Evaluation of Different Data-derived Label Hierarchies in Multi-label Classification

Gjorgji Madjarov, Tomche Delev, Ivica Dimitrovski, and Dejan Gjorgjevikj

Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering,
Rudgjer Boshkovikj 16, 1000 Skopje, Macedonia
{gjorgji.madjarov, tomche.delev, ivica.dimitrovski,
dejan.gjorgjevikj}@finki.ukim.mk

Abstract. Motivated by an increasing number of new applications, the research community is devoting an increasing amount of attention to the task of multi-label classification (MLC). Many different approaches to solving multi-label classification problems have been recently developed. Recent empirical studies have comprehensively evaluated many of these approaches on many datasets using different evaluation measures. The studies have indicated that the predictive performance and efficiency of the approaches could be improved by using data derived (artificial) hierarchies, in the learning and prediction phases. In this paper, we compare different clustering algorithms for constructing the label hierarchies (in a data-driven manner), in multi-label classification. We consider flat label sets and construct the label hierarchies from the label sets that appear in the annotations of the training data by using four different clustering algorithms (balanced k-means, agglomerative clustering with single and complete linkage and predictive clustering trees). The hierarchies are then used in conjunction with global hierarchical multi-label classification (HMC) approaches.

Keywords: multi-label, hierarchical, classification, clustering

1 Introduction

Multi-label learning is concerned with learning from examples, where each example is associated with multiple labels. Multi-label classification (MLC) has received significant attention in the research community over the past few years, motivated by an increasing number of new applications. The latter include semantic annotation of images and video (news clips, movies clips), functional genomics (predicting gene and protein function), music categorization into emotions, text classification (news articles, web pages, patents, e-mails, bookmarks...), directed marketing and others.

Madjarov et al. [1] presented an extensive experimental evaluation of the most popular methods for multi-label learning using a wide range of evaluation measures on a variety of datasets. In particular, the authors have experimentally evaluated 12 methods using 16 evaluation measures over 11 benchmark

datasets. The results reveal that the best performing methods over all evaluation measures are the Hierarchy Of Multi-label classifiERs (HOMER) [2] and Random Forests of Predictive Clustering Trees for Multi-target Classification (RF-PCTs for MTC) [3], followed by Binary Relevance (BR) [4] and Classifier Chains (CC) [5].

We believe that the better predictive performance and efficiency of the HOMER method as compared to BR and CC, is a result of the data derived (artificial) hierarchy, that HOMER defines over the output space of the original MLC problem first, and then uses it in the learning and prediction phases. In particular, HOMER transforms the (original, flat) multi-label learning task into a hierarchy of (simpler) multi-label learning tasks, based on a hierarchy of labels derived from the data. The hierarchy is obtained by applying an unsupervised (clustering) approach to the label part of the data that comes from the original MLC problem. An example hierarchy of labels (and classifiers) produced for a multi-label classification task with 8 labels $\{\lambda_1, \lambda_2, ..., \lambda_8\}$ is given in Figure 1.

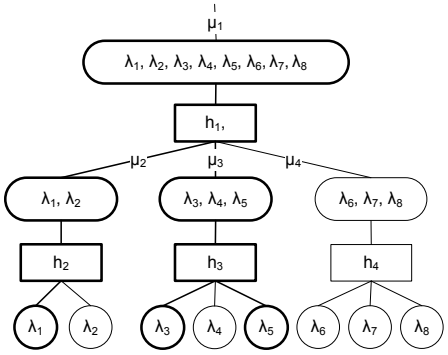

Fig. 1. An example of labels and classifiers considered by HOMER (λ - label, μ - meta-label, h - multi-label classifier).

