway related data from national water quality monitoring networks [10, 34].

The water quality data in our study includes three continuous dependent/ target variables: biological quality of water, phosphorus concentration and nitrate concentration in water. The independent (descriptive) variables include environmental pressure (source) variables, such as different kinds of nutrient inputs (e.g., nitrogen input from fertilizer). They also include pathway (transport) variables, such as the net rainfall. A detailed description of the dataset can be found in Section 3 and Appendix A.

A key property of this dataset is that approximately 50% of the spatial units of analysis (i.e., examples) are only partially labeled, i.e., have missing values for some of the target variables. Therefore, this dataset is appropriate to demonstrate the utility of the advanced machine learning algorithms for semi-supervised learning which can handle partially-labeled data. The majority of methods for MTR cannot handle partially-labeled data and has to discard such partially-labeled examples from the learning process [35, 36, 37]. In our study, this would mean that almost a half of the dataset cannot be used, thus, a lot of potentially useful information would be lost. By using

SSL methods for MTR that can handle incompletely labeled examples, the discarding of partially-labeled examples can be avoided.

1.5. An outline of the paper

Having introduced the tasks of fully supervised MTR, semi-supervised MTR, and MTR from partially labeled examples, we now proceed to describe the recently introduced approach of *predictive clustering trees* (PCTs) for SSL in the context of MTR [13]. While not yet used to handle partially labeled examples, SSL PCTs for MTR can indeed handle all three types of examples in MTR: fully labeled, unlabeled and partially labeled ones. This holds for both individual PCTs for MTR and ensembles thereof: We explain how they handle all three types of examples in Section 2.2.1.

We demonstrate the utility of the methodology on the real-world problem of quality assessment for agricultural waters in Ireland: The dataset at hand indeed contains a large number of partially labeled examples. To the best of our knowledge, this is a rare example of applying SSL for MTR on a practically relevant problem. In the published literature on the subject, the advantages of SSL are demonstrated on benchmark datasets consisting of fully labelled examples, where labels are artificially removed to simulate missing labels.

In our experiments, we use supervised and semi-supervised variants of single PCTs and ensembles of PCTs for regression [11, 12, 13]. We build both multi-target (i.e., global) regression models, where all of the target variables are simultaneously predicted by a single model, and single target (i.e., local) regression models, where a separate predictive model is built for each target variable. We compare the predictive performance of the semi-supervised PCTs models learned from partially-labeled examples with the performance of standard supervised PCTs that use only labeled data instances. Moreover, in order to aid the global decision-making process, we generate predictive water quality maps for the Republic of Ireland from the obtained PCT models and discuss their predictive accuracy.

We address the following research questions:

- How the incomplete examples influence the predictive performance in addition to complete examples?

- Whether better models (in terms of interpretability, model size and predictive performance) can be obtained by using single- or multi-target regression?

- How these modelling methodologies scale from single models to ensemble of models?

The reminder of the paper is structured as follows. Section 2 describes the proposed machine learning methodology we use. Section 3 describes the case study i.e., the problem and the data used. In Section 4, we specify the experimental setup (parameter settings and model evaluation criteria etc.). Next, Section 5 presents the experimental results with interpretation and discussion on the optimal model, as well as maps depicting the original and predicted values for each target. Finally, Section 6 concludes the paper.

2. Machine learning methods

2.1. The machine learning task

In this section, we present the machine learning methodology that was used to obtain the predictive models. To begin with, we formalize the semi-supervised learning task of multi-target regression with both unlabeled and partially labeled examples. We then describe predictive clustering trees (PCTs) for MTR. Next, we present the adaptation of the variance function in PCTs to consider labeled, unlabeled and partially-labeled instances. Finally, we present ensembles of predictive clustering trees for MTR.

Machine learning methods that can handle labeled examples are known as supervised learning methods. On the other hand, methods which use unlabeled examples in addition to labeled ones with the aim to improve the performance of supervised learning methods are called *semi-supervised learning* (SSL) methods [19, 13]. Note that learning from partially-labeled data can be considered as a generalization of semi-supervised learning. Usually, in SSL, the class information for unlabeled examples is entirely missing.

Partially-labeled examples can be considered an additional source of information in the spirit of SSL. Namely, the known information of the descriptive attributes for unlabeled examples and the known information of the targets for partially-labeled examples can be exploited. This would improve the predictive performance of the model and the model itself.

The formal definition of the semi-supervised multi-target regression task learning with partially-labeled examples is as follows.

Given:

- A description (input) space $\mathcal{X}^D$ spanning D descriptive variables, i.e.,

$$\mathcal{X}^D = X_1 \times X_2 \times \cdots \times X_D,$$

where X_i is the set of possible values of the $i-th$ descriptive variable;

- A target (output) space $\mathcal{Y}^T$ spanning T target variables, i.e.,

$$\mathcal{Y}^T = (Y_1 \cup \{?\}) \times (Y_2 \cup \{?\}) \times \cdots \times (Y_T \cup \{?\}),$$

where Y_j is the set of possible values of the $j-th$ target variable, extended with potentially missing value (denoted as ?);

- A set I of N examples (x, y) where $x \in X^D$ and $y \in \mathcal{Y}^T$ and an example (x, y) is

$$\begin{cases} \text{fully- labeled,} & \text{if } \forall i \in \{1, 2, \ldots, T\} : y_i \in Y_i \\ \text{unlabeled,} & \text{if } \forall i \in \{1, 2, \ldots, T\} : y_i = ? \\ \text{partially- labeled,} & \text{otherwise} \end{cases} .$$

- A quality criterion q, which rewards the models with the lowest predictive error.

Find:

- a function $f : \mathcal{X}^D \rightarrow \mathcal{Y}^T$ by using the set of examples I, such that f maximizes the quality criterion q.

Depending on the example set I, changing with respect to the output (target) space $\mathcal{Y}^T$, we can distinguish between different MTR tasks. If there are labeled examples only in I, then we have the classical supervised multi-target regression task (MTR). We define the task of semi-supervised learning for multi-target regression (SSL for MTR) if we have only fully labeled and unlabeled examples in I in addition. [13].

In our study, f is represented with predictive clustering trees or ensembles thereof [12], which we will present in the following subsection.

2.2. Learning predictive clustering trees (PCTs) for MTR from partially-labeled data

2.2.1. Semi-supervised PCTs for multi-target regression

Blockeel [38] proposed the predictive clustering framework, while predictive clustering trees (PCTs) for multi-target regression (MTR) were proposed by Struyf and Džeroski [11]. The PCT framework views a decision tree as a hierarchy of clusters, where the top-node corresponds to one cluster containing all the data. While moving down the tree, clusters are recursively sub-divided into smaller clusters, aiming to minimize the intra-cluster variances. The PCT framework is implemented in the CLUS software toolbox (`http://source.ijs.si/ktclus/clus-public/`) [12, 11].

PCTs are learned with a standard top-down induction of decision trees (TDIDT) algorithm (see **Algorithm** 1) [39]. which takes a set of examples I as input, to produce a tree as output. The procedure starts by selecting a test (t) for the root node by using a heuristic function (h) computed on the training examples. The goal of the heuristic (h) is to select the test (t) that maximizes the variance reduction caused by the partitioning (P) of the examples into subsets according to the test outcome (see the $BestTest$ procedure in **Algorithm** 1).

The recursive procedure for partitioning the examples continues until a stopping criterion is satisfied. Further partitioning of the examples yield a tree with lower quality. In this case, the prediction (output value, calculated by applying a prototype function) is stored in the corresponding leaf of the tree. Satisfying the stopping criteria means pre-pruning of the tree in order to avoid overfitting and provide a more interpretable tree.

In PCTs for MTR, the prototype function calculates the mean values of all target variables Y for the training examples belonging to that leaf. In the prediction phase, for each new example, the algorithm identifies the leaf the example belongs to and returns the value predicted by the prototype function associated to that leaf (see **Algorithm** 1, procedure PCT, line 8).

The most significant feature of the algorithm for learning PCTs and its main difference from a standard decision tree learner, is that it considers the variance function and the prototype function as parameters that can be instantiated for a specific learning task.

Algorithm 1 The top-down induction algorithm for PCTs

```
procedure PCT
Input: A dataset $I$
Output: A predictive clustering tree
1: $(t^*, h^*, \mathcal{P}^*)$ =BestTest($I$)
2: if $t^* \neq none$ then
3:     for each $I_i \in P^*$ do
4:         $tree_i$ =PCT($I_i$)
5:     end for
6:     return node($t^*, \bigcup_i \{tree_i\}$)
7: else
8:     return leaf(Prototype($I$)))
9: end if

procedure BestTest
Input: A dataset $I$
Output: the best test ($t^*$), its heuristic score ($h^*$) and the partition ($\mathcal{P}^*$) induced on $I$ by ($t^*$)
1: $(t^*, h^*, \mathcal{P}^*) = (none, 0, \emptyset)$
2: for each possible test $t$ do
3:     $\mathcal{P}$ = partition induced by $t$ on $I$
4:     $h = Var(I) - \sum_{I_i \in \mathcal{P}} \frac{|I_i|}{|I|} Var(I_i)$
5:     if $(h > h^*) \wedge$ Acceptable $(t, \mathcal{P})$ then
6:         $(t^*, h^*, \mathcal{P}^*) = (t, h, \mathcal{P})$
7:     end if
8: end for
9: return $(t^*, h^*, \mathcal{P}^*)$
```

For the semi-supervised task of MTR, PCTs are learned by using a modified version of variance reduction, where variance is calculated as an aggregation of the variances of both the target attributes and descriptive attributes (see **Algorithm** 1, procedure $BestTest$, line 4). For each set of instances I, the variance function is calculated as a weighted sum of the variance functions over both the target (Var_f^Y) and the descriptive space (Var_f^X), by using the following formula:

$$Var_f(I) = \mathbf{w} \cdot Var_f^Y(I) + (1 - \mathbf{w}) \cdot Var_f^X(I) \tag{1}$$

where the weight parameter **w** controls the level of supervision employed during the tree learning phase. It ranges from 0 (no supervision) to 1 (full supervision). The ability to control the amount of supervision during the tree learning phase by using **w** parameter is important, because there are different types of datasets from different domains with different numbers of labeled examples, which may require different amount of supervision.

The variance of the i^{th} target attribute, that takes into account also 'incomplete' examples is calculated as follows:

$$Var_i(I) = \frac{\sum_{j=1}^{K_i} (y_{i,j})^2 - \frac{1}{K_i}(\sum_{j=1}^{K_i} y_{i,j})^2}{K_i} \tag{2}$$

where $y_{i,j}$, is the value of the i^{th} target variable for the j^{th} example and K_i is the number of examples with non-missing (known) values of the i^{th} target variable [13].

The variance function over the descriptive space ($Var_f^X(I)$) is defined with respect to the different types of descriptive attribute values. They can be either nominal or numerical. Therefore, $Var_f^X(I)$ is a sum of variance scores of the numerical variables and Gini scores of the nominal variables, i.e.,

$$Var_f(I, X) = \frac{1}{D}\left(\sum_{X_i \text{ is numeric}} Var_i(I) + \sum_{X_j \text{ is nominal}} Gini_j(I)\right) \tag{3}$$

Var_i is calculated as in Eq 2, whereas $Gini_j$ is calculated as follows:

$$Gini_j(I) = 1 - \sum_{k=1}^{C_j} \tilde{p_k} \tag{4}$$

where C_j is the number of categorical values of the descriptive attribute X_j and $\tilde{p_k}$ is the empirical probability of the value c_k, calculated by considering only the K_i examples with known/partially-known values [13].

Note that, when inducing a semi-supervised regression tree, some extreme cases may occur. First, a leaf of a tree may contain only examples with unknown values for some target attribute. In that case, the prototype function calculates the prediction by returning the prototype function value from the first parent node of such a leaf that contains labeled instances. The extreme can occur in the procedure *BestTest*, i.e, when the candidate split has to be evaluated and all examples from one of the branches of that test have missing values for a target attribute. This is handled by estimating the variance in the current node with the variance of the parent node.

2.2.2. Ensembles of PCTs for MTR

Random forests of PCTs, proposed by Kocev et al. [12], are constructed following the standard random forest algorithm (Algorithm 2) proposed by Breiman [40]. A random forest (Algorithm 2) represents an ensemble of trees in which the diversity between trees is achieved by using bootstrap replicates of the training dataset and also by taking a random subset of descriptive attributes at each node of the tree during the learning process. Bootstrap replicates are obtained by randomly sampling instances from the training set, with replacements, until the same number of examples as in the original training set is obtained.

Algorithm 2 The learning algorithms for random forests of PCTs (RForest) Here, I is the set of training examples, k is the number of trees in the forest, $f(D)$ is the size of the subset of the descriptive space considered at each node during tree construction.

```
procedure RForest($I, k, f(D)$)
returns Forest
$F = \emptyset$
for $i = 1$ to $k$ do
    $I_i = bootstrap(I)$
    $T_i = PCT\_rnd(I_i, f(D))$
    $F = F \bigcup T_i$
end for
return $F$
```

The difference between the PCT procedure for tree construction in the random forest algorithm Algorithm 2 and the PCT procedure in the TDIDT Algorithm 1 is in the selection of attributes. In Algorithm 2 selection of the attributes is randomized, i.e., the classical *PCT* procedure is replaced by *PCT_rnd*(). Namely, at each node in the decision tree, a random subset of attributes is taken from the descriptive space and the best test is selected using this attribute subset. There are various ways of choosing the number of randomly selected attributes, given the total number of descriptive attributes D, e.g., $f(D) = [\sqrt{D} + 1]$, $f(D) = [\log_2 D + 1]$, etc.

The prediction for a new instance in the random forest algorithm of PCTs is made by combining the predictions of all base predictive models. In the multi-target regression (MTR) task, the prediction of each target is defined as an average of the predictions obtained from each predictive tree. The prediction for all targets is a vector of such predictions.

3. The case study

3.1. Motivation and problem description

One of the main debates about use and management of Ireland's rural environment is based around the impact of agriculture on water quality. According to the European Protection Agency [41], 69% of Irish rivers are classified as 'unpolluted' by European standards, which makes quality of water reasonably good. Nutrient enrichment (by nitrogen and phosphorus), which may lead to eutrophication processes, is the main threat to the water quality directly affecting the aquatic ecosystem [42]. Although the average concentration of nitrogen (N) is typically below the maximum limit value for drinking water ($mg \cdot l^{-1}$), the potential impact of N loss on eutrophication cannot be ignored. The average molybdate reactive phosphorus (P) concentration more frequently violates the maximum limit threshold for eutrophication, $mg \cdot l^{-1}$ [43]. According to the Environmental Protection Agency [44], there is evidence of an increasing trend of P concentration in rivers, in recent years.

To the best of our knowledge, there are only few studies that use the pressure and pathway variables in order to manage the nutrient loads produced by agricultural management practices. None of them uses advanced machine learning techniques that provide explainable models for predicting potential nutrient loss by using the existing data that comprises the most important pressure-pathway factors. Schulte et al. [10] present a pressure-pathway model for control of nutrient enrichment. They quantify and map the agro-meteorological (pressure and pathway) indicators by controlling and evaluating the effect of nutrient input to Irish water bodies. Daly et al. [34] develop an empirical model for predicting only molybdate reactive phosphorus (MRP) by using data collected from 35 different water catchments in Ireland. The data in their study are clustered into two different groups based on soil type. FThe frst group consists of predominantly poorly-drained soil samples and has higher water phosphorus concentrations, while the second group has predominantly well-drained soil samples and slightly lower water phosphorus concentrations. At the end, different models for each group are developed with explained variability in data, i.e., coefficient of determination (R^2), of 62% and 68%, respectively.

Recently, some efforts have been made in using machine learning methods in order to assess water quality. For example, Giri et al. [35] evaluated the impact of land uses on stream integrity by applying several known ensemble techniques (random forests, boosted regression trees, etc.) for learning models from land use and land cover change data from aerial photography. Chen et al. [36] proposed a comparative analysis for assessing the performance of tree learning techniques (decision trees, random forests, deep cascade forest, etc.) for predicting surface water quality. Their data consisted of 30.000 examples taken from 10 national large rivers and lakes in China, where chemical water quality indicators were measured. Furthermore, Sheng-chou et al. [37] proposed a study for predicting water quality in reservoirs. Water quality data collected at stations of 20 reservoirs in Taiwan were used by machine learning algorithms, i.e., artificial neural networks (ANNs) and support vector machines (SVMs) for supervised classification and regression, and basic linear regression in both baseline and ensemble scenarios.

All of the above studies exclude 'incomplete' examples with missing values for the target variables in the learning process.

3.2. Data description

In our study, we use predictive clustering trees (PCTs), an advanced machine learning technique that simultaneously generates a prediction for the response variables (targets) for new examples and groups the existing examples based on values of the important controlling factors (in this case nutrient pressure and/or pathway for loss to water variables). The trees were constructed based on water quality monitoring data collected of the national level in Ireland, generated by the Environment Protection Agency in the period, 2001– 2003, comprised 708 observations corresponding to different parts of Ireland, with associated nutrient pressure and pathway for loss to water variables. The data were provided by the Spatial Analysis Group at TEAGASC, the Irish Agriculture and Food Development Authority in Ireland. These data were harmonized and adjusted for the need of a study by Schulte et al. [10]. The land area of Ireland was divided into grid cells of 10x10km. All target and descriptive variables were expressed as mean values within each grid cell.

These data consist of 3 continuous target (response) variables and 26 continuous descriptive (pressure-pathway) variables. The target (response) variables are the following:

1. *Biological water quality*, expressed by a Q-value, which ranges from 1 (very poor) to 5 (very high quality). The Q-value is based on biological observations;
2. *Phosphorus concentration*, expressed as "Molybdate Reactive P" or P, which is application of phosphorus to the land measured in *mg* per liter ($mg \cdot l^{-1}$).
3. *Nitrate concentration*, expressed as "NO_3" or N, which is application of nitrogen to the land, measured in *mg* per liter ($mg \cdot l^{-1}$);

We deal with these 3 targets either independently, i.e., in a separate per-target analyses (one model per target), or simultaneously, predicting all of them (one model for all targets). Not all of the three targets were measured at every observation point. Three different types of examples are presented in the dataset based on the missing values in their target space.

The Q-value was measured in 706 grid cells, the phosphorus concentration, i.e., P value in 529 grid cells, and the nitrogen concentrations, i.e., N value in 352 grid cells. For 351 examples, the values of all target variables (Q, P, and N) were available. The distributions of the original (measured) values for the target (response) variables are shown in Figure 1.

