The use of the label hierarchy in HMC improves performance: A case study in predicting community structure in ecology

Jurica Levatić[1,2], Dragi Kocev[1], and Sašo Džeroski[1,2]

[1] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
[2] Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
Jurica.Levatic@ijs.si, Dragi.Kocev@ijs.si, Saso.Dzeroski@ijs.si

Abstract. In this article, we address the task of learning models for predicting structured outputs. We consider both global and local prediction of structured outputs, the former based on a single model that predicts the entire output structure and the latter based on a collection of models, each predicting a component of the output structure. More specifically, we investigate whether the global models have better predictive performance than the local predictive models. Moreover, we discuss the interpretability power of the obtained models. Furthermore, we evaluate the predictive models on two case studies from ecological modelling. Finally, we identify the properties of the data and the eco-system under consideration that lead to the differences in the performance.

Keywords: predictive clustering trees, hierarchical multi-label classification, multi-label classification, habitat modelling

1 Introduction

Supervised learning is one of the most widely researched and investigated areas of machine learning. The goal in supervised learning is to learn, from a set of examples with known class, a function that outputs a prediction for the class of a previously unseen example. If the examples belong to two classes (e.g., the example has some property or not) the task is called binary classification. The task where the examples can belong to a single class from a given set of m classes ($m \geq 3$) is known as multi-class classification. The case where the output is a real value is called regression.

However, in many real life problems of predictive modelling the output (i.e., the target) is structured, meaning that there can be dependencies between classes (e.g., classes are organized into a tree-shaped hierarchy or a directed acyclic graph) or some internal relations between the classes (e.g., sequences). These types of problems occur in domains such as life sciences (predicting gene function, finding the most important genes for a given disease, predicting toxicity of molecules, etc.), ecology (analysis of remotely sensed data, habitat modelling), multimedia (annotation and retrieval of images and videos) and the semantic

web (categorization and analysis of text and web pages). Having in mind the needs of these application domains and the increasing quantities of structured data, Kriegel et al. [1] and Dietterich et al. [2] listed the task of "mining complex knowledge from complex data" as one of the most challenging problems in machine learning.

A variety of methods, specialized in predicting a given type of structured output (e.g., a hierarchy of classes [3]), have been proposed [4]. These methods can be categorized into two groups of methods for solving the problem of predicting structured outputs [3, 4]: (1) local methods that predict component(s) of the output and then combine the individual models to get the overall model and (2) global methods that predict the complete structure as a whole (also known as 'big-bang' approaches). The global methods have several advantages over the local methods. First, they exploit and use the dependencies that exist between the components of the structured output in the model learning phase, which can result in better predictive performance. Next, they are typically more efficient: it can easily happen that the number of components in the output is very large (e.g., hierarchies in functional genomics can have several thousands of components), in which case executing a basic method for each component is not feasible. Furthermore, they produce models that are typically smaller than the sum of the sizes of the models built for each of the components.

Albeit the many interesting applications and the developed methods, it is not clear when it is favorable (performance wise) to construct global models and when local models. In this work, we focus on this important issue for the task of hierarchical multi-label classification (HMC). HMC is a variant of classification where a single example may belong to multiple classes at the same time and the classes are organized in a form of hierarchy. An example that belongs to some class c automatically belongs to all super-classes of c: This is called the hierarchical constraint. Problems of this kind can be found in many domains including text classification, functional genomics, and object/scene classification. Silla and Freitas [3] give a detailed overview of the possible application areas and the available approaches to HMC.

We construct four types of predictive models that exploit different amounts of the information provided by the output structure, i.e., the hierarchical organization of the classes. We investigate the predictive performance of simple single-class classification trees, hierarchical single-label classification trees, multi-label classification trees and HMC trees. The first two predictive models are local models, while the last two are global models.

The predictive models that we consider in this article are predictive clustering trees (PCTs). They can be considered as a generalization of standard decision trees towards predicting structured outputs. PCTs offer a unifying approach for dealing with different types of structured outputs and construct the predictive models very efficiently. They are able to make predictions for several types of structured outputs: tuples of continuous/discrete variables, hierarchies of classes, and time series. More details about the PCT framework can be found in [5–7].