In this paper, we experimentally evaluate the influence of different data-derived label hierarchies on the predictive performance of multi-label classifiers. Additionally, we confirmed even stronger, that structuring the output space (label part) of a flat MLC problem, and using this structure by a classifier that can directly handle hierarchical multi-label classification (HMC) problems can improve the predictive performance of a classifier that does not use this structure and directly solves the flat MLC problems. More specifically, we derive a hierarchy from the output space of the (original) flat MLC problem using four different clustering approaches first, and then use PCTs for HMC [6] for solving the newly defined hierarchical multi-label classification problem.

To show the improvements that can be achieved by using the data derived structure on the label space and to evaluate the influence of the different data-derived label hierarchies in multi-label classification, we compare: single PCT [6]

for solving classical MLC problems [3], and single PCT for solving HMC problems [7] (both in global settings). The transformation of the (original) flat MLC problem to HMC problem is made by balanced k-means clustering [2], agglomerative clustering with single and complete linkage [8] and clustering performed by predictive clustering trees for multi-target classification (MTP) [6].

The remainder of this paper is organized as follows. Section 2 defines the tasks of multi-label classification, multi-label ranking and hierarchical multi-label classification. The use of data derived label hierarchies in multi-label classification is presented in Section 3. Section 4 describes the multi-label datasets, the evaluation measures and the experimental setup, while Section 5 presents and discusses the experimental results. Finally, the conclusions and directions for further work are presented in Section 6.

2 Background

In this section, we define the task of multi-label classification and the task of hierarchical multi-label classification.

2.1 The task of multi-label classification (MLC)

Multi-label learning is concerned with learning from examples, where each example is associated with multiple labels. These multiple labels belong to a predefined set of labels. We can distinguish two types of tasks: multi-label classification and multi-label ranking.

In the case of multi-label classification, the goal is to construct a predictive model that will provide a list of relevant labels for a given, previously unseen example. On the other hand, the goal of the task of multi-label ranking is to construct a predictive model that will provide, for each unseen example, a list of preferences (i.e., a ranking) on the labels from the set of possible labels.

The task of multi-label learning is defined as follows [9]:
Given:

- An input space $\mathcal{X}$ that consists of vectors of values of primitive data types (nominal or numeric), i.e., $\forall \mathbf{x_i} \in \mathcal{X}, \mathbf{x_i} = (x_{i_1}, x_{i_2}, ..., x_{i_D})$, where D is the size of the vector (or number of descriptive attributes),
- an output space $\mathcal{Y}$ that is defined as a subset of a finite set of disjoint labels $\mathcal{L} = \{\lambda_1, \lambda_2, ..., \lambda_Q\}$ ($Q > 1$ and $\mathcal{Y} \subseteq \mathcal{L}$)
- a set of examples E, where each example is a pair of a vector and a set from the input and output space respectively, i.e., $E = \{(\mathbf{x_i}, \mathcal{Y}_i) | \mathbf{x_i} \in \mathcal{X}, \mathcal{Y}_i \subset \mathcal{L}, 1 \leq i \leq N\}$ where N is the number of examples of E ($N = |E|$), and
- a quality criterion q, which rewards models with high predictive performance and low computational complexity.

If the task at hand is multi-label classification, then the goal is to
Find: a function h: $\mathcal{X} \rightarrow 2^{\mathcal{L}}$ such that h maximizes q.

On the other hand, if the task is multi-label ranking, then the goal is to
Find: a function f: $\mathcal{X} \times \mathcal{L} \rightarrow \mathcal{R}$, such that f maximizes q, where $\mathcal{R}$ is the ranking on the labels for a given example.

An extensive bibliography of learning methods for solving multi-label learning problems can be found in [10] [4] [11] [1].

2.2 The task of hierarchical multi-label classification (HMC)

Hierarchical classification differs from the multi-label classification in the following: the labels are organized in a hierarchy. An example that is labeled with a given label is automatically labeled with all its parent-labels (this is known as the hierarchy constraint). Furthermore, an example can be labeled simultaneously with multiple labels that can follow multiple paths from the root label. This task is called hierarchical multi-label classification (HMC).