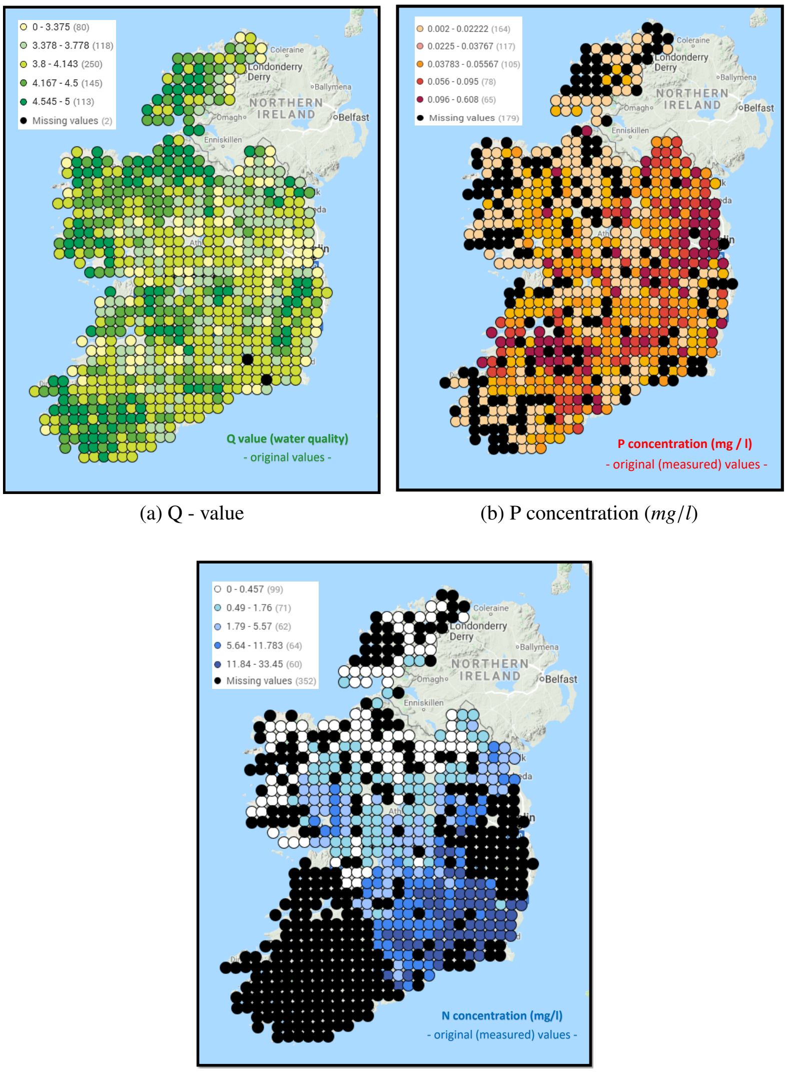
Conceptually, we link water quality to agro-meteorological data and nutrient pressures through a which are controlled of by pressure-pathway variables. In short, this means that the risk of nutrients to water is was highest when the nutrient source pressure of nutrients and the pathways to transport it, are were both present at the same place at the same time. Based on the findings in [10], for N, both the pressures and the pathways are generally highest in the east and south-east of the country, so it is not surprising that the highest nitrate concentrations in water are found here (see Figure 1c). For P, the situation is less clear: the pathways for P loss are were of potentially higher, but source pressure may be lower, intensity in the west and north-west parts of the country. However, the P-pressure follows a pattern similar to the N-pressure, i.e., it is highest in the east and south-east. Apparently, while P source pressures are likely to be higher, the loss pathways for P seem to be adequate also in the east and south- east to transport part of this pressure to water bodies, resulting in highest P concentrations in water in these regions of Ireland, because the highest P concentration, like N, are found in the east and south-east part (see Figure 1b). Thus, as expected, the higher concentration of nutrients (phosphorus and nitrogen) in east and south-east leads to lower biological water quality of the water, i.e. the Q-value, as compared to the west and south-west regions, where the nutrient enrichment is much lower (Figure 1a).

Descriptions of the individual descriptive pressure-pathway variables can be found in the Appendix.

4. Experimental design

4.1. Evaluation Framework: Tasks and algorithms

In this section, we explain the design of the mschine learning experiments performed on the dataset at hand, consisting of 'complete' (i.e., labeled) and 'incomplete' (i.e., unlabeled and/or partially-labeled) examples. In order to distinguish between the different learning tasks considered, we use the prefixes *SL-* (supervised learning), *SSLPL-* (semi-supervised learning with partially-labeled examples) and *SSL-* (semi-supervised learning with unlabeled examples) before the name of the learning task, for both single-target (ST) and multi-target (MT) regression. The difference in *SSLPL* and *SSL* is that, in the former, we consider the MTR task with partially-labeled and fully-labeled examples, and in the latter, we consider the single-target regression task where unlabeled and labeled examples, are used. Furthermore, we use the suffixes *-PCT* and



(a) Q - value

(b) P concentration (mg/l)

(c) N concentration (mg/l)

Figure 1: Distribution of the original (measured) data examples for the target values (Q-value, P and N concentration), including 'incomplete' examples with missing (unknown) values for the target variable.

-RF after the learning task to denote the use of a single PCT or a random forest of PCTs to solve the task.

The overall evaluation framework we adopt is depicted in Figure 2.

In the SSL and SSLPL scenarios, we first optimize the weight parameter **w** in the CLUS algorithm (which controls the level of supervision) by internal 3-fold cross validation on the training set. The final model is then built by using the entire training set and the selected **w** parameter.

We learn the following four types of individual PCTs: *SL-ST-PCT*, *SL-MT-PCT*, *SSL-ST-PCT* and *SSLPL-MT-PCT*.

Analogously, we build four types of random forest models, denoted as follows: *SL-ST-RF*, *SL-MT-RF*, *SSL-ST-RF* and *SSLPL-MT-RF*.

We consider the following values for **w**: 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1.0.

When learning individual trees, we specify the minimal number of examples per leaf in a PCT to be 16, in order to produce smaller, more general trees, where only the most important set of descriptive attributes are in the tree nodes. We construct random forests consisting of 100 predictive clustering trees. Trees in the random forests are not pruned, and the number of selected random features for consideration at each internal node is set to $[\log_2(D) + 1]$, where D is the total number of descriptive attributes [40].

4.2. Evaluating performance: Procedure and metrics

We estimate the predictive performance of the learned models by 10-fold cross validation, where 9 folds are used for training and the remaining one for testing it. The procedure is repeated 10 times so that each fold is used exactly once as a test set. The reported results represent the average performance across all 10 runs.

In order to obtain comparable results, we adjust the folds as followasfor *SL-MT-PCT, SL-ST-PCT*, as well as their respective counterparts in the ensemble setting *SL-MT-RF, SL-ST-RF*, we use only labeled instances, i.e., the folds with only labeled examples *per se*. For *SSLPL-MT-PCT* and *SSLPL-MT-RF*, we remove 'incomplete' instances (with missing values) for each target from the test sets for each fold, accordingly. For *SSL-ST-PCT* and *SSL-ST-RF* we enrich the train sets with the unlabeled examples from the test sets and evaluate only on labeled examples. Note that the same

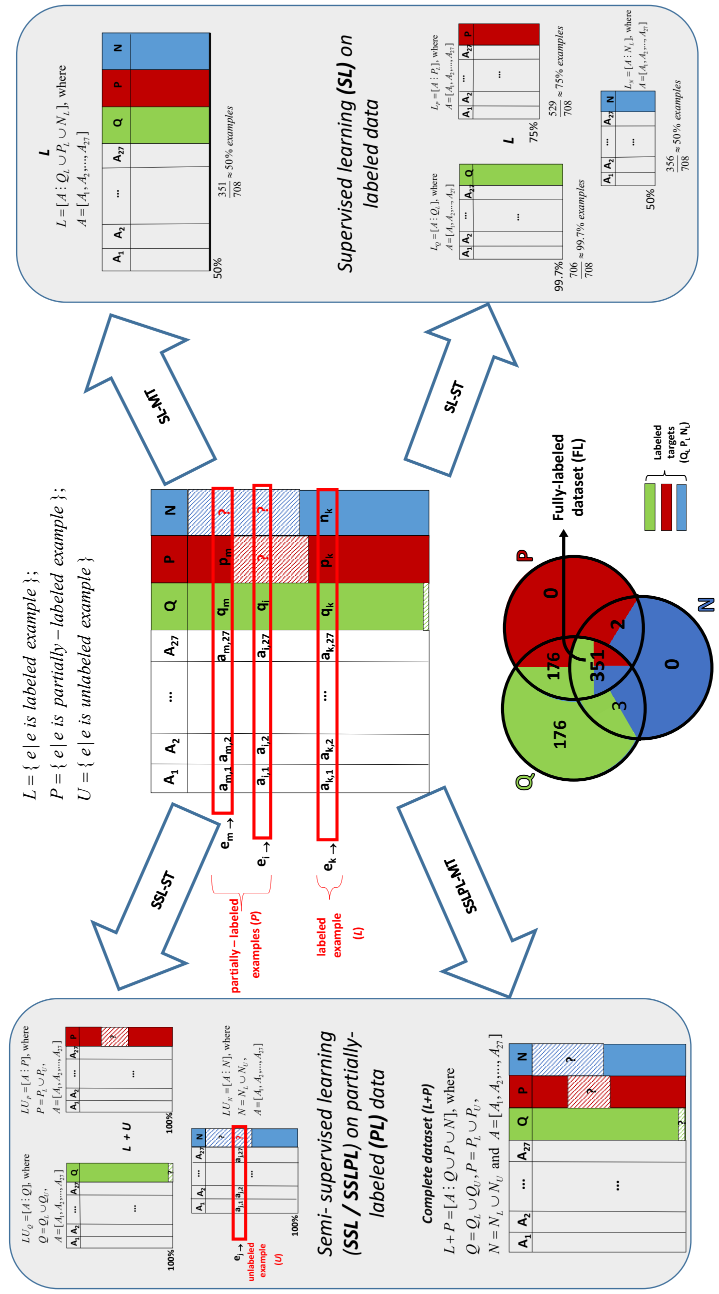

L = { e | e is labeled example };
P = { e | e is partially – labeled example };
U = { e | e is unlabeled example }
A1 A2 … A27 Q P N
em → am,1 am,2 am,27 qm pm ?
el → al,1 al,2 al,27 ql ? ?
ek → ak,1 ak,2 … ak,27 qk pk nk
partially – labeled examples (P)
labeled example (L)
Fully-labeled dataset (FL)
Labeled targets (Q, P, N)
Q P N
176 176 0 351 3 2 0
SL-MT
SL-ST
SSL-ST
SSLPL-MT
Supervised learning (SL) on labeled data
L
L = [A : QL ∪ PL ∪ NL], where A = [A1, A2, …, A27]
351/708 ≈ 50% examples
50%
LQ = [A : QL], where A = [A1, A2, …, A27]
706/708 ≈ 99.7% examples
99.7%
LP = [A : PL], where A = [A1, A2, …, A27]
529/708 ≈ 75% examples
75%
LN = [A : NL], where A = [A1, A2, …, A27]
356/708 ≈ 50% examples
50%
Semi- supervised learning (SSL / SSLPL) on partially-labeled (PL) data
LUQ = [A : Q], where Q = QL ∪ QU, A = [A1, A2, …, A27]
LUP = [A : P], where P = PL ∪ PU, A = [A1, A2, …, A27]
LUN = [A : N], where N = NL ∪ NU, A = [A1, A2, …, A27]
L + U
100%
eu → unlabeled example (U)
Complete dataset (L+P)
L + P = [A : Q ∪ P ∪ N], where Q = QL ∪ QU, P = PL ∪ PU, N = NL ∪ NU and A = [A1, A2, …, A27]

Figure 2: Experimental evaluation framework. The different types of data subsets and corresponding learning tasks are shown in the grey coloured rectangles. The different types of examples in the dataset with respect to their labeling are shown in the table in the center. The data distribution in terms of labeling is depicted in the Venn diagram in the bottom center.

labeled examples are used in the training folds in each scenario (*SL-MT-PCT, SL-ST-PCT, SSL-ST-PCT* and *SSLPL-MT-PCT*).

We assess the predictive performance of the algorithms by using the metric of average relative root-mean-square-error (RRMSE), defined as follows:

$$RRMSE = \frac{1}{T}\sum_{i=1}^{T} RRMSE_i.$$

T is the number of target variables and $RRMSE_i$ is the relative root-mean-square-error of the i^{th} target variable, defined as follows:

$$RRMSE_i = \sqrt{\frac{\sum_{j=1}^{N_{test}}(y_{j,i} - \hat{y}_{j,i})}{\sum_{j=1}^{N_{test}}(y_{j,i} - \bar{y}_i)}}$$

where N_{test} is the number of examples in the test set, $\hat{y}_{j,i}$ is the predicted value of the j^{th} example, $y_{j,i}$ is the actual value of the i^{th} target variable for the j^{th} example of the test set and $\bar{y}_i$ is the mean of the i^{th} target variable on the training set.

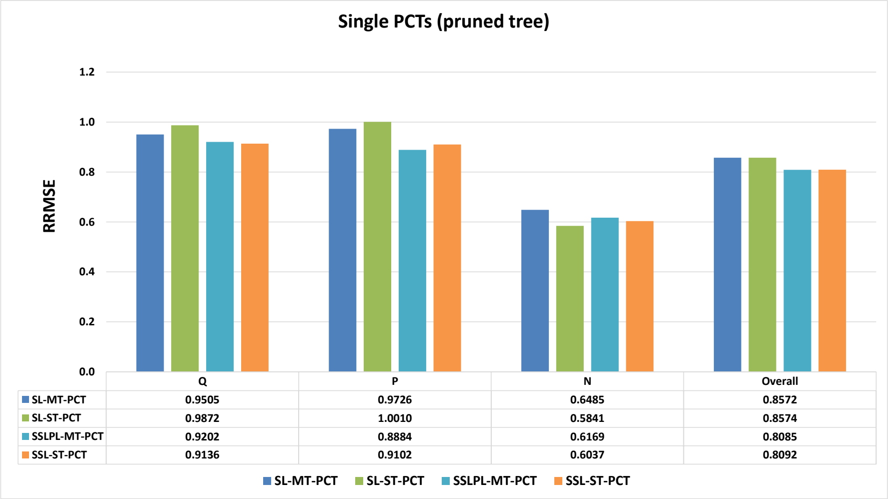
We use the *model size* as a measure of interpretability efficiency of pruned single tree models. For the multi-target PCTs, *model size* is the total number of nodes (internal nodes and leaves) and for the single-target PCTs, the *model size* is the sum of the total number of nodes across the individual trees for the three target variables.

5. Results and discussion

5.1. Predictive performance and model complexity

The predictive performance of the models is shown in Figure 3 (single PCTs) and Figure 4 (random forests). We can observe that the models able to exploit 'incomplete' examples achieve better average predictive performance than the models that discard them and use only complete (fully-labeled) examples. This observation can be made for both local and global single PCTs and local and global random forest of PCTs. Overall, the semi-supervised methods (*SSL-ST-PCT* and *SSLPL-MT-PCT*) perform better than the supervised methods, with the best performance/lowest RRMSE achieved by *SSLPL-MT-PCT* (RRMSE = 0.8085) for single PCTs and *SSLPL-MT-RF* (RRMSE = 0.7558) for random forests of PCTs.

For single PCTs, local and global methods perform almost the same overall, therefore. There is no clear conclusion whether single-target (i.e., local) or multi-target (i.e., global) PCTs perform better. For random forests of PCTs, global models (*SL-MT-RF* and *SSLPL-MT-RF*) perform slightly better than local models (*SL-ST-RF* and *SSLPL-MT-RF*).



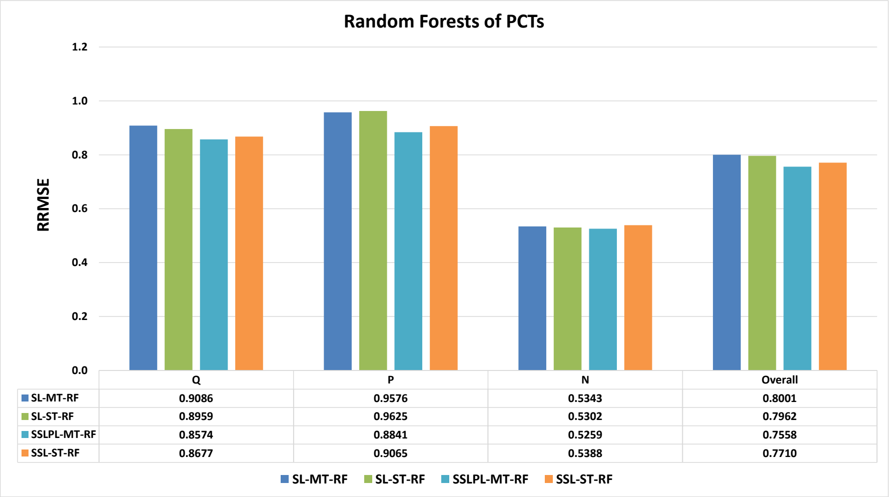
	Q	P	N	Overall
SL-MT-PCT	0.9505	0.9726	0.6485	0.8572
SL-ST-PCT	0.9872	1.0010	0.5841	0.8574
SSLPL-MT-PCT	0.9202	0.8884	0.6169	0.8085
SSL-ST-PCT	0.9136	0.9102	0.6037	0.8092

Figure 3: Predictive performance (RRMSE) of single pruned PCTs, presented separately for each of the target variables Q, P and N, and as an overall value calculated as the average RRMSE over the 3 targets.

Comparing supervised single models obtained from the fully-labeled dataset (*SL-MT-PCT* and *SL-ST-PCT*), we observe that the fully labeled multi-target (i.e., global) model achieves slightly better performance than the fully labeled single-target (i.e., local) models when predicting the Q-value and P concentration and slightly worse when predicting the N concentration. Considering the overall performance, there is no clear conclusion about the best performing supervised model, i.e., both global and local models perform similarly. A similar situation is observed for supervised ensemble models. Semi-supervised local and global single tree models performed similarly, while for semi-supervised ensemble models (i.e., random forests of PCTs), the global model (*SSLPL-MT-RF*) performed better than the local model (*SSL-ST-RF*). Overall, random forests achieve better predictive performance than individual trees.

The main advantage of SSL methods that handle ‘incomplete’ data lies in the fact that they are the most efficient in label frugal conditions where the amount of available labels is much less than 50% (e.g., 5%, 10%). For the SSL task (where the examples can be either completely labeled or completely unlabeled), Levatić et al. [13] performed an extensive experimental evaluation on different benchmark MTR datasets from various domains under different scenarios based on percent of labeled examples in the learning process (5%,10%,20% and 30% of labeled examples). The study revealed that better SSL models can be learned as compared to the supervised models in terms of both predictive performance and model size: The advantage in performance of SSL is the largest, when only a a smaller (e.g., 5%) percent of labeled examples are present in the training data. But even for relatively large proportions of labeled data, SSL still performs better, which is also clearly the case for our case study.

In domains such as environmental sciences, model interpretation plays an important role in identifying the main drivers of the response variables. In order to decide which model is the optimal, we consider a two-dimensional representation of the results, comparing the models by both their predictive performance and model complex-



	Q	P	N	Overall
SL-MT-RF	0.9086	0.9576	0.5343	0.8001
SL-ST-RF	0.8959	0.9625	0.5302	0.7962
SSLPL-MT-RF	0.8574	0.8841	0.5259	0.7558
SSL-ST-RF	0.8677	0.9065	0.5388	0.7710

Figure 4: Predictive performance (RRMSE) of random forests of PCTs, presented separately for each of the target variables Q, P and N, and as an overall value calculated as the average RRMSE over the 3 targets.