We perform the evaluation of the predictive models on two practically relevant datasets from the task of habitat modelling [8]. Habitat modelling focuses on the spatial aspects of the distribution and abundance of plants and animals. It studies the relationships between environmental variables and the presence/abundance of plants and animals. This is typically done under the implicit assumption that both are observed at a single point in time for a given spatial unit (i.e., sampling site). We investigate the effect of environmental conditions on communities of organisms in two different ecosystems. Namely, we consider the Collembola community in the soils of Denmark [9] and organisms living in Slovenian rivers [10]. The structured output space in these case studies is the taxonomic hierarchy of the species under consideration.

The remainder of this paper is organized as follows. Section 2 explains the predictive clustering trees framework and the extensions for the different tasks considered here. The experimental setup is given in Section 3. Section 4 presents the obtained results. Finally, the conclusions are stated in Section 5.

2 Predictive modelling for HMC

In this section, we present the methodology used to construct the predictive models. We first present global predictive models that predict the complete output with a single model (i.e., a single model for all of the species present in the dataset). We then overview local predictive models that construct several models - each one predicting a part of the output (i.e., a model for each species separately).

2.1 Global predictive models

The Predictive Clustering Trees (PCTs) framework views a decision tree as a hierarchy of clusters: the top-node corresponds to one cluster containing all data, which is recursively partitioned into smaller clusters while moving down the tree. The PCT framework is implemented in the CLUS system [11], which is available for download at `http://clus.sourceforge.net`.

PCTs are induced with a standard *top-down induction of decision trees* (TDIDT) algorithm [12]. The algorithm is presented in Table 1. It takes as input a set of examples (E) and outputs a tree. The heuristic (h) that is used for selecting the tests (t) is the reduction in variance caused by partitioning ($\mathcal{P}$) the instances (see line 4 of the BestTest procedure in Table 1). By maximizing the variance reduction, the cluster homogeneity is maximized and the predictive performance is improved.

The main difference between the algorithm for learning PCTs and a standard decision tree learner is that the former considers the variance function and the prototype function, that computes a label for each leaf, as *parameters* that can be instantiated for a given learning task. So far, PCTs have been instantiated for the following tasks: multi-target prediction (which includes multi-label classification) [6], hierarchical multi-label classification [7] and prediction of time-series [13]. In this article, we focus on the first two tasks.

Table 1. The top-down induction algorithm for PCTs.

procedure PCT
Input: A dataset E
Output: A predictive clustering tree
1: $(t^*, h^*, \mathcal{P}^*) = \text{BestTest}(E)$
2: **if** $t^* \neq none$ **then**
3: **for each** $E_i \in \mathcal{P}^*$ **do**
4: $tree_i = \text{PCT}(E_i)$
5: **return** node$(t^*, \bigcup_i \{tree_i\})$
6: **else**
7: **return** leaf(Prototype(E))

procedure BestTest
Input: A dataset E
Output: the best test (t^*), its heuristic score (h^*) and the partition ($\mathcal{P}^*$) it induces on the dataset (E)
1: $(t^*, h^*, \mathcal{P}^*) = (none, 0, \emptyset)$
2: **for each** possible test t **do**
3: $\mathcal{P}$ = partition induced by t on E
4: $h = Var(E) - \sum_{E_i \in \mathcal{P}} \frac{|E_i|}{|E|} Var(E_i)$
5: **if** $(h > h^*) \wedge \text{Acceptable}(t, \mathcal{P})$ **then**
6: $(t^*, h^*, \mathcal{P}^*) = (t, h, \mathcal{P})$
7: **return** $(t^*, h^*, \mathcal{P}^*)$

PCTs for multi-label classification PCTs for multi-label classification can be considered as PCTs that are able to predict multiple discrete targets simultaneously. Therefore, the variance function for the PCTs for MLC is computed as the sum of the Gini indices of the target variables, i.e., $Var(E) = \sum_{i=1}^{T} Gini(E, Y_i)$. Furthermore, one can also use the sum of the entropies of class variables as a variance function, i.e., $Var(E) = \sum_{i=1}^{T} Entropy(E, Y_i)$ (this definition has also been used in the context of multi–label prediction [14]). The CLUS system also implements other variance functions, such as reduced error, gain ratio and the m-estimate. The prototype function returns a vector of probabilities that an instance belongs to a given class for each target variable. Using these probabilities, the most probable (majority) class for each target attribute can be calculated.