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 representing the parent-child relationship ($\forall\ \lambda_1, \lambda_2 \in \mathcal{L} : \lambda_1 \leq_h \lambda_2$ if and only if λ_1 is a parent of λ_2) structured as a tree [9]. 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., $E = \{(\mathbf{x_i}, \mathcal{Y}_i) | \mathbf{x_i} \in \mathcal{X}, \mathcal{Y}_i \subseteq \mathcal{L}, \lambda \in \mathcal{Y}_i \Rightarrow \forall \lambda' \leq_h \lambda : \lambda' \in \mathcal{Y}_i, 1 \leq i \leq N\}$ where N is the number of examples of E ($N = |E|$). The quality criterion q, rewards models with high predictive performance and low complexity as in the task of multi-label classification.

An extensive bibliography of learning methods for hierarchical classification scattered across different application domains is given by [12].

3 The use of data derived label hierarchies in multi-label classification

In this study, we suggest to transform the flat multi-label classification problem into a hierarchical multi-label one and solve it by using an approach for HMC [12]. In particular, one should derive a hierarchy from the label part of the original (flat) multi-label classification problem first, and then use this hierarchy to construct hierarchical classification problem that later solves by using a HMC approach [12].

Table 1 shows an example of a multi-label dataset and its corresponding transformed hierarchical multi-label dataset. The transformation is performed according to the label hierarchy generated by a clustering algorithm that considers only the label part (output space) of the training data. In particular, the third column (*Original label set*) in Table 1 shows the labels of the (original) label space of a multi-label learning dataset with five examples. It is defined over a set of eight labels ($\mathcal{L} = \{\lambda_1, \lambda_2, ..., \lambda_8\}$). The fourth column in the same table (*Hierarchical label set*), shows the corresponding hierarchical label set (for the same dataset), obtained by using the label hierarchy from Figure 1 ($\mathcal{HL} = \{\mu_1, \mu_2, \mu_3, \mu_4, \lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7, \lambda_8\}$). Each example in the HMC

dataset is actually labeled with multiple paths of the hierarchy, defined from the root to the leaves (represented by the relevant labels for the corresponding example in the original MLC dataset).

Table 1. A hierarchical multi-label dataset obtained by transforming an original flat multi-label dataset (the label hierarchy from Figure 1 is used)

Example	Features	Original label set	Hierarchical label set
$\mathbf{x_1}$	$x_{11}, x_{12}, \ldots, x_{1D}$	$\{\boldsymbol{\lambda_1}, \boldsymbol{\lambda_4}, \boldsymbol{\lambda_8}\}$	$\{\mu_1, \mu_2, \mu_3, \mu_4, \boldsymbol{\lambda_1}, \boldsymbol{\lambda_4}, \boldsymbol{\lambda_8}\}$
$\mathbf{x_2}$	$x_{21}, x_{22}, \ldots, x_{2D}$	$\{\boldsymbol{\lambda_3}, \boldsymbol{\lambda_6}\}$	$\{\mu_1, \mu_3, \mu_4, \boldsymbol{\lambda_3}, \boldsymbol{\lambda_6}\}$
$\mathbf{x_3}$	$x_{31}, x_{32}, \ldots, x_{3D}$	$\{\boldsymbol{\lambda_1}\}$	$\{\mu_1, \mu_2, \boldsymbol{\lambda_1}\}$
$\mathbf{x_4}$	$x_{41}, x_{42}, \ldots, x_{4D}$	$\{\boldsymbol{\lambda_2}, \boldsymbol{\lambda_3}, \boldsymbol{\lambda_4}, \boldsymbol{\lambda_8}\}$	$\{\mu_1, \mu_2, \mu_3, \mu_4, \boldsymbol{\lambda_2}, \boldsymbol{\lambda_3}, \boldsymbol{\lambda_4}, \boldsymbol{\lambda_8}\}$
$\mathbf{x_5}$	$x_{51}, x_{52}, \ldots, x_{5D}$	$\{\boldsymbol{\lambda_1}, \boldsymbol{\lambda_4}, \boldsymbol{\lambda_7}\}$	$\{\mu_1, \mu_2, \mu_3, \mu_4, \boldsymbol{\lambda_1}, \boldsymbol{\lambda_4}, \boldsymbol{\lambda_7}\}$