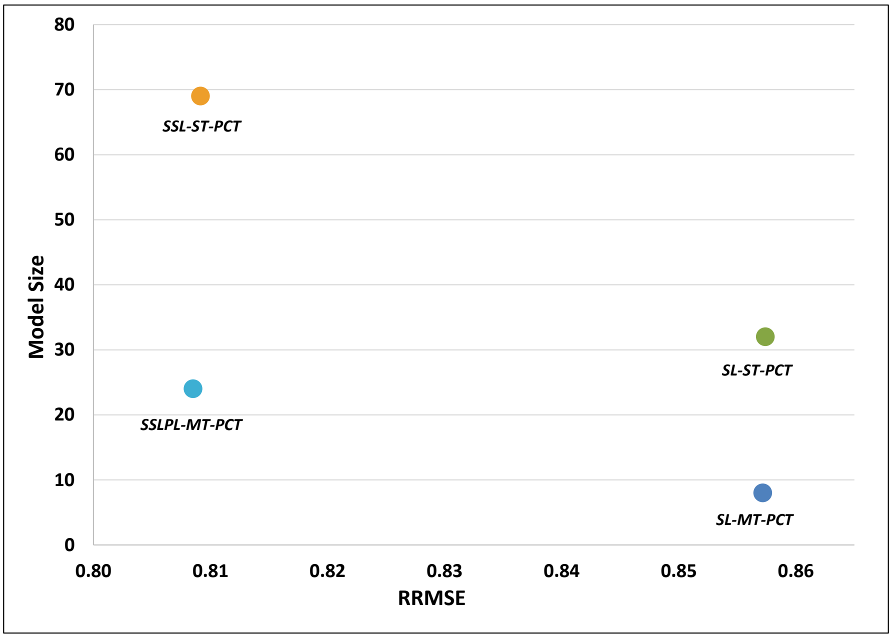


Figure 5: Comparison of single tree models along two criteria. The x-axis ahows the RRMSE values (i.e., predictive performance), while the y-axis shows model size values (i.e., model complexity).

ity (Figure 5). The obtained values for model sizes are taken as per-fold averages. We can see that when exploiting the partially-labeled examples, the obtained multi-target regression model has a per-fold average size of 24. The obtained multi-target regression model that only uses fully-labeled data (supervised model), i.e., *SL-MT-PCT* has an average size of 8. The reason for such a smaller tree is the number of learning examples - the latter approach uses only 50% of all examples. Considering model sizes obtained in the single target (i.e., local) cases, we can see that there are 32 nodes in *SL-ST-PCT* tree and 69 nodes in *SSL-ST-PCT* tree. The main observation here is the difference in model sizes between local and global models: The complexity of the global PCT models in terms of model size is significantly lower.

Moreover, in Figure 5, we observe that global and local models have similar predictive performance in both the supervised and the semi-supervised setting. However, the global model obtained in the semi-supervised setting (*SSLPL-MT-PCT*) is superior in terms of predictive performance as compared to the global model obtained in the supervised setting (*SSLPL-MT-PCT*). However, the multi-target tree learned from the fully-labeled dataset (*SL-MT-PCT*) with 8 nodes is smaller than the tree learned from partially-labeled dataset (*SSLPL-MT-PCT*) with 24 nodes. The latter tree used more data, is larger, and has better predictive performance in terms of RRMSE.

Altogether, the semi-supervised models (*SSL-ST-PCT* and *SSLPL-MT-PCT*) are superior in term of predictive performance. Global models (*SSLPL-MT-PCT* and *SL-MT-PCT*) are superior in terms of model complexity (i.e., model size).

5.2. Interpretation of the predictions through maps

From the application domain perspective, the global models are more practical, since they predict all of the target attributes simultaneously and achieve better predictive performance (i.e., generalize better). Here, we examine the water quality maps of Ireland generated by using the supervised (*SL-MT-PCT*) and semi-supervised *SSPL-MT-PCT* global models to predict the missing values. These maps are shown in Figures 6, 7 and 8. Figures 6A, 7A and 8A show the original/measured values of the water quality variables, with missing values shown in black. It is these missing values than are predicted by the two models, with predictions of *SL-MT-PCT* shown in Figures 6B, 7B and 8B and predictions of *SSLPL-MT-PCT* shown in Figures 6C, 7C and 8C, resp.

For biological water quality (Q-value), we only had two incomplete examples and therefore both models make similar predictions (see Figure 6). As a result, we focus on the maps for P and N concentrations where the differences between the predictions of the models are more notable. Figure 7 shows the maps with original (measured) and predicted values for P concentration (with the *SL-MT-PCT* and *SSLPL-MT-PCT* models). Next, Figure 8 shows the maps with original (measured) and predicted values for N concentration (with the *SL-MT-PCT* and *SSLPL-MT-PCT* models).

Schulte et al. [10] found that the pressure and pathway variables were lowest for the west and north-west sites, therefore, we expect the lowest values for P and N concentration in water in these regions. Examining the left most P concentration map (Figure 7A), there are several neighbouring unmeasured sites at the northernmost, westernmost and south-westernmost parts of Ireland. Examining the maps with the predictions, the expectations are that the unmeasured sites, which are covered/bounded by unpol-

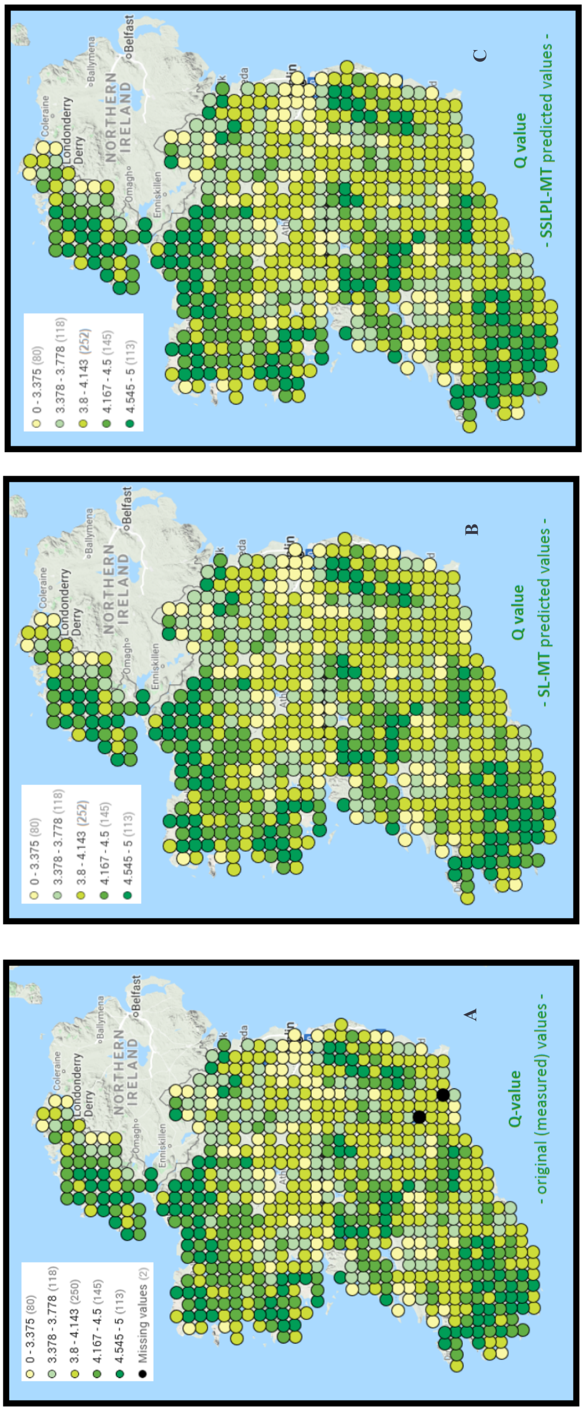


Figure 6: Maps with: (A) The original Q-values; (B) Q-values predicted by the *SL-MT-PCT* model and (C) Q-values predicted by the *SSLPL-MT-PCT* model

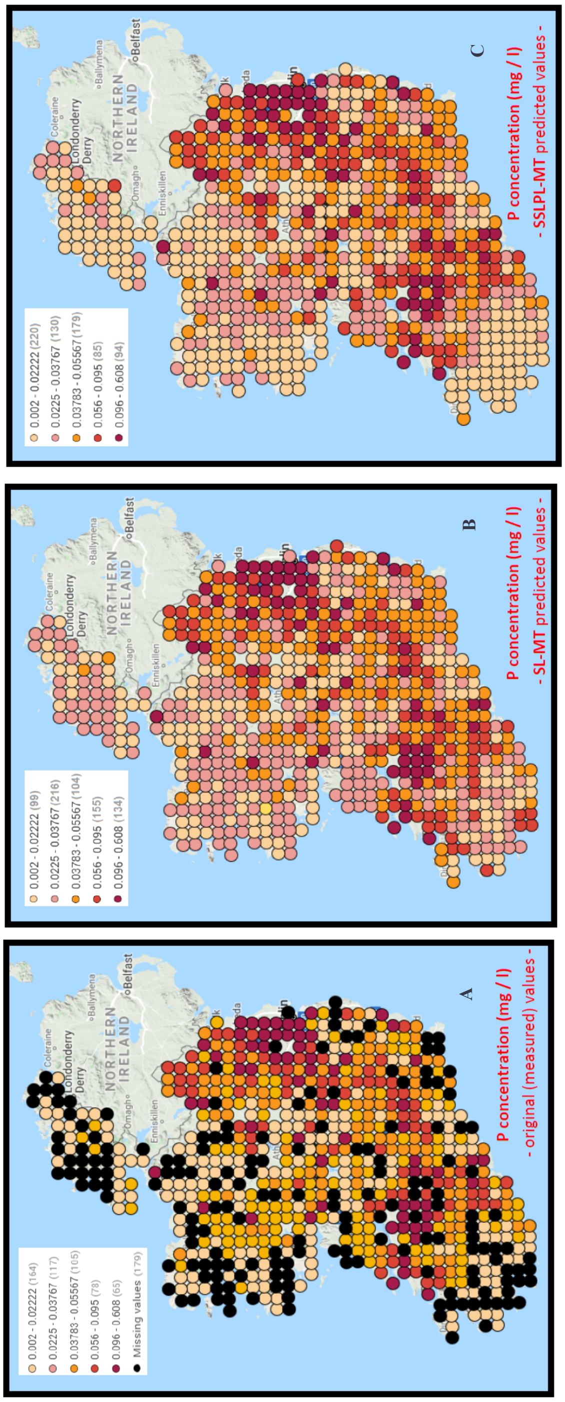


Figure 7: Maps with: (A) The original P concentration values; (B) P concentration values predicted by the *SL-MT-PCT* model and (C) P concentration values predicted by the *SSLPL-MT-PCT* model

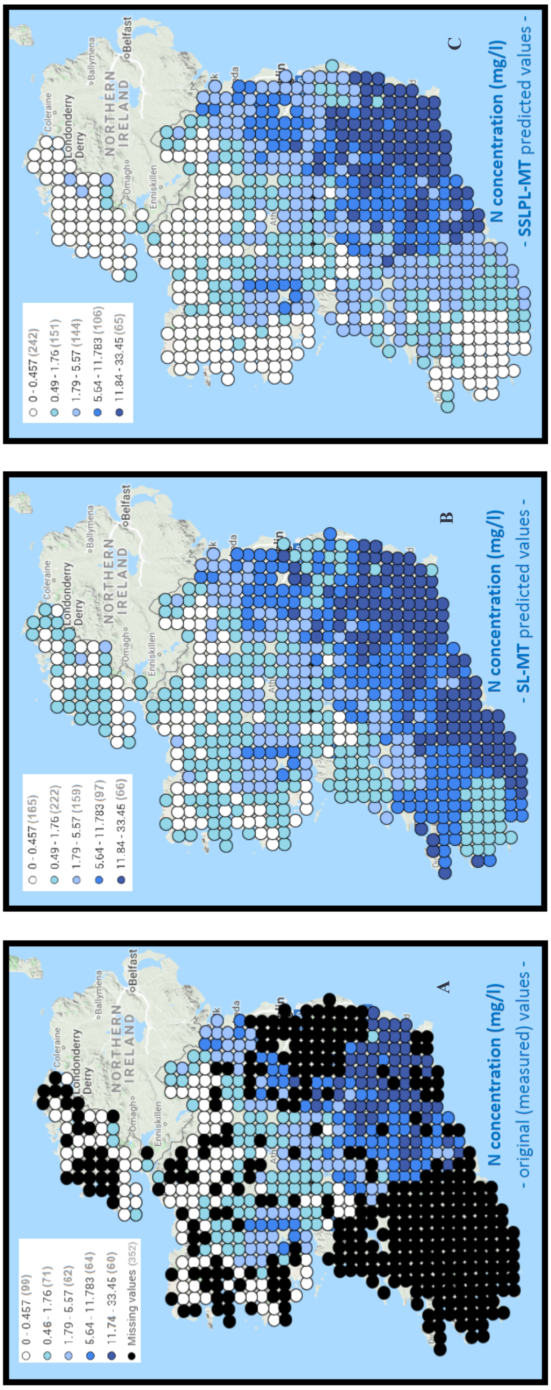


Figure 8: Maps with: (A) The original N concentration values; (B) N concentration values predicted by the *SL-MT-PCT* model and (C) N concentration values predicted by the *SSLPL-MT-PCT* modeln

luted sites, should be also unpolluted, because of the natural neighbourhood influence on water pollution. Examining the Figure 7B map, with predictions obtained by the supervised *SL-MT-PCT* model, we see that unmeasured sites are predicted as more polluted than the measured sites in the neighbourhood, which is not the case in the Figure 7C map, generated by using the predictions of the *SSLPL-MT-PCT* model, where the unmeasured sites are predicted as expected, i.e., to have similar P concentration as the measured neighbouring sites.

We have an even clearer situation in the map with predictions for the N concentration (Figure 8), where almost all sites in the south-western part of Ireland do not have measured values. Again, the *SSLPL-MT-PCT* model (Figure 8C) predicts these sites as less polluted as compared to the predictions generated by using the *SL-MT-PCT* model (Figure 8B). The unmeasured group of sites in the eastern part are expected to have higher P concentration and both maps confirm these expectations, with the difference that the *SSLPL-MT-PCT* model (Figure 8C) predicts those values more in line with the expectations as compared to the *SL-MT-PCT* model (Figure 8B). This confirms the validity of the global *SSLPL-MT-PCT* model, learned from partially-labeled examples.

5.3. Model interpretation

The multi-target (i.e., global) regression trees provide domain information about attribute connections and the main drivers of the processes regulating the water quality in receiving waters in Ireland. The semi-supervised multi-target regression PCT model was the most accurate interpretable model obtained, therefore, we will interpret only the *SSLPL-MT-PCT* model. The supervised multi-target regression model (*SL-MT-PCT*) is provided in Figure 10 in the Appendix for reference.

Figure 9 shows the multi-target regression tree for predicting the mean biological water quality (Q), phosphorus concentration (P) and nitrogen concentration (N) in receiving water bodies. This tree model was learned by using pressure-pathway variables consisting of soil, weather and managements factors from the partially- labeled dataset (scenario *SSLPL-MT-PCT*). By examining the tree, we can see that the data samples, i.e., river water quality monitoring points, were initially split according to the condition $sumPinput\_ha\_total > 14.5 kg/ha$.

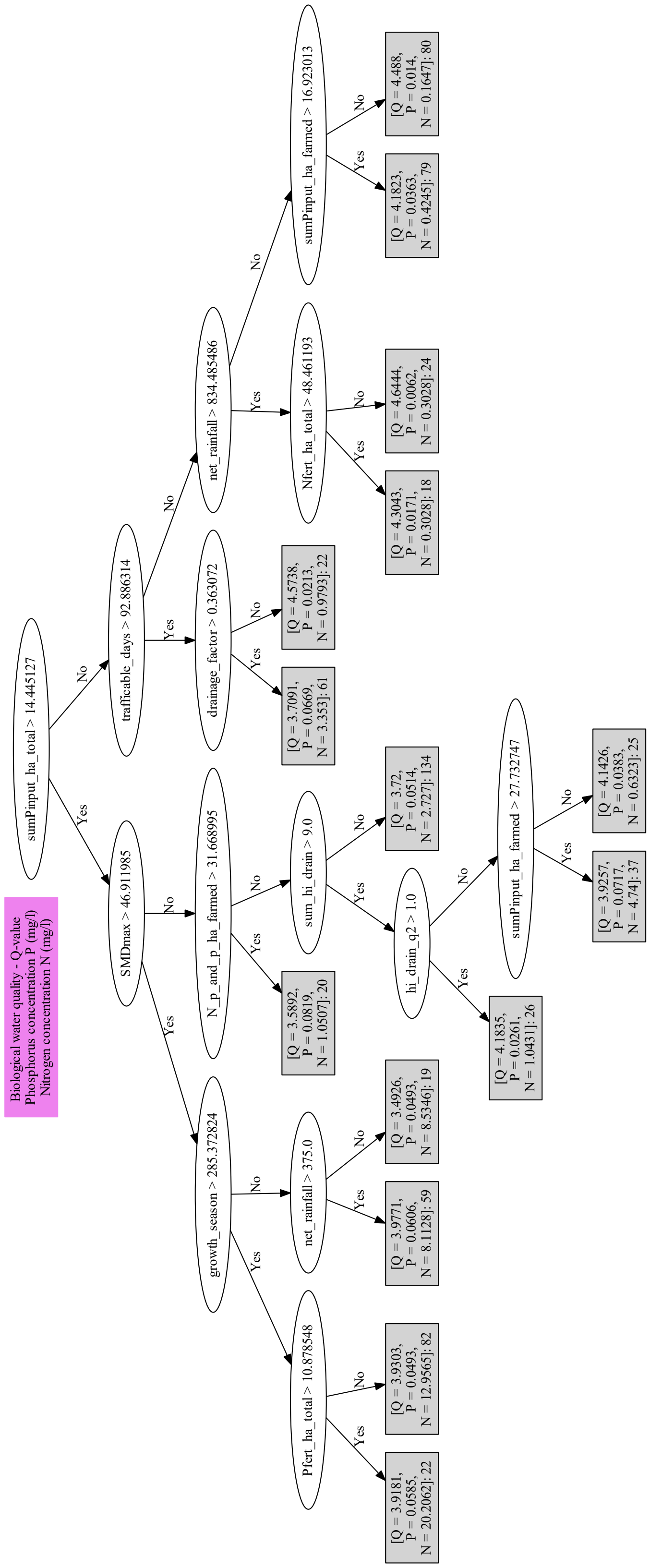

Biological water quality - Q-value
Phosphorus concentration P (mg/l)
Nitrogen concentration N (mg/l)
sumPinput_ha_total > 14.445127
No
trafficable_days > 92.886314
No
net_rainfall > 834.485486
No
sumPinput_ha_farmed > 16.923013
No
[Q = 4.488, P = 0.014, N = 0.1647]: 80
Yes
[Q = 4.1823, P = 0.0363, N = 0.4245]: 79
Yes
Nfert_ha_total > 48.461193
No
[Q = 4.6444, P = 0.0062, N = 0.3028]: 24
Yes
[Q = 4.3043, P = 0.0171, N = 0.3028]: 18
Yes
drainage_factor > 0.363072
No
[Q = 4.5738, P = 0.0213, N = 0.9793]: 22
Yes
[Q = 3.7091, P = 0.0669, N = 3.353]: 61
Yes
SMDmax > 46.911985
No
N_p_and_p_ha_farmed > 31.668995
No
sum_hi_drain > 9.0
No
[Q = 3.72, P = 0.0514, N = 2.727]: 134
Yes
hi_drain_q2 > 1.0
No
sumPinput_ha_farmed > 27.732747
No
[Q = 4.1426, P = 0.0383, N = 0.6323]: 25
Yes
[Q = 3.9257, P = 0.0717, N = 4.74]: 37
Yes
[Q = 4.1835, P = 0.0261, N = 1.0431]: 26
Yes
[Q = 3.5892, P = 0.0819, N = 1.0507]: 20
Yes
growth_season > 285.372824
No
net_rainfall > 375.0
No
[Q = 3.4926, P = 0.0493, N = 8.5346]: 19
Yes
[Q = 3.9771, P = 0.0606, N = 8.1128]: 59
Yes
Pfert_ha_total > 10.878548
No
[Q = 3.9303, P = 0.0493, N = 12.9565]: 82
Yes
[Q = 3.9181, P = 0.0585, N = 20.2062]: 22

Figure 9: Multi-target regression tree learned by exploiting the partially-labeled examples in addition to the fully-labeled examples in the $SSLPL - MT - PCT$ scenario.