PCTs for hierarchical multi–label classification CLUS-HMC is the instantiation (with the distances and prototypes as defined below) of the PCT algorithm for hierarchical classification implemented in the CLUS system [7]. The variance and prototype are defined as follows. First, the set of labels of each example is represented as a vector with binary components; the i'th component of the vector is 1 if the example belongs to class c_i and 0 otherwise. It is easily checked that the arithmetic mean of a set of such vectors contains as i'th component the proportion of examples of the set belonging to class c_i. The variance of a set of examples E is defined as the average squared distance between each example's class vector (L_i) and the set's mean class vector ($\overline{L}$), i.e.,

$$Var(E) = \frac{1}{|E|} \cdot \sum_{E_i \in E} d(L_i, \overline{L})^2.$$

In the HMC context, the similarity at higher levels of the hierarchy is more important than the similarity at lower levels. This is reflected in the distance

measure used in the above formula, which is a weighted Euclidean distance:

$$d(L_1, L_2) = \sqrt{\sum_{l=1}^{|L|} w(c_l) \cdot (L_{1,l} - L_{2,l})^2},$$

where $L_{i,l}$ is the l'th component of the class vector L_i of an instance E_i, $|L|$ is the size of the class vector, and the class weights $w(c)$ decrease with the depth of the class in the hierarchy. More precisely, $w(c) = w_0 \cdot \{w(p(c))\}$, where $p(c)$ denotes the parent of class c and $0 < w_0 < 1$).

For example, consider the toy class hierarchy shown in Figure 1(a,b), and two data examples: (X_1, S_1) and (X_2, S_2) that belong to the classes $S_1 = \{c_1, c_2, c_{2.2}\}$ (boldface in Figure 1(b)) and $S_2 = \{c_2\}$, respectively. We use a vector representation with consecutive components representing membership of class c_1, c_2, $c_{2.1}$, $c_{2.2}$ and c_3, in that order (preorder traversal of the tree of class labels). The distance is then calculated as follows:

$$d(S_1, S_2) = d([1, 1, 0, 1, 0], [0, 1, 0, 0, 0]) = \sqrt{w_0 + w_0^2}.$$

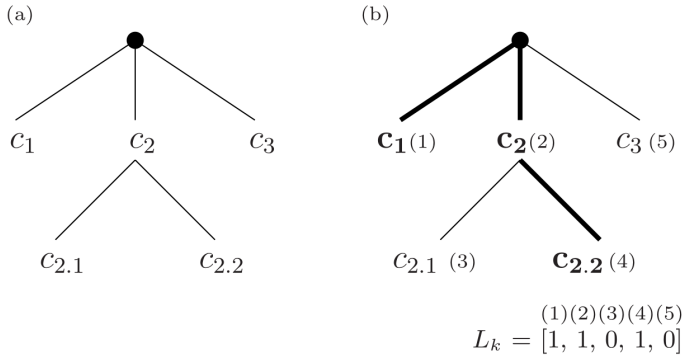


Fig. 1. Toy examples of hierarchies structured as a tree. (a) Class label names contain information about the position in the hierarchy, e.g., $c_{2.1}$ is a subclass of c_2. (b) The set of classes $S_1 = \{c_1, c_2, c_{2.2}\}$, shown in bold in the hierarchy, represented as a vector (L_k).

Recall that the instantiation of PCTs for a given task requires proper instantiation of the variance and prototype functions. The variance function for the HMC task is instantiated by using the weighted Euclidean distance measure (as given above), which is further used to select the best test for a given node by calculating the heuristic score (line 4 from the algorithm in Table 1). We now discuss the instantiation of the prototype function for the HMC task.