3.1 Generating a label hierarchy on a multi-label output space

The process of generating label hierarchies on a multi-label output space is critical for the good performance of the HMC methods on the transformed problems. When we build the hierarchy over the label space, there is only one constraint that we should take care of: the original MLC task should be defined by the leaves of the label hierarchy. In particular, the labels from the original MLC problem represent the leaves of the tree hierarchy (Figure 1), while the labels that represent the internal nodes of the tree hierarchy are so-called meta-labels (that model the correlation among the original labels).

In this study, we use four different clustering approaches (two divisive and two agglomerative) for deriving the hierarchy on the output space of the (original) MLC problem:

- balanced k-means clustering approach [2] (divisive approach),
- predictive clustering trees [6] (divisive approach),
- agglomerative clustering by using complete linkage [8], and
- agglomerative clustering by using single linkage [8].

Balanced k-means creates the label hierarchy by partitioning the original labels recursively in a top-down depth-first fashion. The top node of the hierarchy contains all labels. At each node n, $k <= |\mathcal{L}_n|$ child nodes are created. The labels of the current node are distributed (divided) using a clustering method into k disjoint subsets (k meta-labels) with an explicit constraint on the size of each subset, one for each child of the current node.

In this work, we use a specific setting from the predictive clustering framework as in [13] [3], where the target space is equal to the descriptive space, i.e., the descriptive variables are used to provide descriptions for the obtained clusters. This focuses the predictive clustering setting on the task of clustering instead of classification.

Agglomerative clustering algorithms treat each example as a singleton cluster at the outset and then successively merge pairs of clusters until all clusters have been merged into a single cluster that contains all examples.

The predictive clustering trees and the agglomerative approaches produce binary tree hierarchies, while the balanced k-means clustering approach produces multi-branch tree hierarchies for $k > 2$.

3.2 Solving MLC problems by using classification approaches for HMC

After the transformation of the original MLC problem into a HMC one, the new HMC problem can be solved by a hierarchical multi-label learning approach. The transformed hierarchical multi-label dataset satisfies the hierarchy constraint (an example that is labeled with a given label is automatically labeled with all its parent-labels).

Figure 2 presents the pseudo-code of the algorithm for solving a MLC problem by using data-derived label hierarchies and a classification approach for HMC. The algorithm first defines the hierarchy, then solves the HMC problem by using a classification approach for HMC. It finally extracts the predictions for the leaves of the hierarchy (that are actually the predictions for the original labels) and evaluates the performance.

E^{train} and E^{test} denote the training and testing examples, while $\mathbf{W}^{train}$ is only the label part (label data) of the training set. Using the label hierarchy derived from the label data, $\mathbf{W}^{train}$ is transformed into new hierarchically organized label data $\mathbf{W}_H^{train}$. E_H^{train} and E_H^{test} denote the corresponding hierarchical multi-label datasets obtained by transforming the original (flat) multi-label datasets (E^{train} and E^{test}) into hierarchical form.

P_H denotes the predictions for the examples of the hierarchical multi-label dataset E_H^{test}, while P denotes the predictions for the original labels. The latter are obtained by extracting the probabilities in the leaves of the label tree from the predictions P_H. The predictions P_H are represented as vectors of probabilities (one vector for one example), where each probability is associated to only one label from the hierarchy (meta-label representing an internal node or original label representing a leaf). Predictions P in the original multi-label scenario can be obtained by using different approaches for transforming the hierarchical multi-label predictions P_H. In this work, we use the simplest approach: only the probabilities for the leaves from the hierarchical predictions P_H are evaluated, while the other probabilities (for the meta-labels) are simply ignored.