In general, the group of monitoring points that had lower total P input per ha (*No* branch) had higher water quality, as indicated by their Q values and average P and N concentrations. When we consider the eutrophication indicator based on the MRP threshold ($0.035 mg/l$) [43, 42], the 284 sites with P inputs $\leq 14.5 kg/ha$ had a mean water P concentration below this threshold ($MRP = 0.0317 mg/l$), while the 424 sites with P inputs $> 14.5 kg/ha$ had a mean water P concentration greater than the MRP threshold ($MRP = 0.0534 mg/l$).

At the node in the left branch of the tree, the areas associated with water quality monitoring points with higher fertilizer inputs ($sumPinput\_ha\_total > 14.5 kg/ha$, were further divided based on a *SMDmax* threshold of 46.91, related to sites with high or lower soil moisture deficit (SMD) during the summer period. High SMD during summer has been shown to contribute to increased N-losses during the subsequent winter period from managed soils [45], with higher fertilizer inputs. The monitoring points corresponding to *SMDmax*>46.91 had higher overall N concentrations. Moreover, according to the next splitting node, $growth\_season > 285$ days, the land, corresponding to the monitoring points that follow the left (*Yes*) branch, are intensively managed (high fertilizer inputs), have relatively lower rainfall, and are likely to have a high proportion of freely draining soils conductive to a long growing season. The final splitting node for these monitoring points (for left), shows that areas with higher fertilization ($Pfert\_ha\_total > 10.88 kg/ha$ had higher average water N and P concentrations ($P = 0.0585 mg/l; N = 20.2 mg/l$) but similar average Q value of 3.9 as compared to sites with somewhat lower fertilization ($Pfert\_ha\_total \leq 10.88 kg/ha; Q = 3.9, P = 0.0493 mg/l, N = 12.96 mg/l$). Both of these kinds of areas had much higher associated average water N concentrations than areas with shorter growing season ($growth\_season \leq 285$ days). At the monitoring points corresponding to areas with shorter growing season, the final split shows that higher rainfall ($net\_rainfall > 375$) leads to lower water N concentration, possibly due to dilution, slightly higher average P concentration, and overall higher average Q value (3.97), as compared to similar areas where the net rainfall is lower ($net\_rainfall \leq 375$).

Exploring the tree for areas with lower SMD ($SMDmax \leq 46.91$), i.e., areas with wetter soil conditions, we find changes in water P concentrations and Q-values dominate. Lower SMD in these areas is possibly due to a combination of poorly drained soils, higher water holding capacity of the soils and more frequent or higher rainfall levels in these areas. The predictions of water N concentrations show lower average values (less than $5mg/l$) compared to dryer areas that go to the left side of the tree, discussed previously ($SMDmax > 46.91$). At the next node on this branch of the tree, the monitoring points are split based on the organic N input (organic manure) from pig and poultry ($N\_p\_and\_p\_ha\_farmed$) and threshold of $31.67kgN/ha$). As expected, where there is higher average N input from pig and poultry manure ($> 31.7kgN/ha$) the average water quality was lower in terms of Q-value and P concentration ($Q = 3.5892; P = 0.0819mg/l$) compared to where N input from pig and poultry manure was lower ($\leq 31.67kgN/ha$).

The areas receiving lower pig and poultry inputs ($N\_p\_and\_p\_ha\_farmed \leq 31.67$ kgN/ha) are further split, by the number of intensive soil drainage events ($> 15mm$ per day) occurring ($sum\_hi\_drain > 9$). The soil drainage events indicator has been closely related to P loss, and could account for up to 90% of the total annual P loss within 4 or 5 events ([10]). For sites with few intensive soil drainage events ($sum\_hi\_drain \leq 9$), the average water P concentration was relatively high of $P = 0.0514mg/l$ ($n = 134$ monitoring points), when compared to the MRP threshold of $0.035mg/l$. In general, these sites had lower average Q-values ($Q = 3.72$) compared to sites with higher number of intensive drainage events. The sites with many intensive drainage events, where a higher proportion of these intensive drainage events took place in the spring ($hi\_drain\_q2 > 1$) had lower N nutrient concentrations and higher biological water quality ($Q = 4.1835; P = 0.0261mg/l; N = 1.04mg/l$). Among the sites with a lower proportion of these events in the spring period ($sum\_hi\_drain > 9$ and $hi\_drain\_q2 \leq 1$), those which received high levels of fertilizer ($sumPinput\_ha\_farmed > 27.73kgP/ha$) had concentrations of both P and N that were much higher ($n = 37; Q = 3.93; P = 0.0717mg/l; N = 4.74mg/l$) as compared to sites where fertilizer levels were lower ($n = 25; Q = 4.1426; P = 0.0383mg/l; N = 0.63mg/l$).

Examining the right side of the tree, where the sites received lower overall fertilizer inputs ($n = 294; sumPinput\_ha\_total < 14.445kgP/ha$) indicating less intensive agricultural management, shows that these sites had generally lower N and P concentration in water and higher Q-values. At the first node on this branch, the monitoring points are split based on the number of $trafficable\_days$ with threshold of 92.9 days. Trafficable days is an indicator of wetness of the soil in the landscape: A higher number of trafficable days enables more intensive farm machinery operations and/or grazing by animals. The monitoring sites associated with areas that have higher number of trafficable days (> 92.9) are further split by soil drainage ($drainage\_factor$) with a threshold of 0.363. Areas with higher proportion of well-drained soils ($drainage\_factor > 0.363$) had higher water P and N concentrations ($n = 61; P = 0.0669mg/l; N = 3.35mg/l$) as compared to the areas with higher proportion of poorly-drained soils ($drainage\_factor \leq 0.363; n = 22; P = 0.0213mg/l; N = 0.98mg/l$). Additional analysis of the 61 areas with higher proportion of well-drained soils showed that these areas had received relatively high N inputs by animals, i.e., excreted N and also P loading typically directly deposited when animals are grazing (average $N\_c\_and\_s\_ha\_farmed = 94kgN/ha$ and $P\_c\_and\_s\_ha\_farmed = 14.2kgP/ha$). While these areas have lower overall intensity relative to the areas described by the main left branch of the tree, this analysis indicates that soils that are dryer for longer periods and that have better drainage properties are more likely to be farmed with higher management intensity and therefore have higher nutrient source pressures. The more poorly-drained areas ($drainage\_factor \leq 0.363$), had much lower total nutrient input pressure ($sumNinput\_ha\_total = 74kgN/ha$ and $sumPinput\_ha\_total = 9.75kgP/ha$) and hence had higher water quality ($Q = 4.5738; P = 0.0213mg/l; N = 0.9793mg/l$).

For the areas with lower number of trafficable days (i.e., $trafficable\_days \leq 92.88$ in general, the concentrations of N and P in water at the monitoring points were low and water quality was high ($Q > 4.18$). Areas with wetter soil are typically much less intensively farmed due to soil type, high rainfall and drainage limitations. Where $net\_rainfall > 834mm$, in the absence of high nutrient source pressure, further dilution of nutrient concentration in receiving water is likely and we have relatively high water quality ($n = 42; Q > 4.3$). However, the model further splits these sites based on N fertilizer inputs ($Nfert\_ha\_total$) with the threshold of $48.46kg/ha$. Although this threshold represents a relatively low annual N fertilizer rate, the monitoring points associated with higher N fertilizer inputs ($> 48.46kg/ha$) showed similar N concentrations, higher P concentrations and lower Q values ($n = 18; P = 0.0171 mg/l; N = 0.30 mg/l$) as compared to the areas receiving lower N fertilizer inputs ($n = 24; P = 0.0006 mg/l; N = 0.30 mg/l$). The slightly elevated water P concentrations indicate that increased grazing animal production, as a result of the N fertilizer applications for grass production, may be leading to mobilization of soil P, or P from grazing animal feces or particulate P in sediments disturbed by these animals. Finally, when $net\_rainfall \leq 834 mm$, the 159 monitoring points remaining are split based on the total amount of P input ($sumPinput\_ha\_farmed$). As expected, higher P fertilizer input ($sumPinput\_ha\_farmed > 16.9 kgP/ha$) resulted in slightly higher P loss (79 sites with $P = 0.0363 mg/l$) and lower Q-values ($Q = 4.1$) as compared to areas with lower P inputs ($sumPinput\_ha\_farmed \leq 16.9 kgP/ha$) (80 sites with $P = 0.014 mg/l; Q = 4.5$).

Overall, the predictive clustering tree generated from the data collected for the national water quality monitoring network points was in accordance with existing findings. It confirms and extends the knowledge of the domain experts on the influence of agricultural water quality in different parts of Ireland. The PCT model can also be used to identify and support recommendations for appropriate management practices on farms to help improve water quality and limit diffuse pollution (due to nutrient losses) in the future.

6. Conclusions

Multi-target regression (MTR) is a structured output prediction task where multiple continuous target variables are predicted simultaneously. In the context of this task, parts of the data can often have missing values for some variables, i.e., the data can be 'incomplete' (unlabeled or partially-labeled). Most machine learning methods for multi-target regression are not able to handle such incomplete data at all, let alone exploit it to the full extent possible.

In this paper, we use the predictive clustering trees (PCTs) approach to address the task of MTR on partially-labeled data. We use PCTs to predict three continuous variables related to water quality in Ireland. Approximately 50% of examples in this dataset are 'incomplete'. We compare the performance of semi-supervised predictive clustering trees, which are able to exploit 'incomplete' data, to the performance of standard supervised predictive clustering trees, which can use only complete (fully-labeled) data. We consider single trees and random forests as well as single-target (i.e., local) and multi-target (i.e., global) regression models. The results of the experimental evaluation in different scenarios clearly show that better predictive performance can be achieved if 'incomplete' data are exploited by PCTs, rather than discarded. Furthermore, we can conclude that ensemble methods (random forests of PCTs), give the best predictive performance.

Moreover, in our study, we also show that the multi-target regression tree can be easily interpreted not only because of the compact size of the model, but also because it provides a joint prediction of multiple targets simultaneously. Namely, one tree predicts all of the targets at the same time. In the singe-target regression case, we have to interpret as many trees as we have targets.

From the domain perspective, this approach has a number of advantages and implications for future water quality mitigation and advice for farmers. It has helped to identify the most important attributes which drive the process of controlling the water quality. These are: chemical and organic (N and P) fertilizer input, the (duration of the) grass growing season, as well as almost all of the pathway attributes i.e., trafficable days, soil moisture deficit, drainage factor and net rainfall. The implications resulting from the *SSLPL-MT-PCT* tree (learned from partially-labeled data) are in accordance with known findings on this topic. Namely, the process of eutrophication is mostly controlled by N and P loss and the risk of potential water pollution is higher if the nutrient input (chemical and organic) is higher. Nutrient input could be caused either by pathway variables or by some management practices with controlled fertilization or manure input. For example, wet poorly drained soil has a higher runoff potential, i.e., carries higher risk for potential water pollution, while the application of chemical and organic fertilizers under these conditions also plays a key role in water pollution.

For further work, the models and maps produced in our study could be further updated by using the most recent water quality data from the national monitoring program. Various additional machine learning methods for predictive modeling (bagging of PCTs, deep neural networks etc.) could also be applied, especially for creating even more accurate maps. Furthermore, it would be interesting to apply existing methods for feature importance estimation (ranking) in order to check whether the features that appear in the tree nodes have high feature importance scores, i.e., are highly ranked.

Acknowledgments

We would like to acknowledge the support of the European Commission through the project LANDMARK - Land management, assessment, research, knowledge base (H2020 Grant number 635201). We also acknowledge the Environment Soils and Land-use Department at TEAGASC Johnstown Castle, Co Wexford Ireland, and Spatial Analysis Unit at TEAGASC, Ashtown, Co Dublin Ireland for providing the data and for constructive comments and suggestions. As their fellow, SN would like to thank the TEAGASC Walsh Fellowship Programme.

References

References

[1] T. Mitchell, *Machine Learning (1 ed.).*, McGraw-Hill, Inc., New York, NY, USA, 1997.

[2] I. H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2005.

[3] S. Džeroski, *Towards a General Framework for Data Mining*, in: Proceedings of the 5th International Conference on Knowledge Discovery in Inductive Databases, KDID'06, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 259–300.

[4] P. Panov, L. N. Soldatova, S. Džeroski, *Generic Ontology of Datatypes*, Inf. Sci. (2016) 900–920doi:10.1016/j.ins.2015.08.006.

[5] G. H. Bakır, T. Hofmann, B. Schölkopf, A. J. Smola, B. Taskar, S. V. N. Vishwanathan, *Predicting Structured Data*, Neural Information Processing. The MIT Press., 2007.

[6] D. Demšar, S. Džeroski, T. Larsen, J. Struyf, J. Axelsen, M. Pedersen, P. Krogh, *Using multi-objective classification to model communities of soil.*, Ecological Modelling 191 (2006) 131–143.

[7] D. Kocev, S. Džeroski, M. White, G. Newell, P. Griffioen, *Using single and multi-target regression trees and ensembles to model a compound index of vegetation condition*, Ecological Modelling 220 (2009) 1159–1168.

[8] J. Levatić, D. Kocev, M. Debeljak, S. Džeroski, *Community structure models are improved by exploiting taxonomic rank with predictive clustering trees*, Ecological Modelling 306 (2015) 294–304.

[9] S. Nikoloski, P. Murphy, D. Kocev, S. Džeroski, D. P. Wall, *Using machine learning to estimate herbage production and nutrient uptake on Irish dairy farms*, Journal of Dairy Science 102 (2019) 10639–10656. doi:https://doi.org/10.3168/jds.2019-16575.

[10] R. Schulte, K. Richards, D. Kurz, E. McDonald, N. Holden, *Agriculture, meteorology and water quality in Ireland: A regional evaluation of pressures and pathways of nutrient loss to water*, Biology and Environment: Proceedings of the Royal Irish Academy 106b (2006) 117–133.

[11] J. Struyf, S. Džeroski, *Constraint based induction of multi-objective regression trees*, Knowledge Discovery in Inductive Databases, LNCS 3933 (2006) 222–333.

[12] D. Kocev, C. Vens, J. Struyf, D. S., *Tree ensembles for predicting structured outputs*, Pattern Recognition 46 (2013) 817–833.

[13] J. Levatić, D. Kocev, M. Ceci, S. Džeroski, *Semi-supervised trees for multi-target regression*, Information Sciences 450 (2018) 109 – 127. doi:https://doi.org/10.1016/j.ins.2018.03.033.

[14] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, H. Blockeel, *Decision trees for hierarchical multi-label classification*, Machine Learning 73 (2008) 185–214.

[15] J. Levatić, D. Kocev, S. Džeroski, *The importance of the label hierarchy in hierarchical multi-label classification*, Journal of Intelligent Information Systems 45 (2015) 247–271.

[16] V. Mileski, S. Džeroski, D. Kocev, *Predictive Clustering Trees for Hierarchical Multi-Target Regression*, in: N. Adams, A. Tucker, D. Weston (Eds.), Advances in Intelligent Data Analysis XVI, Springer International Publishing, 2017, pp. 223–234.

[17] I. Slavkov, V. Gjorgjioski, J. Struyf, S. Džeroski, *Finding explained groups of time-course gene expression profiles with predictive clustering trees*, Molecular BioSystems 6 (2010) 729–740.

[18] M. Debeljak, G. Squire, D. Kocev, C. Hawes, M. Young, S. Džeroski, *Analysis of time series data on agroecosystem vegetation using predictive clustering trees*, Ecological modelling 222 (2011) 2524–2529.

[19] O. Chapelle, B. Schoelkopf, A. Zien, *Semi-supervised Learning*, MIT Press, Cambridge, MA 2.

[20] R. Navaratnam, A. W. Fitzgibbon, R. Cipolla, *The Joint Manifold Model for Semi-supervised Multi-valued Regression*, in: 2007 IEEE 11th International Conference on Computer Vision, 2007, pp. 1–8.

[21] Y. Zhang, D.-Y. Yeung, *Semi-Supervised Multi-Task Regression*, in: W. Buntine, M. Grobelnik, D. Mladenić, J. Shawe-Taylor (Eds.), Machine Learning and Knowledge Discovery in Databases, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 617–631.

[22] H. D. V. Cardona, M. A. Álvarez, Á. A. Orozco, *Convolved Multi-output Gaussian Processes for Semi-Supervised Learning*, in: V. Murino, E. Puppo (Eds.), Image Analysis and Processing — ICIAP 2015, Springer International Publishing, Cham, 2015, pp. 109–118.

[23] U. Brefeld, *Semi-supervised structured prediction models*, Ph.D. thesis, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät II (2008). doi:http://dx.doi.org/10.18452/15748.

[24] Z. Abraham, P. Tan, *A Semi-supervised Framework for Simultaneous Classification and Regression of Zero-Inflated Time Series Data with Application to Precipitation Prediction*, in: 2009 IEEE International Conference on Data Mining Workshops, 2009, pp. 644–649.