A classification tree stores in a leaf the majority class for that leaf, which will be the tree's prediction for all examples that will arrive in the leaf. In the

case of HMC, an example may have multiple classes, thus the notion of *majority class* does not apply in a straightforward manner. Instead, the mean $\bar{L}$ of the class vectors of the examples in the leaf is stored as a prediction. Note that the value for the i-th component of $\bar{L}$ can be interpreted as the probability that an example arriving at the given leaf belongs to class c_i.

The prediction for an example that arrives at the leaf can be obtained by applying a user defined threshold τ to the probability; if the i-th component of $\bar{L}$ is above τ then the examples belong to class c_i. When a PCT is making a prediction, it preserves the hierarchy constraint (the predictions comply with the parent-child relationships from the hierarchy) if the values for the thresholds τ are chosen as follows: $\tau_i \leq \tau_j$ whenever $c_i \leq_h c_j$ (c_i is ancestor of c_j). The threshold τ is selected depending on the context. The user may set the threshold such that the resulting classifier has high precision at the cost of lower recall or vice versa, to maximize the F-score, to maximize the interpretability or plausibility of the resulting model etc. In this work, we use a threshold-independent measure (precision-recall curves) to evaluate the performance of the HMC models.

2.2 Local habitat models

Local predictive models of structured outputs use a collection of predictive models, each predicting a component of the overall structure that needs to be predicted. The local predictive models for the task of predicting multiple targets are constructed by learning a predictive model for each of the targets separately. In the task of hierarchical multi-label classification, however, there are four different approaches that can be used: flat classification, local classifiers per level, local classifiers per node, and local classifiers per parent node (see [3] for details).

Vens et al. [7] investigated the performance of the last two approaches with local classifiers over a large collection of datasets from functional genomics. The conclusion of the study was that the last approach (called hierarchical single-label classification - HSC) performs better in terms of predictive performance, smaller total model size and faster induction times.

In particular, the CLUS-HSC algorithm by Vens et al. [7] constructs a decision tree classifier for each edge (connecting a class c with a parent class $par(c)$) in the hierarchy, thus creating an architecture of classifiers. The corresponding tree predicts membership to class c, using the instances that belong to $par(c)$. The construction of this type of trees uses few instances, as only instances labeled with $par(c)$ are used for training. The instances labeled with class c are positive instances, while the ones that are labeled with $par(c)$, but not with c are negative.

The resulting HSC tree predicts the conditional probability $P(c|par(c))$. A new instance is predicted by recursive application of the product rule $P(c) = P(c \,|\, par(c)) \cdot P(par(c))$, starting from the tree for the top-level class. Again, the probabilities are thresholded to obtain the set of predicted classes. To satisfy the hierarchy constraint, the threshold τ should be chosen as in the case of CLUS-HMC.

In this work, we also construct single-label classification trees. We construct these models by setting the number of labels to 1 and use the same algorithm as for the multi-label classification models.

3 Experimental design

In this section, we present the design of the experimental evaluation. We begin by describing the data used in the case study. Next, we outline the specific experimental setup for constructing the predictive models. Finally, we give the evaluation measure for assessing the predictive performance of the predictive models.

3.1 Data description

We use datasets from two studies that concern two eco-systems: river and soil. Namely, we construct habitat models for river water organisms living in the Slovenian rivers [10] and for soil microarthropods from Danish farms [9].

The data for the water organisms from Slovenian rivers come from the Hydrometeorological Institute of Slovenia (now Environmental Agency of Slovenia) that performs water quality monitoring for Slovenian rivers and maintains a database of water quality samples. The data provided cover a six year period of monitoring, starting from 1990 until 1995. Biological samples were taken twice a year, once in summer and once in winter, while physical and chemical samples were taken several times a year for each sampling site. In total, there are 1060 samples, each is described with 16 attributes corresponding to physical and chemical properties of water. Presence of 491 species is recorded at each sampling site, with an average of 25 species per site. Species are organized in taxonomic hierarchy with 724 nodes, with maximal depth of 4 (most general taxonomic rank is 'order').