3.3 Classification approaches for HMC

Based on the existing literature, Silla et al. [12] propose a unifying framework for hierarchical classification, including a taxonomy of hierarchical classification problems and methods. One of the dimensions along which the hierarchical classification methods differ is the way of using (exploring) the hierarchical label

procedure MLCToHMC(E^{train} ,E^{test}) returns performance

1: $\mathbf{W}^{train}$ = ExtractLabelSet(E^{train});
2: $\mathbf{W}_H^{train}$ = DefineHierarchy($\mathbf{W}^{train}$);
3:
4: *//transform multi-label dataset to hierarchical multi-label one*
5: E_H^{train} = MLCToHMCTrainDataset(E^{train}, $\mathbf{W}_H^{train}$);
6: E_H^{test} = MLCToHMCTestDataset(E^{test}, $\mathbf{W}_H^{train}$);
7:
8: *//solve transformed hierarchical multi-label problem*
9: *//by using approach for HMC*
10: HMCModel = HMCMetod(E_H^{train});
11:
12: *//generate HMC predictions*
13: P_H = HMCModel(E_H^{test});
14:
15: *//Extract predictions only for the leaves from the HMC predictions* P_H
16: P = ExtractLeavesPredictionsFromHMCPredictions(P_H, $\mathbf{W}_H^{train}$, $\mathbf{W}^{train}$);
17: **return** EvaluatePredictions(P);

Fig. 2. Solving flat MLC problems by using classification approaches for HMC.

structure in the learning and prediction phases. They reviewed two different approaches that utilize the hierarchy: the top-down (or local) approach that uses local information to create a set of local classifiers and the global (or big-bang) approach.

The recent research show that learning a single global model for all labels (in the hierarchy) can have some advantages [3] [14] over the local approaches. The total size of the global classification model is typically smaller as compared to the total size of all the local models learned by local classifier approaches. Also, in the global classifier approach, a single classification model is built from the training set, taking into account the label hierarchy and relationships. During the prediction phase, each test example is classified using the induced model, in a process that can assign labels to a test example at potentially every level of the hierarchy. Because of that, in this study we compare PCTs for MTP (as flat, global MLC approach) and PCTs for HMC (in a global setting) [3].

4 Experimental design

4.1 Datasets and evaluation measures

We use four multi-label classification benchmark problems used in previous studies and evaluations of methods for multi-label learning. Table 2 presents the basic statistics of the datasets. The datasets come from the domain of text categorization and pre-divided into training and testing parts as used by other researchers.

In any multi-label experiment, it is essential to include multiple and contrasting measures because of the additional degrees of freedom that the multi-label

Table 2. Description of the benchmark problems in terms of number of training (*#tr.e.*) and test (*#t.e.*) examples, number of features (D), total number of labels (Q) and label cardinality - average number of labels per example (l_c).

	Reference	*#tr.e.*	*#t.e.*	D	Q	l_c
tmc2007	[15]	21519	7077	500	22	2.16
bibtex	[16]	4880	2515	1836	159	2.40
bookmarks	[16]	60000	27856	2150	208	2.03
delicious	[2]	12920	3185	500	983	19.02

setting introduces. In our experiments, we used various evaluation measures that have been suggested by [11] In particular, we used 12 *bipartitions-based* evaluation measures: six *example-based* evaluation measures (*hamming loss*, *accuracy*, *precision*, *recall*, *F measure* and *subset accuracy*) and six *label-based* evaluation measures (*micro precision*, *micro recall*, *micro* F_1, *macro precision*, *macro recall* and *macro* F_1). Note that these evaluation measures require predictions stating that a given label is present or not (binary 1/0 predictions). However, most predictive models predict a numerical value for each label and the label is predicted as present if that numerical value exceeds some pre-defined threshold τ. The performance of the predictive model thus directly depends on the selection of an appropriate value of τ.