[25] M. Herrera, S. Canu, A. Karatzoglou, R. Pérez-García, J. Izquierdo, *An approach to water supply clusters by semisupervised learning*, International Congress on Environmental Modelling and Software. 496.
URL https://scholarsarchive.byu.edu/iemssconference/2010/all/496

[26] S. Carpenter, N. Caraco, D. Correl, R. Howarth, A. Sharpley, S. V.H., *Nonpoint pollution of surface waters with phosphorus and nitrogen.*, Ecological Applications 8 (1998) 559–568.

[27] V. H. Smith, *Eutrophication of Freshwater and Coastal Marine Ecosystems. A Global Problem.*, Environmental Science and Pollution Research 10 (2003) 1–14.

[28] V. H. Smith, S. B. Joye, R. W. Howarth, *Eutrophication of freshwater and marine ecosystems.*, Limnology and Oceanography 51 (2006) 351–355.

[29] D. W. Schindler, *Recent advances in the understanding and management of eutrophication.*, Limnology and Oceanography 51 (2006) 356–363.

[30] W. K. Dodds, *Trophic state, eutrophication and nutrient criteria in streams.*, Trends in Ecology and Evolution 22 (2007) 669–676.

[31] R. Dupas, M. Delmas, J.-M. Dorios, J. Garnier, F. Moatar, C. Gascuel-Odoux, *Assessing the impact of agricultural pressures on N and P loads and eutrophication risk.*, Ecological Indicators 48 (2015) 396–407.

[32] D. Wall, P. Jordan, A. Melland, P.-E. Mellander, C. Buckley, S. Reaney, G. Shortle, *Using the nutrient transfer continuum concept to evaluate the European Union Nitrates Directive National Action Programme*, Journal of Environmental Science and Policy 14 (2011) 664–674.

[33] S. N. Longphuirt, S. O'Boyle, D. B. Stengel, *Environmental response of an Irish estuary to changing land management practices.*, Science of The Total Environment 521-522 (2015) 388–399.

[34] K. Daly, P. Mills, B. Coulter, M. McGarrigle, *Modeling Phosphorus Concentrations in Irish Rivers Using Land Use, Soil Type, and Soil Phosphorus Data.*, Journal of Environmental Quality 31 (2002) 590–599.

[35] S. Giri, Z. Zhang, D. Krasnuk, R. G. Lathrop, *Evaluating the impact of land uses on stream integrity using machine learning algorithms*, Science of The Total Environment 696 (2019) 133858.

[36] K. Chen, H. Chen, C. Zhou, Y. Huang, X. Qi, R. Shen, F. Liu, M. Zuo, X. Zou, J. Wang, Y. Zhang, D. Chen, X. Chen, Y. Deng, H. Ren, *Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data*, Water Research 171 (2020) 115454.

[37] J.-S. Chou, C.-C. Ho, H.-S. Hoang, *Determining quality of water in reservoir using machine learning*, Ecological Informatics 44 (2018) 57 – 75.

[38] H. Blockeel, *Top-down induction of first order logical decision trees*, Ph.D. thesis, Katholieke Universiteit Leuven, Leuven, Belgium (1998).

[39] L. Breiman, J. Friedman, R. Olshen, C. Stone, *Classification and Regression Trees*, Chapman and Hall/CRC (1984).

[40] L. Breiman, *Random forests*, Machine Learning 45 (2001) 5–32.

[41] EPA2016, *Ireland's Environment 2016. Environmental Protection Agency* (2016).

[42] EPA2004, *Ireland's Environment 2004. Environmental Protection Agency* (2004).

[43] M. McGarrigle, J. Bowman, K. Clabby, J. Lucey, P. Cunningham, M. MacCarthaigh, M. Keegan, B. Cantrell, C. Lehane, M. nad Clenaghan, P. Toner, *Water quality in Ireland 1998 - 2000. 2nd ed.*, Wexford. Environmental Protection Agency (2002).

[44] EPA2017, *Water Quality in 2017: an indicators report. Environmental Protection Agency* (2017).

[45] R. Schulte, J. Diamond, K. Finkele, N. Holden, A. Brereton, *Predicting soil moisture conditions of Irish grasslands*, Irish Journal of Agriculture and Food Research 44 (2005) 95–110.

[46] D. Scholefield, D. Lockyer, D. Whitehead, K. Tyson, *A model to predict transformations and losses of nitrogen in UK pastures grazed by beef cattle.*, Plant and Soil 132 (1991) 165–177.

[47] P. Haygarth, S. Jarvis, *Soil derived phosphorus in surface runoff from grazed grassland lysimeters.*, Water Research (1997) 140–148.

[48] R. Uusitalo, E. Turtola, T. Kauppila, T. Lilja, *Particulate phosphorus and sediment in surface runoff and drainflow from clayey soils.*, Journal of Environmental Quality (2001) 589–595.

[49] D. Scholefield, K. Tyson, E. Garwood, A. Armstrong, J. Hawkins, A. Stone, *Nitrate leaching from grazed grassland lysimeters: effects of fertilizer input, field drainage, age of sward and patterns of weather.*, Journal of Soil Science 44 (1993) 601–613.

[50] K. Tyson, D. Scholefield, S. Jarvis, A. Stone, *A comparison of animal output and nitrogen leaching losses recorded from drained fertilized grass and grass/clover pasture.*, Journal of Agricultural Science 129 (1997) 315–323.

[51] R. Earl, *Prediction of trafficability and workability from soil moisture deficit.*, Soil and Tillage Research 40 (1997) 155–168.

Appendix. Supplementary material

A. Description of the pressure-pathway controlling factors

The input (descriptive) space comprises 26 numeric continuous variables, which are described below:

- **Environmental pressures (source) variables:**

  1. *Growth season (days):* Length of the growth season. As long as the grass grows, nutrients are taken up by the grass crop and hence protected from loss. Therefore, shorter grass growth seasons increase N and P pressures [46].
  2. *N c&s / ha farmed (kg/ha):* N input from the excreta of cattle and sheep, expressed as average kg N per hectare farmed within the grid cell.
  3. *N p&p / ha farmed (kg/ha):* N input from the excreta of pig and poultry, expressed as average kg N per hectare farmed within the grid cell.
  4. *N fert / ha farmed (kg/ha):* N fertilizer input, expressed as average kg N per hectare farmed within the grid cell.
  5. *P c&s / ha farmed (kg/ha):* P input from the excreta of cattle and sheep, expressed as average kg P per hectare farmed within the grid cell.
  6. *P p&p / ha farmed (kg/ha):* P input from the excreta of pig and poultry, expressed as average kg P per hectare farmed within the grid cell.
  7. *P fert / ha farmed (kg/ha):* P fertilizer input, expressed as average kg P per hectare farmed within the grid cell.
  8. *Sum N-input / hectare farmed (kg/ha)* = N c&s + N p&p + N fert / ha farmed
  9. *Sum P-input / hectare farmed (kg/ha)* = P c&s + P p&p + P fert / ha farmed
  10. *N c&s / ha total (kg/ha):* N input from the excreta of cattle and sheep, expressed as average kg N per hectare farmed and non-farmed within the grid cell.
  11. *N p&p / ha total (kg/ha):* N input from the excreta of pig and poultry, expressed as average kg N per hectare farmed and non-farmed within the grid cell.

12. *N fert / ha total (kg/ha):* N fertilizer input, expressed as average kg N per hectare farmed and unfarmed within the grid cell;
13. *P c&s / ha total (kg/ha):* P input from the excreta of cattle and sheep, expressed as average kg P per hectare farmed and unfarmed within the grid cell.
14. *P p&p / ha total (kg/ha):* P input from the excreta of pig and poultry, expressed as average kg P per hectare farmed and unfarmed within the grid cell.
15. *P fert / ha total (kg/ha):* P fertilizer input, expressed as average kg P per hectare farmed and unfarmed within the grid cell.
16. *Sum N-input / hectare total (kg/ha)* = N c&s + N p&p + N fert / ha total
17. *Sum P-input / hectare total (kg/ha)* = P c&s + P p&p + P fert / ha total

- **Pathway (transport) variables:**

  1. *Drainage factor:* Describes the average infiltration capacity of the soil, ranging from 0 (entire grid cell is poorly-drained) to 1 (entire grid cell is well-drained). Poorly drained soils (low drainage factor) are prone to overland flow and hence prone to P-loss. Well-drained soil (high drainage factor) are prone to leaching of N [47, 48].
  2. *SMDmax:* Soil moisture deficit is calculated from the hybrid SMD model given by Schulte et al. [45], as a cumulative balance from precipitation, evapotranspiration and drainage. High SMDmax, i.e., maximum soil moisture deficit, during summer is related to N-loss in the subsequent winter: the higher the SMD in summer, the higher the N concentrations in autumn/winter [45].
  3. *Net rainfall*, is calculated as rainfall minus evapotranspiration. Net rainfall is relatively high everywhere to ensure full recharge in winter, and to transport residual nitrates in soils below the rooting zone. Therefore the main effect of net rainfall is to dilute nitrate concentrations. For P, it may lead to higher losses of P through overland flow [49, 50].

4. *Hi drain q1, q2, q3, q4, sum:* The number of intense drainage events (>15mm per day) in the 1st, 2nd, 3rd and 4th quarter of the year, or (sum) the total number of intense drainage events per year. Intense drainage events are the main pathway of P-loss (90% of P is lost in 4-5 drainage events) [10]
5. *Trafficable days:* The number of days on which the soil conditions allow traffic by animal and/or machinery. Trafficking of soils when soil moisture deficits is lower than 10mm greatly increase the risk of nutrient-loss[51].

Note that the N and P in animal excreta (c&s and p&p) are both derived from the same animal numbers and therefore show almost perfect correlation. It will be difficult to discriminate between these two descriptive variables.

B. Supervised multi-target regression PCT model: SL-MT-PCT

In Figure 10, we show the *SL-MT-PCT* tree learned from the fully-labeled data with supervised PCTs for MTR. The *SL-MT-PCT* model is obtained by learning from 50% of the available data, therefore, the tree is smaller. The predictive performance of the PCT model is worse than the *SSLPL-MT-PCT* model learned from the partially-labeled data.

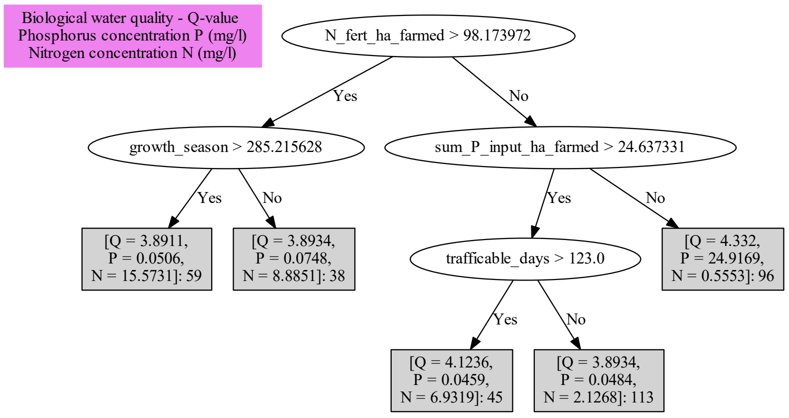

Biological water quality - Q-value
Phosphorus concentration P (mg/l)
Nitrogen concentration N (mg/l)
N_fert_ha_farmed > 98.173972
Yes
No
growth_season > 285.215628
sum_P_input_ha_farmed > 24.637331
Yes
No
Yes
No
[Q = 3.8911, P = 0.0506, N = 15.5731]: 59
[Q = 3.8934, P = 0.0748, N = 8.8851]: 38
trafficable_days > 123.0
[Q = 4.332, P = 24.9169, N = 0.5553]: 96
Yes
No
[Q = 4.1236, P = 0.0459, N = 6.9319]: 45
[Q = 3.8934, P = 0.0484, N = 2.1268]: 113

Figure 10: Multi-target regression tree learned from the fully-labeled examples in the *SL-MT-PCT* scenario.

Chapter 7

Conclusions

In this thesis, we have presented a method for structured output prediction that finds the dependencies in the output space and structures the target attributes into a hierarchy. We have also applied machine learning methods for structured output prediction (SOP) in two case studies in the domain of dairy and soil science. We have improved upon existing methods for SOP by identifying hierarchical target space representations. In the case studies, we have exploited the new human-centered paradigm called explainable artificial intelligence (XAI), by focusing on the understandability and interpretability of the learned predictive models that will facilitate their use by decision makers.

From a machine learning point of view, we have proposed a novel method for data-driven structuring of the output space in multi-target prediction (MTP). For the structuring process, we have used two different representations of the output space, one consisting of the values of the target attributes themselves and another consisting of the feature importance scores for each target attribute. We address the limitations of the existing methods for multi-target prediction that do not take into account similarities among the target attributes. Our proposed algorithm transforms the classical MTP tasks to hierarchical MTP tasks by using the structure (i.e., hierarchy) of the target attributes reconstructed in a data-driven manner. Our results show improvement in the predictive performance when this structure is used on a target space, especially for problems with a large number of target attributes.

From the perspective of environmental sciences and modeling of soil ecosystem services, we have introduced two novel case studies. We have applied supervised and semi-supervised predictive clustering trees (PCTs) for multi-target regression on data related to modeling specific aspects and outcomes for three soil functions. These functions are (1) primary productivity (as referred by total grass yield and nutrient uptake) on Irish dairy farms; and a combination of (2) water purification and regulation and (3) regulation and provision of nutrients (biological water quality and nitrogen/ phosphorus concentration) in Irish agricultural catchments.

In our first case study, concerned with estimating total grass yield and nutrient uptake, we have presented the advantages of using PCTs for MTR as a prominent and well-known representative of explainable AI. The trees learned from the data provided useful information in terms of explaining herbage production potential and nutrient uptake across a large number of sites from 16 Irish dairy farms. The training data consisted of many soil, environmental and management data variables and four target variables (total herbage production, N, P and K uptake), all measured at each sample point, i.e. data example. Our results show that the performance of the single MTR tree is essentially the same as that of the four models for predicting each target separately, while the single MTR model can be more easily interpreted by the domain experts. We have also improved the predictive performance of individual PCT models by using ensembles of PCTs, i.e., random forests of PCTs, but at the cost of losing interpretability. Furthermore, the obtained tree models have shown significant practical implications for the domain of use, such as: (1) guiding towards a more balanced fertilizer input; (2) identifying poorer herbage potential in specific fields/regions of the country; and (3) monitoring and assessing grassland productivity with only a few most important input variables which appear in the tree model.

In our second case study, we have addressed the limitations of existing MTR methods related to data availability. Namely, the data for this study is incomplete, i.e., not all of the target variables have been measured/assessed during data collection. Supervised machine learning methods, which cannot handle incomplete data, discard this data, and learn models from the small amount of complete (i.e, fully-labeled) data. In order to address this limitation, we have proposed the use of semi-supervised PCTs for MTR, with the incomplete (i.e., partially-labeled and unlabeled) data, where not all of the target attributes have been measured for each data example. Our results have shown improvement in the predictive performance of the learned PCTs. The obtained PCT models were also understandable/explainable to the domain experts, confirmed the existing domain knowledge and suggested new recommendations with practical implications for the domain. We achieved even larger improvements in predictive performance, when learning ensembles of semi-supervised PCTs for MTR.

7.1 Summary of Contributions

In this dissertation, we have made the following contributions to science:

- In the context of multi-target prediction (MTP), we have introduced novel methods for data-driven exploitation of the structure of the output space which discover a target/label hierarchy during the learning process. We transform the problem of MTP to the problem of hierarchical MTP. In particular, we transform MLC tasks to HMLC tasks (Chapter 5.1) and MTR tasks to HMTR tasks (Chapter 5.2).

  We evaluate our approach on various benchmark datasets of both MLC and MTR tasks and in both cases show improvements as compared to the classical MTP approaches, where no target relations and dependencies are considered. The improvement is noticeable for tasks with large output spaces (more than 100 target attributes). We thus confirm both parts of Hypothesis 1, with MLC models and experiments referring to Hypothesis 1a and MTR methods and experiments referring to Hypothesis 1b.

- We have used ML methods for MTR on a case study of modeling primary productivity soil function, specifically, for estimating total herbage production and nutrient uptake in Irish dairy farms. We have used the existing soil, environment and management data to learn models interpretable by human experts and contribute to the domain of dairy science (Chapter 6.1). We have used PCTs for multi-target regression which predict multiple targets simultaneously, as well as single target PCTs. Our results show that PCTs for MTR are better in terms of both predictive power and interpretability: Instead of interpreting and understanding of three (one for each target variable) different single-target (STR) trees, the domain expert has to interpret and explain only one MTR tree. They confirm Hypothesis 2.

- We have applied semi-supervised learning to model several aspects and outcomes of two soil functions, i.e., water regulation and purification, and regulation and provision of nutrients. Specifically, we have modeled water quality in agricultural catchments in Ireland by using existing partially-labeled data from the Irish national monitoring program, moving beyond classical supervised learning methods that learn from complete, i.e., fully-labeled data (Chapter 6.2). We have proposed the use of a variant of semi-supervised learning that uses incomplete, i.e., partially-labeled data, where not all of the target attributes are measured at every observation point; all, none, or several of the targets can be measured for each data point. On the considered case study, we have obtained better predictive performance and maintained interpretability of the learned models from partially-labeled data as compared to learning from fully-labeled data only. We have thus completely confirmed Hypothesis 3.

7.2 Further Work

An immediate possibility for the extension of the proposed methodology for structuring the output space would consider the use of different feature ranking methods for MLC/MTR for creating the space of feature importance scores used to describe the targets. These descriptions would be then used to hierarchically structure the target space. There are several well-known feature ranking methods that could be used for this purpose, such as: the ReliefF method (Kononenko et al., 1997; Robnik-Sikonja & Kononenko, 1997) and feature ranking by using generalized genetic programming with symbolic regression for high-dimensional spaces (Q. Chen et al., 2017). Moreover, we can consider different representations of the output space such as reduced dimensionality versions of the target space produced by existing algorithms for dimensionality reduction, such as matrix factorization with principal component analysis (PCA)(Jolliffe, 2002), non-negative matrix factorization (NMF) (Sra & Dhillon, 2006; Tandon & Sra, 2010) and singular value decomposition (SVD) (Stewart, 1993) as well as their evaluation/ application to MTP problems with large output spaces.

In this dissertation, we have presented two case studies, one related to modeling the primary productivity soil function, i.e., estimation of total herbage potential and nutrient uptake, and another related to modeling water quality in agricultural catchments in the Republic of Ireland, covering some aspects of two different soil functions: water purification and regulation, and regulation and provision of nutrients. The data provider in both case studies was TEAGASC, Environment Soils and Land-use Department from Ireland. For further work, there is a possibility to apply similar machine learning approaches to different datasets for the same soil functions, as well as to data for the remaining three soil functions: soil biodiversity, nutrient cycling and climate regulation, and carbon sequestration. Moreover, we can apply our proposed modeling approach, either on each soil function separately or for two and more soil functions simultaneously, once the data are available. In this way, we can consider the interactions among the different soil functions.