The data for the soil microarthropods from Danish farms describes four experimental farming systems (observed during the period 1989-1993) and a number of organic farms in Denmark (observed during the period 2002-2003). Soil samples were collected within a $20m \times 20m$ area of the field, with a distance of $5m$ between the individual samples. Sampling was performed in the upper 5.5 cm soil layer and the sampling containers measured 6 cm in diameter. The data concerns the *Collembola* species community in the soil samples. These species can be used as indicators of the soil quality (in particular soil dessication) and some are considered as pests for the plants. Also, they are one of the main biological factors responsible for the control of the soil microorganisms [15]. In total, there are 1944 sites, each described with 137 attributes corresponding to various agricultural events and soil biological parameters. Presence of 35 species is recorded at each site, with an average of 7 species per site. Species are organized in taxonomic hierarchy with 72 nodes, with maximal depth of 3 (most general taxonomic rank is 'family').

3.2 Experimental design

We constructed four types of predictive models described in the previous section for each of the case studies. First, we constructed single-label classification trees for each species separately. Next, we constructed hierarchical single-label classification tree for each species. Furthermore, we constructed multi-label classification tree for all of the species, but without using the hierarchy. Finally, we constructed a hierarchical multi-label classification tree for all of the species with using the hierarchy.

We used F-test pruning to ensure that the produced models are not overfitted and have better predictive performance. This pruning procedure uses the exact Fisher test to check whether a given split/test in an internal node of the tree results in a reduction in variance that is statistically significant at a given significance level. If there is no split/test that can satisfy this, then the node is converted to a leaf. An optimal significance level was selected by using internal 3-fold cross validation, from the following values: 0.125, 0.1, 0.05, 0.01, 0.005 and 0.001.

The w_0 parameter determines the class weights with respect to the depth of the class within the hierarchy used for learning the CLUS-HMC model. We considered 3 different values of w_0: 0.75, 1 and 1.25, meaning that the classes at higher levels of hierarchy are more important, all classes are equally important and classes at lower levels of the hierarchy are more important, respectively. Different choices of w_0 yielded similar results, with $w_0 = 1$ performing slightly better. Performance measures presented in Section 4 correspond to the CLUS-HMC model learned with $w_0 = 1$.

3.3 Evaluation measures

We evaluate the algorithms using the Area Under the Precision-Recall Curve (AUPRC), and in particular, the Area Under the Average Precision-Recall Curve (AU$\overline{\text{PRC}}$) as suggested by Vens et al. [7]. The points in the PR space are obtained by varying the value for the threshold τ from 0 to 1 with step 0.02. For each value of the threshold τ, precision and recall are micro-averaged as follows:

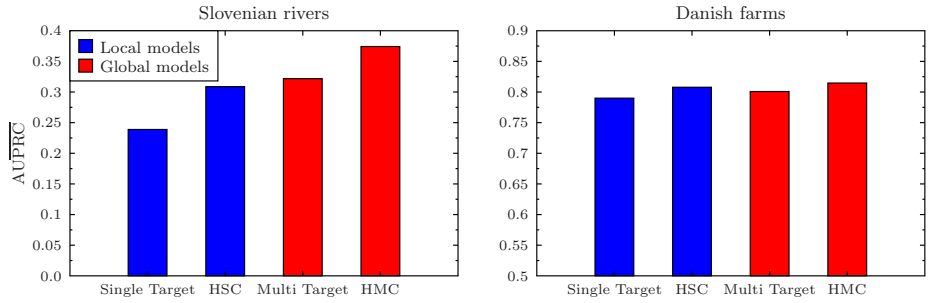
$$\overline{Prec} = \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FP_i}, \quad \text{and} \quad \overline{Rec} = \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FN_i}$$

where i ranges over all classes. In the case of hierarchical classification, only the performance on the classes which correspond to leafs in the taxonomic hierarchy (i.e., species) are taken into account. To estimate the predictive performance of the obtained models, we used the 10-fold cross-validation procedure.

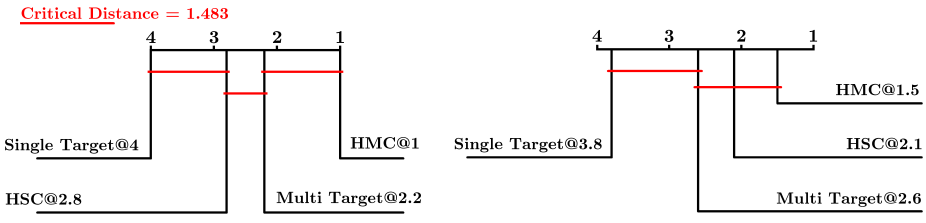
4 Results and discussion

In this section, we present the results from the experimental evaluation. We discuss the obtained models first by their predictive performance and then by their interpretability power.

The predictive performance of the models is given in Figure 2a. A quick inspection of the performance shows that the global models are better than the local models on the river communities study and both types of models perform equally well on the soil communities study (with the note that the HMC model performs slightly better than rest of the models). To test whether the observed differences are statistically significant, we followed the methodology proposed by Demšar [16]: Applying the Friedman test to the per-fold performance figures for each dataset separately, shows that the AU$\overline{\text{PRC}}$s are statistically different with a *p-value* $= 3.3 \times 10^{-16}$ (Slovenian rivers) and 3.5×10^{-5} (Danish farms). Figure 2b shows the average ranks for all models together, obtained with the Nemenyi post-hoc test. We can see that the HMC model performs significantly better than the Single Target models at both datasets, and better than HSC for the Slovenian rivers dataset. The Multi Target model performs significantly better than the Single Target model for the Slovenian rivers dataset. We explain the findings from a machine learning perspective and from an ecological perspective.



(a)



(b)

Fig. 2. (a) The area under the average precision-recall curve (AU$\overline{\text{PRC}}$) scores for the constructed predictive models. (b) Average ranks diagrams of AU$\overline{\text{PRC}}$s. Better algorithms are on the right-hand side, the ones that differ by less than critical distance for a *p-value* $= 0.05$ are connected with a horizontal bar. The number after the name of an algorithm indicates its average rank.

From a machine learning point of view, the data from the two studies have quite different properties. The average number of species per sample for the river communities (25 labels per example) is much larger than the one for the soil communities (7 labels per example). Therefore, the taxonomic hierarchy is much more populated in the former case. Arguably, such a scenario enables the HMC tree to fully exploit the hierarchy and thus produce a predictive model with a better predictive performance. We also hypothesize that the type of the descriptive attributes influenced the performance. In the study of river organisms, all of the descriptive attributes are continuous, while in the study of the soil organisms, the majority of the descriptive attributes are discrete. These continuous descriptive attributes, in this case, enable the tree construction algorithm to perform the test selection more granularly, thus obtaining more optimal tests.

From an ecological point of view, the two eco-systems studies here have different properties. More specifically, the soil eco-system is more stable than the river eco-system. This means that the soil eco-system can be more efficiently described with simpler measurements (i.e., it is much easier to monitor it). Hence, the constructed habitat models (constructed with any method) are of high quality and predictive performance (the average value of AU$\overline{\text{PRC}}$ is 0.80). Including additional information from the taxonomic ranks does not increase much the predictive performance of the constructed models. On the other hand, the river eco-system is difficult to monitor. In order to describe the state of the system, one needs to perform time-course large-scale measurements (i.e., the measurements need to cover more than a few parameters). This is the main reason because the predictive models for the river eco-system have lower predictive performance than the one from the soil eco-system (the average value of AU$\overline{\text{PRC}}$ is 0.31). Therefore, including additional information (i.e., the taxonomic ranks) helps to significantly improve the predictive performance.

In habitat modelling, besides the predictive power of the models, their interpretability is also a highly desired property. The predictive models that we consider here (PCTs) are readily interpretable. However, the difference in the interpretability of the local and global models is easy to notice. In Figure 3, we present illustrative examples of the predictive models for the Slovenian rivers dataset. We show the PCTs for single-label classification, multi-label classification and hierarchical multi-label classification.

We can immediately notice the different between the local and global predictive models. The local models[3] offer an information only for a part for the output space, i.e., they are valid just for a single species. In order to reconstruct the complete community model, one needs to look at the separate models and then try to make some overall conclusions. However, this could be very tedious or even impossible in domains with high biodiversity and where there are hundreds of species present, such as the domain we consider here - Slovenian rivers.

[3] Note that the hierarchical single-label classification models will be much similar to the single-label classification models, with the difference that the predictive models are organized into an hierarchical architecture. This makes the interpretation of the HSC models even more difficult task.