In addition to the aforementioned possibilities for further work, we could extend the existing decision support tool for modeling soil functions, called the Soil Navigator (Debeljak et al., 2019). This tool simultaneously assesses several soil functions as was developed within the LANDMARK EU H2020 project (LANDMARK, 2019). In the data-driven extension of the Soil navigator, we would replace the decision rules from human-made DEXi models for each soil function by predictive rules derived from the interpretable models learned from data by PCTs in the appropriate decision tables.

Finally, we can envisage the application of the methods and approaches proposed in this thesis to other domains, not only the domain of soil sciences. For example, other problems from the area of sustainable food production and environmental sciences can be considered. In this context, we expect the ability of learning explainable models in the spirit of explainable AI.

References

Aarts, H., de Haan, M., Schroder, J., Holster, H., de Boer, J., Reijs, J., Oenema, J., Hilhorst, G., Sebek, L., Verhoeven, F., & Meerkerk, B. (2015). Quantifying the environmental performance of individual dairy farms - the annual nutrient cycling assessment (anca). In *Grassland and forages in high output dairy farming systems* (pp. 377–380). Wageningen Academic Publishers.

Abraham, Z., & Tan, P. (2009). A semi-supervised framework for simultaneous classification and regression of zero-inflated time series data with application to precipitation prediction, In *2009 IEEE International Conference on Data Mining Workshops.*

Abrahamsen, P., & Hansen, S. (2000). Daisy: An open soil-crop-atmosphere system model. *Environmental Modelling and Software*, *15*(3), 313–330.

Agrawal, R., Gupta, A., & Sarawagi, S. (1997). Modeling multidimensional databases. *In Proceedings of the 13th International Conference on Data Engineering (IEEE Computer Society)*, 232–243.

Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., Ehteram, M., & Elshafie, A. (2019). Machine learning methods for better water quality prediction. *Journal of Hydrology*, *578*, 124084.

Aho, T., Ženko, B., & Džeroski, S. (2009). Rule ensembles for multi-target regression, In *In Proc. of Ninth IEEE International Conference on Data Mining*, IEEE Press.

Akansu, A. N., & Haddad, R. A. (1992). *Multiresolution signal decomposition: Transforms, subbands, wavelets.* San Diego: Academic Press.

Alali, A., & Kubat, M. (2015). Prudent: A pruned and confident stacking approach for multi-label classification. *IEEE Transactions on Knowledge and Data Engineering*, *27*(9), 2480–2493.

Alaydie, N., Reddy, C. K., & Fotouhi, F. (2012). Exploiting label dependency for hierarchical multi-label classification., In *Advances in Knowledge Discovery and Data Mining. PAKDD 2012. Lecture Notes in Computer Science, vol 7301.* Springer, Berlin, Heidelberg.

Ali, I., Cawkwell, F., Dwyer, E., & Green, S. (2016). Modeling managed grassland biomass estimation by using multitemporal remote sensing data—a machine learning approach. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *10*(7), 3254–3264.

Allison, L. (2003). Types and classes of machine learning and data mining, In *26th Australasian Computer Science Conference (ACSC), Adelaide, ACS Series Conferences in Research and Practice in Information Technology V16.*

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175–185.

Appice, A., & Džeroski, S. (2007). Stepwise induction of multi-target model trees, In *Proceedings of 18th ECML 2007, Warsaw, Poland.*

Aron, A., & Aron, E. (1999). *Statistics for psychology.* Prentice Hall, Upper Saddle River, NJ.

Arumugam, A. (2017). A predictive modeling approach for improving paddy crop productivity using data mining techniques. *Turkish Journal of Electrical Engineering and Computer Sciences*, *25*, 4777–4787.

Bakır, G. H., Hofmann, T., Schölkopf, B., Smola, A. J., Taskar, B., & Vishwanathan, S. V. N. (2007). *Predicting structured data*. Neural Information Processing. The MIT Press.

Balkovič, J., van der Velde, M., Schmid, E., Skalský, R., Khabarov, N., Obersteiner, M., Stürmer, B., & Xiong, W. (2013). Pan-european crop modelling with epic: Implementation, up-scaling and regional crop yield validation. *Agricultural Systems*, *120*, 61–75.

Bampa, F., O'Sullivan, L., Madena, K., Sandén, T., Spiegel, H., Henriksen, C. B., Ghaley, B. B., Jones, A., Staes, J., Sturel, S., Trajanov, A., Creamer, R. E., & Debeljak, M. (2019). Harvesting european knowledge on soil functions and land management using multi-criteria decision analysis. *Soil Use and Management*, *35*(1), 6–20.

Barros, R. C., Cerri, R., Freitas, A. A., & de Carvalho, A. C. P. L. F. (2012). Probabilistic clustering for hierarchical multi-label classification of protein functions., In *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2013. Lecture Notes in Computer Science, vol 8189*, Springer, Berlin, Heidelberg.

Barutcuoglu, Z., Schapire, R. E., & Troyanskaya, O. G. (2006). Hierarchical multi-label prediction of gene function. *Bioinformatics*, *22*(7), 830–836.

Blockeel, H., Raedt, L. D., & Ramon, J. (1998). Top-down induction of clustering trees, In *Proc. of the 15th International Conference on Machine Learning*.

Blockeel, H., Schietgat, L., Struyf, J., Džeroski, S., & Clare, A. (2006). Decision trees for hierarchical multilabel classification: A case study in functional genomics., In *Knowledge Discovery in Databases: PKDD 2006. PKDD 2006. Lecture Notes in Computer Science, vol 4213*. Springer, Berlin, Heidelberg.

Bockstaller, C., & Girardin, P. (2003). How to validate environmental indicators. *Agricultural Systems*, *76*(2), 639–653.

Bohanec, M. (2014). *Dexi: Program for multi-attribute decision making, user's manual, version 4.01*. IJS Report DP-11739 (Ljubljana: Jožef Stefan Institute).

Bohanec, M. (2017). Multi-criteria dex models: An overview and analysis.

Bohanec, M., Boshkoska, B. M., Prins, T. W., & Kok, E. J. (2017). Sigmo: A decision support system for identification of genetically modified food or feed products. *Food Control*, *71*, 168–177.

Bohanec, M., & Rajkovič, V. (1990). Dex: An expert system shell for decision support. *Sistemica*, 145–157.

Bondi, G., Creamer, R., Ferrari, A., Fenton, O., & Wall, D. (2018). Using machine learning to predict soil bulk density on the basis of visual parameters: Tools for in-field and post-field evaluation. *Geoderma*, *318*, 137–147.

Borchani, H., Varando, G., Bielza, C., & Larrañaga, P. (2015). A survey on multi-output regression. *WIREs Data Mining and Knowledge Discovery*, *5*(5), 216–233.

Boutell, M. R., Luo, J., Shen, X., & Brown, C. M. (2004). Learning multi-label scene classification. *Pattern Recognition*, *37*(9), 1757–1771.

Brefeld, U. (2008). *Semi-supervised structured prediction models* (Doctoral dissertation). Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät II.

Breiman, L., & Friedman, J. H. (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *59*(1), 3–54.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(1), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. J. (1984). *Classification and regression trees*. Chapman & Hall/CRC.

Breskvar, M., Kocev, D., & Džeroski, S. (2018). Ensembles for multi-target regression with random output selections. *Machine Learning*, *107*, 1673–1709.

Brisson, N., Gary, C., Justes, E., Roche, R., Mary, B., Ripoche, D., Zimmer, D., Sierra, J., Bertuzzi, P., Burger, P., Bussière, F., Cabidoche, Y., Cellier, P., Debaeke, P., Gaudillère, J., Hénault, C., Maraux, F., Seguin, B., & Sinoquet, H. (2003). An overview of the crop model STICS. *European Journal of Agronomy*, *18*(3), 309–332.

Brouard, C., Szafranski, M., & d'Alché-Buc, F. (2016). Input output kernel regression: Supervised and semi-supervised structured output prediction with operator-valued kernels. *Journal of Machine Learning Research*, *17*(176), 1–48.

Brouwer, W. J., Kubicki, J. D., Sofo, J. O., & Gilesd, C. L. (2014). An investigation of machine learning methods applied to structure prediction in condensed matter. *arXiv preprint arXiv:1405.3564*.

Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H., & Kazakis, N. (2020). Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Science of The Total Environment*, *721*, 137612.

Bui, E., Henderson, B., & Viergever, K. (2009). Using knowledge discovery with data mining from the australian soil resource information system database to inform soil carbon mapping in australia. *Global Biogeochemical Cycles*, *23*(4).

Cardona, H. D. V., Álvarez, M. A., & Orozco, Á. A. (2015). Convolved multi-output gaussian processes for semi-supervised learning (V. Murino & E. Puppo, Eds.). In V. Murino & E. Puppo (Eds.), *Image Analysis and Processing – ICIAP 2015*, Cham, Springer International Publishing.

Castrillo, M., & García, Á. L. (2020). Estimation of high frequency nutrient concentrations from water quality surrogates using machine learning methods. *Water Research*, *172*, 115490.

Cerri, R., Barros, R. C., & [de Carvalho], A. C. P. L. F. (2012). A genetic algorithm for hierarchical multi-label classification, In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, Trento, Italy, Association for Computing Machinery.

Cerri, R., Barros, R. C., & Carvalho], A. C. [(2014). Hierarchical multi-label classification using local neural networks. *Journal of Computer and System Sciences*, *80*(1), 39–56.

Cesa-Bianchi, N., Gentile, C., & Zaniboni, L. (2006). Incremental algorithms for hierarchical classification. *The Journal of Machine Learning Research*, *7*, 31–54.

Chapelle, O., Schoelkopf, B., & Zien, A. (2006). Semi-supervised learning. *MIT Press, Cambridge, MA*, *2*.

Chen, B., Li, W., Zhang, Y., & Hu, J. (2016). Enhancing multi-label classification based on local label constraints and classifier chains, In *2016 International Joint Conference on Neural Networks (IJCNN)*, IEEE.

Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., Liu, F., Zuo, M., Zou, X., Wang, J., Zhang, Y., Chen, D., Chen, X., Deng, Y., & Ren, H. (2020). Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Research*, *171*, 115454.

Chen, Q., Zhang, M., & Xue, B. (2017). Feature selection to improve generalization of genetic programming for high-dimensional symbolic regression. *IEEE Transactions on Evolutionary Computation*, *21*(5), 792–806.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *CoRR*, arXiv.

Chen, W.-J., Shao, Y.-H., Li, C.-N., & Deng, N.-Y. (2016). Mltsvm: A novel twin support vector machine to multi-label learning. *Pattern Recognition*, *52*, 61–74.

Cheng, W., & Hüllermeier, E. (2009). Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning*, *76*, 211–225.

Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, *151*, 61–69.

Chou, J.-S., Ho, C.-C., & Hoang, H.-S. (2018). Determining quality of water in reservoir using machine learning. *Ecological Informatics*, *44*, 57–75.

Clare, A., & King, R. D. (2003). Predicting gene function in saccharomyces cerevisiae. *Bioinformatics*, *19*, ii42–ii49.

Clare, A. (2003). *Machine learning and data mining for yeast functional genomics.* (Doctoral dissertation). University of Wales, Aberystwyth. CA.

Clare, A., & King, R. D. (2001). Knowledge discovery in multi-label phenotype data., In *De Raedt L., Siebes A. (eds) Principles of Data Mining and Knowledge Discovery. PKDD 2001. Lecture Notes in Computer Science, vol 2168.*

Cluzeau, D., Guernion, M., Chaussod, R., Martin-Laurent, F., Villenave, C., Cortet, J., Ruiz-Camacho, N., Pernin, C., Mateille, T., Philippot, L., Bellido, A., Rougé, L., Arrouays, D., Bispo, A., & Pérès, G. (2012). Integration of biodiversity in soil quality monitoring: Baselines for microbial and soil fauna parameters for different land-use types. *European Journal of Soil Biology*, *49*, 63–72.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27.

Craheix, D., Angevin, F., Doré, T., & de Tourdonnet, S. (2016). Using a multicriteria assessment model to evaluate the sustainability of conservation agriculture at the cropping system level in france. *European Journal of Agronomy*, *76*, 75–86.

Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters*, *13*(11), 114003.

Daly, K., Richards, K., Mellander, P.-E., Jordan, P., hUallacháin, D. Ó., Sheriff, S., Vero, S. E., & Fenton, O. (2018). Soils and water quality. In R. Creamer & L. O'Sullivan (Eds.), *The soils of ireland. world soils book series* (pp. 235–243). Cham, Springer.

de Leeuw, J., & Meijer, E. (2008). *Handbook of multilevel analysis.* Springer, New York, NY.

Debeljak, M., Cortet, J., Demšar, D., Krogh, P. H., & Džeroski, S. (2007). Hierarchical classification of environmental factors and agricultural practices affecting soil fauna under cropping systems using bt maize. *Pedobiologia*, *51*(3), 229–238.

Debeljak, M., Squire, G. R., Kocev, D., Hawes, C., Young, M. W., & Džeroski, S. (2011). Analysis of time series data on agroecosystem vegetation using predictive clustering trees. *Ecological Modelling*, *222*(14), 2524–2529.

Debeljak, M., Trajanov, A., Kuzmanovski, V., Schröder, J., Sandén, T., Spiegel, H., Wall, D. P., Van de Broek, M., Rutgers, M., Bampa, F., Creamer, R. E., & Henriksen, C. B. (2019). A field-scale decision support system for assessment and management of soil functions. *Frontiers in Environmental Science*, *7*, 115.

DECATHLON. (2016). Development of Cost efficient Advanced DNA-based methods for specific Traceability issues and High Level On-site applicatioNs [Online; accessed 25.03.2020].

Demšar, D., Džeroski, S., Larsen, T., Struyf, J., Axelsen, J., Pedersen, M., & Krogh, P. (2006). Using multi-objective classification to model communities of soil. *Ecological Modelling*, *191*, 131–143.

Deutsch, C., & Journel, A. (1998). *GSLIB: Geostatistical Software Library and User's Guide. 2nd Edition.* Oxford University Press, New York.

Djerdj, T., Hackenberger, D. K., Hackenberger, D. K., & Hackenberger, B. K. (2020). Observing earthworm behavior using deep learning. *Geoderma*, *358*, 113977.

Djodjic, F., Montas, H., Shirmohammadi, A., Bergström, L., & Ulén, B. (2002). A decision support system for phosphorus management at a watershed scale. *Journal of Environmental Quality*, *31*(3), 937–945.

Dong, Y., Fu, Z., Peng, Y., Zheng, Y., Yan, H., & Li, X. (2020). Precision fertilization method of field crops based on the wavelet-bp neural network in china. *Journal of Cleaner Production*, *246*, 118735.

Dou, X., & Yang, Y. (2018). Evapotranspiration estimation using four different machine learning approaches in different terrestrial ecosystems. *Computers and Electronics in Agriculture*, *148*, 95–106.

Duivesteijn, W., Mencía, E. L., Fürnkranz, J., & Knobbe, A. (2012). Multi-label lego — enhancing multi-label classifiers with local patterns., In *Advances in Intelligent Data Analysis XI. IDA 2012. Lecture Notes in Computer Science, vol 7619.*

Džeroski, S. (2006). Towards a general framework for data mining, In *Proceedings of the 5th International Conference on Knowledge Discovery in Inductive Databases*, Berlin, Germany, Springer-Verlag.

Edwards, L., & Veale, M. (2017). Slave to the algorithm? why a 'right to an explanation' is probably not the remedy you are looking for. *Duke Law and Technology Review*, *16*(1), 1–65.

Efroymson, M. A. (1960). *Multiple regression analysis. mathematical methods for digital computers, ralston a. and wilf,h. s., (eds.)* Wiley, New York.

Elisseeff, A., & Weston, J. (2001). A kernel method for multi-labelled classification, In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, Vancouver, British Columbia, Canada, MIT Press.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813.

Enrique Sucar, L., Bielza, C., Morales, E. F., Hernandez-Leal, P., Zaragoza, J. H., & Larrañaga, P. (2014). Multi-label classification with bayesian network-based chain classifiers. *Pattern Recognition Letters*, *41*, 14–22.

FAO. (2003). *Food and Agriculture Organization of the United Nations (FAO), World agriculture: towards 2015/2030. An FAO Perspective.* J. Bruinsma (Ed.), London: Earthscan Publications Ltd.

Faris, H., Aljarah, I., & Mirjalili, S. (2017). Chapter 28 - Evolving Radial Basis Function Networks Using Moth–Flame Optimizer. In P. Samui, S. Sekhar, & V. E. Balas (Eds.), *Handbook of neural computation* (pp. 537–550). Academic Press.

Fjodorova, N., Vračko, M., Jezierska, A., & Novič, M. (2010). Counter propagation artificial neural network categorical models for prediction of carcinogenicity for non-congeneric chemicals. *SAR and QSAR in Environmental Research*, *21*(1-2), 57–75.

Friedman, J. (2002). Stochastic gradient boosting. *Computational Statistical Data Analysis*, *38*(4), 367–378.

Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics*, *11*, 86–92.

Fürnkranz, J., Hüllermeier, E., Mencía, E. L., & Brinker, K. (2008). Multilabel classification via calibrated label ranking. *Machine learning*, *73*, 133–153.

Gardi, C., Montanarella, L., Arrouays, D., Bispo, A., Lemanceau, P., Jolivet, C., Mulder, C., Ranjard, L., Römbke, J., Rutgers, M., & Menta, C. (2009). Soil biodiversity monitoring in europe: Ongoing activities and challenges. *European Journal of Soil Science*, *60*(5), 807–819.

Gardi, C., Visioli, G., Conti, F. D., Scotti, M., Menta, C., & Bodini, A. (2016). High nature value farmland: Assessment of soil organic carbon in europe. *Frontiers in Environmental Science*, *4*, 47.

Gelman, A. (2006). Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics*, *48*, 432–435.

Gilhespy, S. L., Anthony, S., Cardenas, L., Chadwick, D., del Prado, A., Li, C., Misselbrook, T., Rees, R. M., Salas, W., Sanz-Cobena, A., Smith, P., Tilston, E. L., Topp, C. F., Vetter, S., & Yeluripati, J. B. (2014). First 20 years of DNDC (DeNitrification DeComposition): Model evolution. *Ecological Modelling*, *292*, 51–62.

Giri, S., Zhang, Z., Krasnuk, D., & Lathrop, R. G. (2019). Evaluating the impact of land uses on stream integrity using machine learning algorithms. *Science of The Total Environment*, *696*, 133858.

Giusti, E., & Marsili-Libelli, S. (2015). A fuzzy decision support system for irrigation and water conservation in agriculture. *Environmental Modelling and Software*, *63*, 73–86.

Gjorgjevic, D., Madjarov, G., & Džeroski, S. (2013). Hybrid decision tree architecture utilizing local svms for efficient multi-label learning. *International Journal of Pattern Recognition and Artificial Intelligence*, *27*, 1351004.

Godbole, S., & Sarawagi, S. (2004). Discriminative methods for multi-labeled classification. In *Advances in knowledge discovery and data mining* (pp. 22–30). Springer Berlin, Heidelberg.

Goldberg, D., & Holland, J. (1988). Genetic algorithms and machine learning. *Machine Learning*, *3*, 95–99.

Goldberger, A. S. (1964). *Classical linear regression. econometric theory.* John Wiley; Sons, New York.

Gonçalves, E. C., Plastino, A., & Freitas, A. A. (2013). A genetic algorithm for optimizing the label ordering in multi-label classifier chains, In *2013 IEEE 25th International Conference on Tools with Artificial Intelligence.*

Gonçalves, T., & Quaresma, P. (2004). Using ir techniques to improve automated text classification., In *Meziane F., Métais E. (eds) Natural Language Processing and Information Systems. NLDB 2004. Lecture Notes in Computer Science, vol 3136*, Springer, Berlin, Heidelberg.

Gönen, M., & Kaski, S. (2014). Kernelized bayesian matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(10), 2047–2060.

Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., & Schmidhuber, J. (2009). A novel connectionist system for improved unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*, 855–868.

Griffiths, B., Rombke, J., Schmelz, R., Scheffczyk, A., Faber, J., Bloem, J., Pérès, G., Cluzeau, D., Chabbi, A., Suhadolc, M., Sousa, J., da Silva, P. M., Carvalho, F., Mendes, S., Morais, P., Francisco, R., Pereira, C., Bonkowski, M., Geisen, S., . . . Stone, D. (2016). Selecting cost effective and policy-relevant biological indicators for european monitoring of soil biodiversity and ecosystem function. *Ecological Indicators*, *69*, 213–223.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI explainable artificial intelligence. *Science Robotics*, *4*(37).

Guo, X., Bian, Z., Wang, S., Wang, Q., Zhang, Y., Zhou, J., & Lin, L. (2020). Prediction of the spatial distribution of soil arthropods using a random forest model: A case study in changtu county, northeast china. *Agriculture, Ecosystems and Environment*, *292*, 106818.

Hamouda, M. A., Anderson, W. B., & Huck, P. M. (2009). Decision support systems in water and wastewater treatment process selection and design: a review. *Water Science and Technology*, *60*(7), 1757–1770.

Hastie, T., Friedman, J., & Tibshirani, R. (2001). Additive models, trees, and related methods. *In The Elements of Statistical Learning (Springer)*, 321–329.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction.* Springer, NY.

Herrera, M., Canu, S., Karatzoglou, A., Pérez-García, R., & Izquierdo, J. (2010). An approach to water supply clusters by semisupervised learning. *International Congress on Environmental Modelling and Software*, *496*.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. *The Annals of Statistics*, *36*, 1171–1220.

Hopfield, J., & Tank, D. (1985). 'neural' computation of decisions in optimization problems. *Biological Cybernetics*, *52*, 141–152.

Hosseinzadeh, A., Baziar, M., Alidadi, H., Zhou, J. L., Altaee, A., Najafpoor, A. A., & Jafarpour, S. (2020). Application of artificial neural network and multiple linear regression in modeling nutrient recovery in vermicompost under different conditions. *Bioresource Technology*, *303*, 122926.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, *15*, 651–674.

Huang, G.-B., Zhou, H., Ding, X., & Zhang, R. (2012). Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, *42*(2), 513–529.

Huynh-Thu, V. A., Irrthum, L., Wehenkel, & Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLos One*, *5*(9), 1–10.

Iman, R. L., & Davenport, J. M. (1980). Approximations of the critical region of the friedman statistic. *Communications in Statistics - Theory and Methods*, *9*, 571–595.

Izenman, A. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, *5*, 248–264.

Jang, J. .-.-R. (1993). ANFIS: Adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man, and Cybernetics*, *23*(3), 665–685.

Jayadeva, Khemchandani, R., & Chandra, S. (2007). Twin support vector machines for pattern classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(5), 905–910.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features., In *Nédellec C., Rouveirol C. (eds) Machine Learning: ECML-98. ECML 1998. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol 1398.* Springer, Berlin, Heidelberg.

Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *31*(3), 300–303.

Jolliffe, I. T. (2002). *Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed.* Springer, NY.

Joly, A. (2017). Exploiting random projections and sparsity with random forests and gradient boosting methods—application to multi-label and multi-output learning, random forest model compression and leveraging input sparsity. *arXiv preprint arXiv:1704.08067.*

Joly, A., Geurts, P., & Wehenkel, L. (2014). Random forests with random projections of the output space for high dimensional multi-label classification. *In Joint European conference on machine learning and knowledge discovery in databases*, 607–622.

Jónsson, J. Ö. G., & Davídsdóttir, B. (2016). Classification and valuation of soil ecosystem services. *Agricultural Systems*, *145*, 24–38.

Kiritchenko, S., Matwin, S., Nock, R., & Famili, A. F. (2006). Learning and evaluation in the presence of class hierarchies: Application to text categorization., In *Advances in Artificial Intelligence. Canadian AI 2006. Lecture Notes in Computer Science, vol 4013.* Springer, Berlin, Heidelberg.

Kocev, D., Džeroski, S., White, M. D., Newell, G. R., & Griffioen, P. (2009). Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition. *Ecological Modelling*, *220*(8), 1159–1168.

Kocev, D., Naumoski, A., Mitreski, K., Krstić, S., & Džeroski, S. (2010). Learning habitat models for the diatom community in lake prespa. *Ecological Modelling*, *221*(2), 330–337.

Kocev, D., Vens, C., Struyf, J., & Džeroski, S. (2013). Tree ensembles for predicting structured outputs. *Pattern Recognition*, *46*(3), 817–833.

Kohavi, R. (1995). *Wrappers for performance enhancement and oblivious decision graphs.* (Doctoral dissertation). Department of Computer Science, Stanford University. CA.

Koller, D., & Sahami, M. (1997). Hierarchically classifying documents using very few words, In *Proceedings of the Fourteenth International Conference on Machine Learning*, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc.

Kononenko, I., Simec, E., & Sikonja, M. R. (1997). Overcoming the myopia of inductive learning algorithms with relieff. *Applied Intelligence*, *7*, 39–55.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Commun. ACM*, *60*, 84–90.

Kuhn, M., Weston, S., Keefer, C., Coulter, N., & Quinlan, R. (2014). Cubist: Rule-and instance-based regression modeling [R package version 0.0.18;CRAN: Vienna, Austria].

Kung, H.-Y., Kuo, T.-H., Chen, C.-H., & Tsai, P.-Y. (2016). Accuracy analysis mechanism for agriculture data using the ensemble neural network method. *Sustainability*, *8*(8).

Kuo, M., Mohler, B., Raudenbush, S. L., & Earls, F. J. (2000). Assessing exposure to violence using multiple informants: Application of hierarchical linear model. *Journal of Child Psychology and Psychiatry*, *41*(8), 1049–1056.

LANDMARK. (2019). LAND, Management, Assessment, Research, Knowledge base [[Online; accessed 25.03.2020]].

Le Page, M., Berjamy, B., Fakir, Y., Bourgin, R., Jarlan, L., Abourida, A., Benrhanem, M., Jacob, G., Huber, M., Sghrer, F., Simonneaux, V., & Chehbouni, G. (2012). An integrated dss for groundwater management based on remote sensing. the case of a semi-arid aquifer in morocco. *Water Resources Management*, *26*, 3209–3230.

Letcher, R. (2005). Implementation of a water allocation decision support system in the namoi and gwydir valleys.

Levatić, J., Kocev, D., Debeljak, M., & Džeroski, S. (2015). Community structure models are improved by exploiting taxonomic rank with predictive clustering trees. *Ecological Modelling*, *306*, 294–304.

Levatić, J., Kocev, D., & Džeroski, S. (2015). The importance of the label hierarchy in hierarchical multi-label classification. *Journal of Intelligent Information Systems*, *45*, 247–271.

Levatić, J., Ceci, M., Kocev, D., & Džeroski, S. (2017). Self-training for multi-target regression with tree ensembles. *Knowledge-Based Systems*, *123*, 41–60.

Levatić, J., Kocev, D., Ceci, M., & Džeroski, S. (2018). Semi-supervised trees for multi-target regression. *Information Sciences*, *450*, 109–127.

Liu, S., Ryu, D., Webb, J., Lintern, A., Waters, D., Guo, D., & Western, A. (2018). Characterisation of spatial variability in water quality in the great barrier reef catchments using multivariate statistical analysis. *Marine Pollution Bulletin*, *137*, 137–151.

Madjarov, G., & Gjorgjevikj, D. (2011). Hybrid decision tree architecture utilizing local svms for multi-label classification., In *Hybrid Artificial Intelligent Systems. HAIS 2012. Lecture Notes in Computer Science, vol 7209.*

Madjarov, G., Gjorgjevikj, D., Dimitrovski, I., & Džeroski, S. (2016). The use of data-derived label hierarchies in multi-label classification. *Journal of Intelligent Information Systems*, *47*(1), 57–90.

Madjarov, G., Gjorgjevikj, D., & Džeroski, S. (2012). Two stage architecture for multi-label learning. *Pattern Recognition*, *45*(3), 1019–1034.

Madjarov, G., Vidulin, V., Dimitrovski, I., & Kocev, D. (2019). Web genre classification with methods for structured output prediction. *Information Sciences*, *503*, 551–573.

Mahmoudzadeh, H., Matinfar, H. R., Taghizadeh-Mehrjardi, R., & Kerry, R. (2020). Spatial prediction of soil organic carbon using machine learning techniques in western iran. *Geoderma Regional*, *21*, e00260.

Marinković, B., Crnobarac, J., Brdar, S., Antić, B., Jaćimović, G., & Crnojević, V. (2009). Data mining approach for predictive modeling of agricultural yield data., In *First Int Workshop on Sensing Technologies in Agriculture, Forestry and Environment (BioSense09).*

Melssen, W., Wehrens, R., & Buydens, L. (2006). Supervised kohonen networks for classification problems. *Chemometrics and Intelligent Laboratory Systems*, *83*(2), 99–113.

Metzger, M. J., Bunce, R. G. H., Jongman, R. H. G., Mücher, C. A., & Watkins, J. W. (2005). A climatic stratification of the environment of europe. *Global Ecology and Biogeography*, *14*(6), 549–563.

Mevik, B.-H., & Wehrens, R. (2007). The pls package: Principal component and partial least squares regression in r. *Journal of Statistical Software*, *18*(2), 1–24.

Mileski, V., Džeroski, S., & Kocev, D. (2017). Predictive clustering trees for hierarchical multi-target regression (N. Adams, A. Tucker, & D. Weston, Eds.). In N. Adams, A. Tucker, & D. Weston (Eds.), *Advances in Intelligent Data Analysis XVI*, Springer International Publishing.

Mirschel, W., & Wenkel, K. (2007). Modelling soil–crop interactions with agrosim model family., In *Modelling water and nutrient dynamics in soil–crop systems.* Springer, Dordrecht.

Mitchell, T. (1997). *Machine learning (1 ed.).* McGraw-Hill, Inc., New York, NY, USA.

Moura, P., Barraud, S., Baptista, M. B., & Malard, F. (2011). Multicriteria decision-aid method to evaluate the performance of stormwater infiltration systems over the time. *Water Science and Technology*, *64*(10), 1993–2000.

Mouron, P., Aubert, U., Heijne, B., Naef, A., Strassemeyer, J., Hayer, F., Gaillard, G., Mack, G., Hernandez, J., Avilla, J., Sole, J., Sauphanor, B., Alaphilippe, A., Patocchi, A., Samietz, J., Hohn, H., Bravin, E., Lavigne, C., Bohanec, M., & Bigler, F. (2012). A multi-attribute decision method assessing the overall sustainability of crop protection strategies: A case on apple production in europe. In *Methods and procedures for building sustainable farming systems* (280 p.). Editions Springer.

Mueller, L., Schindler, U., Mirschel, W., Shepherd, T. G., Ball, B. C., Helming, K., Rogasik, J., Eulenstein, F., & Wiggering, H. (2010). Assessing the productivity function of soils. a review. *Agronomy for Sustainable Development*, *30*, 601–614.

Mulder, C., Cohen, J. E., Setälä, H., Bloem, J., & Breure, A. M. (2005). Bacterial traits, organism mass, and numerical abundance in the detrital soil food web of dutch agricultural grasslands. *Ecology Letters*, *8*(1), 80–90.

Navaratnam, R., Fitzgibbon, A. W., & Cipolla, R. (2007). The joint manifold model for semi-supervised multi-valued regression, In *2007 IEEE 11th International Conference on Computer Vision.*

Navarro-Hellín, H., Martínez-del-Rincon, J., Domingo-Miguel, R., Soto-Valles, F., & Torres-Sánchez, R. (2016). A decision support system for managing irrigation in agriculture. *Computers and Electronics in Agriculture*, *124*, 121–131.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, *135*(3), 370–384.

Nemenyi, P. B. (1963). *Distribution-free multiple comparisons* (Doctoral dissertation). Princeton University, Princeton, NY, USA.

Nguyen, T., Tjoa, A. M., & Wagner, R. (2000). An object oriented multidimensional data model for olap. *In Proceedings of the 1st International Conference on Web-Age Information Management (WAIM) in LNCS (Springer-Verlag)*, *1846*, 69–69.

Nikoloski, S., Kocev, D., & Džeroski, S. (2018). Structuring the output space in multi-label classification using feature ranking, In *New Frontiers in Mining Complex Patterns, 6th International Workshop, NFMCP 2017 Held in Conjunction with ECML-PKDD 2017 Skopje, Macedonia, September 18–22, 2017*, Springer LNAI 10785. https://doi.org/10.1007/978-3-319-78680-3_11

Nikoloski, S., Kocev, D., & Džeroski, S. (2019). Data-Driven Structuring of the Output Space Improves the Performance of Multi-Target Regressors [JCR IF = 4.098]. *IEEE Access*, *7*, 145177–145198. https://doi.org/10.1109/ACCESS.2019.2945084

Nikoloski, S., Kocev, D., Levatić, J., Wall, D. P., & Džeroski, S. (2020). Exploiting partially-labeled data in learning predictive clustering trees for multi-target regression: A case study of water quality assessment in ireland. [JCR IF = 2.31]. *Ecological Informatics*. https://doi.org/10.1016/j.ecoinf.2020.101161

Nikoloski, S., Murphy, P., Kocev, D., Džeroski, S., & Wall, D. P. (2019). Using machine learning to estimate herbage production and nutrient uptake on Irish dairy farms [JCR IF = 3.082]. *Journal of Dairy Science*, *102*, 10639–10656. https://doi.org/10.3168/jds.2019-16575

O'Brien, A., James, & Marakas, M., George. (2010). *Management information systems*. McGraw-Hill/Irwin.

OECD. (2001). *Environmental Indicators for Agriculture; Methods and Results*. Paris, France: OECD.

Okujeni, A., Canters, F., Cooper, S. D., Degerickx, J., Heiden, U., Hostert, P., Priem, F., Roberts, D. A., Somers, B., & der Linden], S. [(2018). Generalizing machine learning regression models using multi-site spectral libraries for mapping vegetation-impervious-soil fractions across multiple cities. *Remote Sensing of Environment*, *216*, 482–496.

Orgiazzi, A., Panagos, P., Yigini, Y., Dunbar, M. B., Gardi, C., Montanarella, L., & Ballabio, C. (2016). A knowledge-based approach to estimating the magnitude and spatial patterns of potential threats to soil biodiversity. *Science of The Total Environment*, *545-546*, 11–20.

Ottoy, S., Meerbeek, K. V., Sindayihebura, A., Hermy, M., & Orshoven, J. V. (2017). Assessing top- and subsoil organic carbon stocks of low-input high-diversity systems using soil and vegetation characteristics. *Science of The Total Environment*, *589*, 153–164.

Panov, P., Soldatova, L. N., & Džeroski, S. (2014). Ontology of core data mining entities. *Data Mining and Knowledge Discovery*, *28*, 1222–1256.

Panov, P., Soldatova, L. N., & Džeroski, S. (2016). Generic ontology of datatypes. *Information Sciences*, *329*, 900–920.

Pantazi, X., Moshou, D., Alexandridis, T., Whetton, R., & Mouazen, A. (2016). Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and Electronics in Agriculture*, *121*, 57–65.

Parton, W. J., & Rasmussen, P. E. (1994). Long-term effects of crop management in wheat-fallow: II. CENTURY model simulations. *Soil Science Society of America Journal*, *58*, 530–536.

Parton, W. J., Hartman, M., Ojima, D., & Schimel, D. (1998). Daycent and its land surface submodel: Description and testing. *Global and Planetary Change*, *19*(1), 35–48.

Pelzer, E., Fortino, G., Bockstaller, C., Angevin, F., Lamine, C., Moonen, C., Vasileiadis, V., Guérin, D., Guichard, L., Reau, R., & Messéan, A. (2012). Assessing innovative cropping systems with dexipm, a qualitative multi-criteria assessment tool derived from dexi. *Ecological Indicators*, *18*, 171–182.

Pereira, P., Bogunovic, I., Muñoz-Rojas, M., & Brevik, E. C. (2018). Soil ecosystem services, sustainability, valuation and management. *Current Opinion in Environmental Science & Health*, *5*, 7–13.

Pierce, S. A., Sharp, J. M., & Eaton, D. J. (2016). Decision support systems and processes for groundwater. In A. J. Jakeman, O. Barreteau, R. J. Hunt, J.-D. Rinaudo, & A. Ross (Eds.), *Integrated groundwater management: Concepts, approaches and challenges* (pp. 639–665). Cham, Springer International Publishing.

Pugelj, M., & Džeroski, S. (2011). Predicting structured outputs k-nearest neighbours method, In *Discovery Science, LNCS vol. 6926.*

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* San Francisco, CA, USA, Morgan Kaufmann Publishers Inc.

Quinlan, R. J. (1992). Learning with continuous classes., In *5th Australian Joint Conference on Artificial Intelligence, Singapore*, World Scientific, Singapore.

Ransom, C. J., Kitchen, N. R., Camberato, J. J., Carter, P. R., Ferguson, R. B., Fernández, F. G., Franzen, D. W., Laboski, C. A., Myers, D. B., Nafziger, E. D., Sawyer, J. E., & Shanahan, J. F. (2019). Statistical and machine learning methods evaluated for incorporating soil and weather into corn nitrogen recommendations. *Computers and Electronics in Agriculture*, *164*, 104872.

Rasmussen, C. E., & Williams, C. K. (2006). Gaussian processes for machine learning. *The MIT Press, Cambridge, MA, USA*, *38*, 715–719.

Read, J., Pfahringer, B., & Holmes, G. (2008). Multi-label classification using ensembles of pruned sets, In *2008 Eighth IEEE International Conference on Data Mining.*

Read, J., Martino, L., Olmos, P. M., & Luengo, D. (2015). Scalable multi-output label prediction: From classifier chains to classifier trellises. *Pattern Recognition*, *48*(6), 2096–2109.

Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2011). Classifier chains for multi-label classification. *Machine learning*, *85*, 333.

Recio, B., Ibáñez, J., Rubio, F., & Criado, J. (2005). A decision support system for analysing the impact of water restriction policies. *Decision Support Systems*, *39*(3), 385–402.