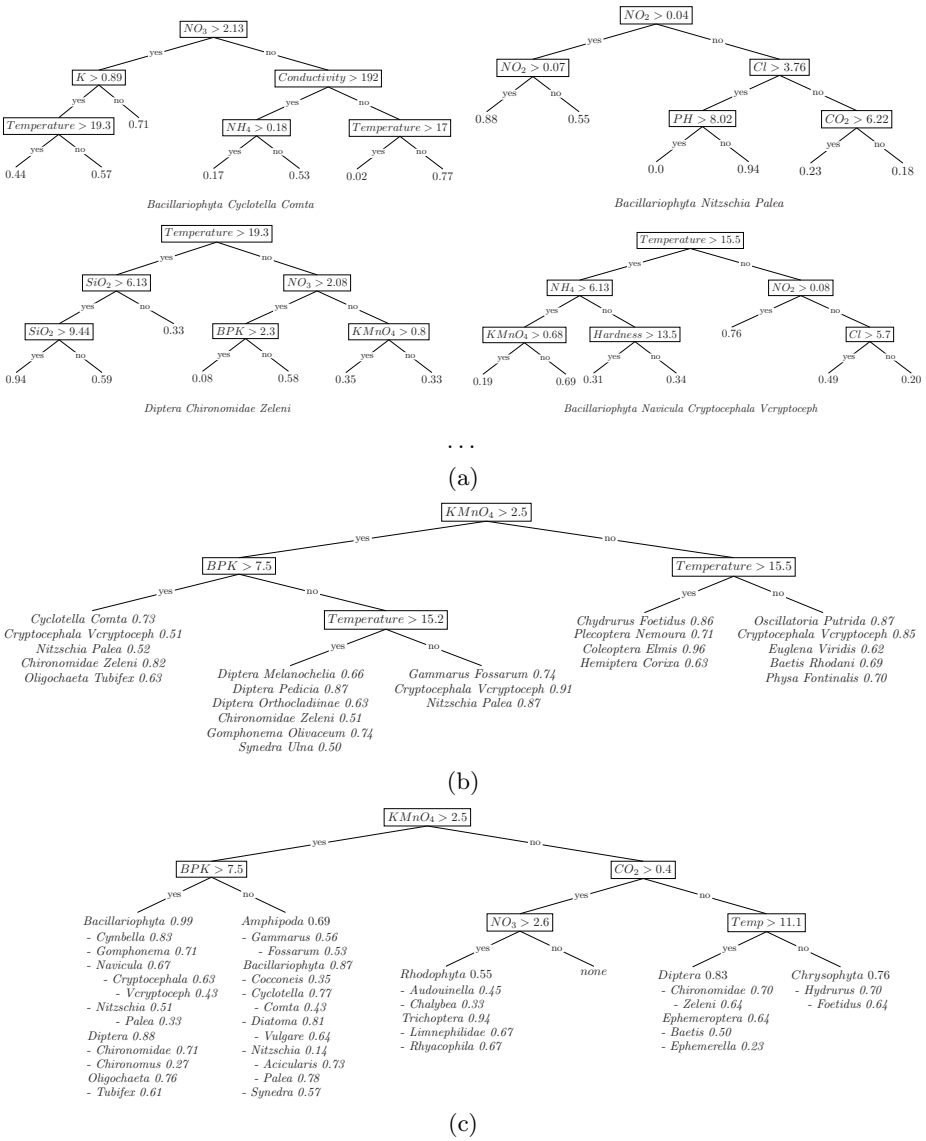


Fig. 3. Illustrative examples of decision trees for Slovenian rivers dataset constructed with PCTs. Single target classification (a) produces a separate model for each of the species, whereas multi target classification (b) and hierarchical multi-label classification (c) consider all of the species in single tree.

On the other hand, the global models are much easier to interpret. The single global model is valid for the complete structured output, i.e., for the whole community of species present in the ecosystem. The global models are able to capture the interactions present between the species, i.e., which species can co-exist at

a locations with given physico-chemical properties. Moreover, the HMC models, as compared to the multi-label models, offer additional information about the higher taxonomic ranks. For example, the HMC model could state that there is a low chance that the species *Diptera chironomus* could be present under the given environmental conditions, however the genus *Diptera* could be.