Riedmiller, M., & Braun, H. (1993). A direct adaptive method for faster backpropagation learning: The rprop algorithm. *In IEEE International Conference on Neural Networks (IEEE)*, 586–591.

Robnik-Sikonja, M., & Kononenko, I. (1997). An adaptation of relief for attribute estimation in regression.

Rosenberg, C., Hebert, M., & Schneiderman, H. (2005). Semi-supervised self-training of object detection models, In *Proceedings of the Seventh IEEE Workshops on Application of Computer Vision - Volume 01*, USA, IEEE Computer Society.

Rousu, J., Saunders, C., Szedmak, S., & Shawe-Taylor, J. (2006). Kernel-based learning of hierarchical multilabel classification models. *Journal of Machine Learning Research*, *7*, 1601–1626.

Russell, S., & Norvig, P. (2009). *Artificial intelligence: A modern approach* (3rd). USA, Prentice Hall Press.

Rutgers, M., Schouten, A. J., Bloem, J., Van Eekeren, N., De Goede, R. G. M., Jagersop Akkerhuis, G. A. J. M., Van der Wal, A., Mulder, C., Brussaard, L., & Breure, A. M. (2009). Biological measurements in a nationwide soil monitoring network. *European Journal of Soil Science*, *60*(5), 820–832.

Sak, H., Senior, A. W., & Beaufays, F. (2014). Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *CoRR*, *abs/1402.1128*arXiv.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & R., M. K. (2019). *Xplainable ai: Interpreting, explaining and visualizing deep learning*. Switzerland AG., Springer Nature.

Sánchez-Fernández, M., de-Prado-Cumplido, M., Arenas-García, J., & Pérez-Cruz, F. (2004). Svm multiregression for nonlinear channel estimation in multiple-input multiple output systems. *IEEE Transactions on Signal Processing*, *52*(8), 2298–2307.

Sandén, T., Trajanov, A., Spiegel, H., Kuzmanovski, V., Saby, N. P. A., Picaud, C., Henriksen, C. B., & Debeljak, M. (2019). Development of an agricultural primary productivity decision support model: A case study in france. *Frontiers in Environmental Science*, *7*, 58.

Schillaci, C., Acutis, M., Lombardo, L., Lipani, A., Fantappiè, M., Märker, M., & Saia, S. (2017). Spatio-temporal topsoil organic carbon mapping of a semi-arid mediterranean region: The role of land use, soil texture, topographic indices and the influence of remote sensing data to modelling. *Science of The Total Environment*, *601-602*, 821–832.

Schröder, J. J., Schulte, R. P. O., Creamer, R. E., Delgado, A., van Leeuwen, J., Lehtinen, T., Rutgers, M., Spiegel, H., Staes, J., Tóth, G., & Wall, D. P. (2016). The elusive role of soil quality in nutrient cycling: A review. *Soil Use and Management*, *32*(4), 476–486.

Schulte, R., Creamer, R. E., Donnellan, T., Farrelly, N., Fealy, R., ODonoghue, C., & OhUallachain, D. (2014). Functional land management: A framework for managing soil-based ecosystem services for the sustainable intensification of agriculture. *Environmental Science and Policy*, *38*, 45–58.

Schulte, R., Richards, K., Daly, K., Kurz, I., McDonald, E., & Holden, N. (2006). Agriculture, meteorology and water quality in Ireland: A regional evaluation of pressures and pathways of nutrient loss to water. *Biology and Environment: Proceedings of the Royal Irish Academy*, *106B*, 117–133.

Shaw, M. L., & Woodward, J. B. (1990). Modeling expert knowledge. *Knowledge Acquisition*, *2*(3), 179–206.

Silla Jr., C. N., & Freitas, A. A. (2009). A global-model naive bayes approach to the hierarchical prediction of protein functions, In *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*, USA, IEEE Computer Society.

Silla, C. N., & Freitas, A. (2011). A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, *22*, 31–72.

Snijders, T. A. B. (2011). Multilevel analysis. In M. Lovric (Ed.), *International encyclopedia of statistical science* (pp. 879–882). Berlin, Heidelberg, Springer Berlin Heidelberg.

Spyromitros-Xioufis, E., G., T., & I., V. (2008). An empirical study of lazy multilabel classification algorithms., In *Darzentas J., Vouros G.A., Vosinakis S., Arnellos A. (eds) Artificial Intelligence: Theories, Models and Applications. SETN 2008. Lecture Notes in Computer Science, vol 5138.*

Spyromitros-Xioufis, E., Groves, W., Tsoumakas, G., & Vlahavas, I. (2012). Multi-label classification methods for multi-target regression. *arXiv preprint arXiv:1211.6581 Cornall University Library*, 1159–1168.

Spyromitros-Xioufis, E., Tsoumakas, G., Groves, W., & Vlahavas, I. (2016). Multi-target regression via input space expansion: Treating targets as inputs. *Machine Learning*, *104*(1), 55–98.

Sra, S., & Dhillon, I. S. (2006). Generalized nonnegative matrix approximations with bregman divergences. In Y. Weiss, B. Schölkopf, & J. C. Platt (Eds.), *Advances in neural information processing systems 18* (pp. 283–290). MIT Press.

Stewart, G. W. (1993). On the early history of the singular value decomposition. *SIAM Review*, *35*(4), 551–566.

Struyf, J., & Džeroski, S. (2006). Constraint based induction of multi-objective regression trees, In *Proc. of the 4th International Workshop on Knowledge Discovery in Inductive Databases KDID - LNCS 3933*, Springer.

Suchithra, M., & Pai, M. L. (2020). Improving the prediction accuracy of soil nutrient classification by optimizing extreme learning machine parameters. *Information Processing in Agriculture*, *7*(1), 72–82.

Sulaeman, Y., Nursyamsi, D., Widowati, L., Husnaen, & Sarwani, M. (2012). Phosphorus and potassium decision support system: Bridging soil database and fertilizer application., In *Proceedings of the International Workshop on Soil Information System-oriented Nutrient Management for Major Asian Crops*, Science City of Munoz, Nueva Ecija, Philippines.

Suykens, J., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, *9*, 293–300.

Szymanski, P., Kajdanowicz, T., & Kersting, K. (2016). How is a data-driven approach better than random choice in label space division for multi-label classification? *Entropy*, *18*, 282.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics. (4th edition).* Needham Heights, MA: Allyn; Bacon.

Taki, M., Mehdizadeh, S. A., Rohani, A., Rahnama, M., & Rahmati-Joneidabad, M. (2018). Applied machine learning in greenhouse simulation; new application and analysis. *Information Processing in Agriculture*, *5*(2), 253–268.

Tandon, R., & Sra, S. (2010). *Sparse nonnegative matrix approximation: New formulations and algorithms* (tech. rep. No. 193). Max Planck Institute for Biological Cybernetics, Tübingen, Germany.

Teng, H.-f., Hu, J., Zhou, Y., Zhou, L.-q., & Shi, Z. (2019). Modelling and mapping soil erosion potential in china. *Journal of Integrative Agriculture*, *18*(2), 251–264.

Thoumazeau, A., Bessou, C., Renevier, M.-S., Trap, J., Marichal, R., Mareschal, L., Decaëns, T., Bottinelli, N., Jaillard, B., Chevallier, T., Suvannang, N., Sajjaphan, K., Thaler, P., Gay, F., & Brauman, A. (2019). Biofunctool®: A new framework to assess the impact of land management on soil quality. part a: Concept and validation of the set of indicators. *Ecological Indicators*, *97*, 100–110.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Toth, G., Gardi, C., Bodis, K., Ivits, E., Aksoy, E., Jones, A., Jeffrey, S., Petursdottir, T., & Montanarella, L. (2013). Continental-scale assessment of provisioning soil function in europe. *Ecological Processes*, *2*, 1–18.

Trajanov, A., Spiegel, H., Debeljak, M., & Sandén, T. (2018). Using data mining techniques to model primary productivity from international long-term ecological research (ilter) agricultural experiments in austria. *Regional Environmental Change*, *19*, 325–337.

Tsoumakas, G., & Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, *3*(3), 1–13.

Tsoumakas, G., Katakis, I., & Vlahavas, I. P. (2008). Effective and efficient multilabel classification in domains with large number of labels, In *Proceedings of the ecml/pkdd workshop on mining multidimensional data*.

Tsoumakas, G., & Vlahavas, I. (2008). Random k-labelsets: An ensemble method for multilabel classification., In *Kok J.N., Koronacki J., Mantaras R.L.., Matwin S., Mladenič D., Skowron A. (eds) Machine Learning: ECML 2007. ECML 2007. Lecture Notes in Computer Science.*

Turbé, A., De Toni, A., Benito, P., Lavelle, P., Lavelle, P., Ruiz, N., Van der Putten, W., Labouze, E., & Mudgal, S. (2010). *Soil biodiversity: Functions threats and tools for policy makers. bio intelligence service, ird and nioo, report for european commission (DG Environment)*. European Commission, Brussels.

Turunen, V., Sorvari, J., & Mikola, A. (2018). A decision support tool for selecting the optimal sewage sludge treatment. *Chemosphere*, *193*, 521–529.

Van de Broek, M., Henriksen, C. B., Ghaley, B. B., Lugato, E., Kuzmanovski, V., Trajanov, A., Debeljak, M., Sandén, T., Spiegel, H., Decock, C., Creamer, R., & Six, J. (2019). Assessing the climate regulation potential of agricultural soils using a decision support tool adapted to stakeholders' needs and possibilities. *Frontiers in Environmental Science*, *7*, 131.

van der Merwe, A., & Zidek, J. V. (1980). Multivariate regression analysis and canonical variates. *Canadian Journal of Statistics*, *8*, 27–39.

van der Tol, C., Verhoef, W., Timmermans, J., Verhoef, A., & Su, Z. (2009). An integrated model of soil-canopy spectral radiances, photosynthesis, fluorescence, temperature and energy balance. *Biogeosciences*, *6*, 3109–3129.

van Leeuwen, J. P., Creamer, R. E., Cluzeau, D., Debeljak, M., Gatti, F., Henriksen, C. B., Kuzmanovski, V., Menta, C., Pérès, G., Picaud, C., Saby, N. P. A., Trajanov, A., Trinsoutrot-Gattin, I., Visioli, G., & Rutgers, M. (2019). Modeling of soil functions for assessing soil quality: Soil biodiversity and habitat provisioning. *Frontiers in Environmental Science*, *7*, 113.

Vens, C., Struyf, J., Schietgat, L., Džeroski, S., & Blockeel, H. (2008). Decision trees for hierarchical multi-label classification. *Machine Learning*, *73*(2), 185–214.

Vogel, H.-J., Bartke, S., Daedlow, K., Helming, K., Kögel-Knabner, I., Lang, B., Rabot, E., Russell, D., Stößel, B., Weller, U., Wiesmeier, M., & Wollschläger, U. (2018). A systemic approach for modeling soil functions. *SOIL*, *4*(1), 83–92.

Wall, D., O'Sullivan, L., Debeljak, M., Trajanov, A., Schroder, J., Henriksen, C. B., Creamer, R. E., Cacovean, H., & Delgado, A. (2019). *Key indicators and management strategies for water purification and regulation.* LANDMARK: Land Management Assessment Research Knowledge base (EU H2020 project).

Wang, J., Chen, Z., Sun, K., Li, H., & Deng, X. (2019). Multi-target regression via target specific features. *Knowledge-Based Systems*, *170*, 70–78.

Wang, S., Wang, J., Wang, Z., & Ji, Q. (2014). Enhancing multi-label classification by modeling dependencies among labels. *Pattern Recognition*, *47*, 3405–3413.

Wang, Y., & Witten, I. H. (1997). Induction of model trees for predicting continuous classes., In *Poster papers of the 9th european conference on machine learning, 1997.* ECML97, Prague, Czech Republic.

WFD. (2000). *Council directive of 23 october 2000 establishing a framework for community action in the field of water policy, 2000/60/ec.* Water Framework Directive, Brussels, European Commission.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques.* Morgan Kaufmann.

Wolanin, A., Camps-Valls, G., Gómez-Chova, L., Mateo-García, G., van der Tol, C., Zhang, Y., & Guanter, L. (2019). Estimating crop primary productivity with sentinel-2 and landsat 8 using machine learning methods trained with radiative transfer simulations. *Remote Sensing of Environment*, *225*, 441–457.

Wold, H. (1997). Estimation of principal components and related models by iterative least squares., In *In Multivariate Analysis. Proceedings of an International Symposium held in Dayton, Ohio, June 14-19, 1965, edited by P. R. Krishnaiah. Academic Press.*

Wu, Q., Tan, M., Song, H., Chen, J., & Ng, M. K. (2016). Ml-forest: A multi-label tree ensemble method for multi-label classification. *IEEE Transactions on Knowledge and Data Engineering*, *28*(10), 2665–2680.

Wu, Q., Ye, Y., Zhang, H., Chow, T. W. S., & Ho, S. (2015). Ml-tree: A tree-structure-based approach to multilabel learning. *IEEE Transactions on Neural Networks and Learning Systems*, *26*(3), 430–443.

WWAP. (2015). *The United Nations World Water Development Report 2015: Water for a Sustainable World.* UNESCO, Paris.

Xu, L., Yu, G., He, N., Wang, Q., Gao, Y., Wen, D., Li, S., Niu, S., & Ge, J. (2018). Carbon storage in china's terrestrial ecosystems: A synthesis. *Scientific Reports*, *8*, 2806.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods, In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, Cambridge, Massachusetts, Association for Computational Linguistics.

Ying-xue, S., Huan, X., & Li-jiao, Y. (2017). Support vector machine-based open crop model (SBOCM): Case of rice production in china [Computational Intelligence Research and Approaches in Bioinformatics and Biocomputing]. *Saudi Journal of Biological Sciences*, *24*(3), 537–547.

Zhang, M.-L. (2009). Ml-rbf: Rbf neural networks for multi-label learning. *Neural Processing Letters volume*, *29*, 61–74.

Zhang, M.-L., Li, Y.-K., Liu, X.-Y., & Geng, X. (2018). Binary relevance for multi-label learning: An overview. *Frontiers of Computer Science*, *12*, 191–202.

Zhang, M.-L., & Zhou, Z.-H. (2006). Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, *18*(10), 1338–1351.

Zhang, M.-L., & Zhou, Z.-H. (2007). Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, *40*, 2038–2048.

Zhang, W., Liu, X., Ding, Y., & Shi, D. (2012). Multi-output ls-svr machine in extended feature space., In *Proc. of the 2012 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications*.

Zhang, Y., & Yeung, D.-Y. (2009). Semi-supervised multi-task regression (W. Buntine, M. Grobelnik, D. Mladenić, & J. Shawe-Taylor, Eds.). In W. Buntine, M. Grobelnik, D. Mladenić, & J. Shawe-Taylor (Eds.), *Machine Learning and Knowledge Discovery in Databases*, Berlin, Heidelberg, Springer Berlin Heidelberg.

Zhen, X., Yu, M., He, X., & Li, S. (2018). Multi-target regression via robust low-rank learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(2), 497–504.

Zhou, T., Tao, D., & Wu, X. (2012). Compressed labeling on distilled labelsets for multi-label learning. *Machine learning*, *88*, 69–126.

Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *3*(1), 1–130.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.

Bibliography

Publications Related to the Thesis

Journal Articles

Nikoloski, S., Kocev, D., & Džeroski, S. (2019). Data-Driven Structuring of the Output Space Improves the Performance of Multi-Target Regressors [JCR IF = 4.098]. *IEEE Access*, *7*, 145177–145198. https://doi.org/10.1109/ACCESS.2019.2945084

Nikoloski, S., Kocev, D., Levatić, J., Wall, D. P., & Džeroski, S. (2020). Exploiting partially-labeled data in learning predictive clustering trees for multi-target regression: A case study of water quality assessment in ireland. [JCR IF = 2.31]. *Ecological Informatics*. https://doi.org/10.1016/j.ecoinf.2020.101161

Nikoloski, S., Murphy, P., Kocev, D., Džeroski, S., & Wall, D. P. (2019). Using machine learning to estimate herbage production and nutrient uptake on Irish dairy farms [JCR IF = 3.082]. *Journal of Dairy Science*, *102*, 10639–10656. https://doi.org/10.3168/jds.2019-16575

Conference Paper

Nikoloski, S., Kocev, D., & Džeroski, S. (2018). Structuring the output space in multi-label classification using feature ranking, In *New Frontiers in Mining Complex Patterns, 6th International Workshop, NFMCP 2017 Held in Conjunction with ECML-PKDD 2017 Skopje, Macedonia, September 18–22, 2017*, Springer LNAI 10785. https://doi.org/10.1007/978-3-319-78680-3_11

Other Publications:

Conference Abstracts

Nikoloski, S., Debeljak, M., Creamer, R., Wall, D. P., Džeroski, S., & Trajanov, A. (2017). Prediction of mineralizable nitrogen (N) in soils using ensembles of regression models, In *Abstract book Pedometrics 2017, Wageningen, 26 June - 1 July 2017*, Wageningen. https://static1.squarespace.com/static/5653202ee4b037d305e7fd3e/t/594a53f7e4fcb553cd43b9c0/1498043388899/Abstract+Book+Pedometrics+2017.pdf

Biography

Stevanche Nikoloski was born on 9 January 1987 in Prilep, Macedonia, where he finished elementary and secondary school. In 2005, he enrolled in the undergraduate studies of mathematics at the Institute of Mathematics, Faculty of Natural Sciences and Mathematics at the "Ss. Cyril and Methodius" University in Skopje. During his studies, he received a scholarship for talented students from the Ministry of Education of Macedonia. He graduated on 28 October 2009. His diploma thesis was entitled "An overview of the Riesz's representation theorem".

In 2010, he enrolled in the postgraduate study program of Applied Mathematics at the Institute of Mathematics, as a part of the Faculty of Natural Sciences and Mathematics at the "Ss. Cyril and Methodius" University in Skopje. On 4 July 2012, he defended his master thesis entitled "Sequential quadratic programming (SQP). An application of SQP in approximation and design of spline curves" thus obtained the degree "Master of mathematical sciences and their applications".

Since 2015, he has been a PhD student at the Jožef Stefan International Postgraduate School (MPŠ), enrolled in the program of Information and Communication Technologies. In the period October 2015 – September 2019, Nikoloski held a scholarship from the MPŠ, as a part of the TEAGASC Walsh Fellowship program within the LANDMARK EU project. Since September 2019, he has been employed at the Department of Knowledge Technologies at the Jožef Stefan Institute. His research is mostly in the field of machine learning, and its applications in the domain of environmental and soil sciences. More specifically, his research topic is based on structuring the output space in multi-target prediction tasks (multi-label classification and multi-target regression) and application of the methods for structured output prediction in the environmental domain.

During his PhD studies, he has participated in several workshops and conferences, such as PEDOMETRICS 2017 and ECML PKDD 2017. His work has been published in several journal and conference/workshop papers. These covers both the areas of computer science and environmental sciences.