5 Conclusions

In this article, we address the task of learning predictive models for structured output learning, which takes as input a tuple of attribute values and produces a structured object. In contrast to standard classification and regression, where the output is a single scalar value, in structured output learning the output is a data structure, such as a tuple or a directed acyclic graph. We consider both global and local prediction of structured outputs, the former based on a single model that predicts the entire output structure and the latter based on a collection of models, each predicting a component of the output structure.

We investigate the differences in performance and interpretability of the local and global models. More specifically, we research whether including information in the form of a taxonomic rank helps to improve the predictive performance of the predictive models. We compare the performance of local and global predictive models on a practically relevant task from ecology - habitat modelling.

The results show that the global models perform better than the local models. This performance improvement is more pronounced on domains that have more populated hierarchy. On the other hand, the improvement is less visible on well described and stable domains where any predictive model has good predictive performance. Furthermore, the global models are much easier to interpret than the local models and offer an overview of the complete eco-system.

We plan to extend this work along several dimensions. We will start by including more datasets with different properties in the evaluation procedure. Next, we will generate artificial datasets to further check the results of this study. Finally, we will include other types of local and global models to check whether these findings carry over other predictive modelling methods.

References

1. Kriegel, H.P., Borgwardt, K., Kröger, P., Pryakhin, A., Schubert, M., Zimek, A.: Future trends in data mining. Data Mining and Knowledge Discovery **15** (2007) 87–97
2. Dietterich, T.G., Domingos, P., Getoor, L., Muggleton, S., Tadepalli, P.: Structured machine learning: the next ten years. Machine Learning **73**(1) (2008) 3–23
3. Silla, C., Freitas, A.: A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery **22**(1-2) (2011) 31–72
4. Bakır, G.H., Hofmann, T., Schölkopf, B., Smola, A.J., Taskar, B., Vishwanathan, S.V.N.: Predicting structured data. Neural Information Processing. The MIT Press (2007)

5. Blockeel, H.: Top-down induction of first order logical decision trees. PhD thesis, Katholieke Universiteit Leuven, Leuven, Belgium (1998)
6. Kocev, D., Vens, C., Struyf, J., Džeroski, S.: Tree ensembles for predicting structured outputs. Pattern Recognition **46**(3) (2013) 817–833
7. Vens, C., Struyf, J., Schietgat, L., Džeroski, S., Blockeel, H.: Decision trees for hierarchical multi-label classification. Machine Learning **73**(2) (2008) 185–214
8. Džeroski, S.: Machine learning applications in habitat suitability modeling. In Haupt, S.E., Pasini, A., Marzban, C., eds.: Artificial Intelligence Methods in the Environmental Sciences. Springer (2009) 397–412
9. Demšar, D., Džeroski, S., Larsen, T., Struyf, J., Axelsen, J., Bruns-Pedersen, M., Krogh, P.H.: Using multi-objective classification to model communities of soil. Ecological Modelling **191**(1) (2006) 131–143
10. Džeroski, S., Demšar, D., Grbović, J.: Predicting chemical parameters of river water quality from bioindicator data. Applied Intelligence **13**(1) (2000) 7–17
11. Blockeel, H., Struyf, J.: Efficient algorithms for decision tree cross-validation. Journal of Machine Learning Research **3** (2002) 621–650
12. Breiman, L., Friedman, J., Olshen, R., Stone, C.J.: Classification and Regression Trees. Chapman & Hall/CRC (1984)
13. Slavkov, I., Gjorgjioski, V., Struyf, J., Džeroski, S.: Finding explained groups of time-course gene expression profiles with predictive clustering trees. Molecular BioSystems **6**(4) (2010) 729–740
14. Clare, A.: Machine learning and data mining for yeast functional genomics. PhD thesis, University of Wales Aberystwyth, Aberystwyth, Wales, UK (2003)
15. Ponge, J.F.: Food resources and diets of soil animals in a small area of Scots pine litter. Geoderma **49**(1-2) (1991) 33–62
16. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research **7** (2006) 130