Also, we used four *ranking-based* evaluation measures (*one-error*, *coverage*, *ranking loss* and *average precision*) that compare the predicted ranking of the labels with the ground truth ranking. A detailed description of the evaluation measures can be found in [1].

4.2 Experimental setup

The comparison of the multi-label learning methods was performed using the CLUS[1] system for predictive clustering. All experiments were performed on a server with an Intel Xeon processor at 2.5GHz and 64GB of RAM with the Fedora 14 operating system. We used the default settings of CLUS to learn the single PCT approaches (PCTs for MTP - as flat MLC approach, and PCTs for HMC). The threshold τ for the *bipartitions-based* evaluation measures was set to 0.5 for all compared methods.

The balanced k-means clustering method requires to be configured the number of clusters k in each node of the hierarchy. For this parameter, five different values (2-6) were considered in the cross-validation phase [2]. After determining the best value of k on every dataset (via cross-validation on the training dataset), the PCT for HMC was trained using all available training examples and was evaluated by recognizing all test examples from the corresponding dataset. The values of the parameter k are 3 for the tmc2007, bibtex and bookmarks

[1] http://clus.sourceforge.net

datasets, and 4 for the delicious dataset. Also, for the balanced k-means and the agglomerative methods, Euclidean distance was used as a distance measure.

5 Results and discussion

In this section, we present the results from the experimental evaluation. Table 3 shows the predictive performance of the compared methods:

- PCTs for MTP, that don't use a hierarchy for solving the original MLC problem (labeled as *no hierarchy (flat MLC)*)
- PCTs for HMC, that use data-derived label hierarchies, defined by:
  - balanced k-means clustering approach (labeled as *balanced-k-means*)
  - agglomerative clustering by using complete linkage (labeled as *agglomerative (complete)*)
  - agglomerative clustering by using single linkage (labeled as *agglomerative (single)*)
  - predictive clustering trees (labeled as *PCTs*)

The first column of the table describes the methods used for defining the hierarchies, while the other columns show the predictive performance of the compared methods and hierarchies in terms of the 16 performance evaluation measures. The best results per dataset are shown in boldface.

Inspecting Table 3, we note that PCTs for HMC outperform PCTs for MLC on all datasets and on almost all evaluation measures. The instantiation of PCTs for MTP (for solving flat multi-label classification problems) shows better predictive performance only on *micro precision* evaluation measure on the *bibtex* and *bookmarks* datasets.

PCTs for HMC that use balanced k-means clustering for deriving the label hierarchies outperform PCTs for HMC that use agglomerative clustering with single and complete linkage and PCTs for deriving the label hierarchies on datasets with higher number of labels (*bibtex*, *bookmarks* and *delicious*). PCTs for HMC with agglomerative clustering perform the best on *tmc2007* dataset. The two agglomerative clustering methods (single and complete linkage) derived identical label hierarchies on all MLC problems, which result in same predictive performance in the experimental evaluation.

The highest improvement of utilizing the data-derived hierarchies is obtained on *delicious* dataset, as a result of the largest number of labels and the largest label cardinality (average number of labels per example). A large number of labels and large label cardinality yields a larger hierarchy that emphasizes the relations between labels, and improves the process of learning and prediction.

Table 3. The predictive performances of PCTs for MLC obtained on the original (flat) MLC problems and PCTs for HMC obtained on the transformed (newly) defined HMC problems by using four different clustering approaches (balanced k-means, predictive clustering trees, and agglomerative clustering with complete and single linkage) along 16 performance evaluation measures.

	HammingLoss	Accuracy	Precision	Recall	Fmeasure	SubsetAccuracy	MicroPrecision	MicroRecall	MicroF1	MacroPrecision	MacroRecall	MacroF1	OneError	Coverage	RankingLoss	AvgPrecision
tmc2007																
no hierarchy (flat MLC)	0.075	0.436	0.659	0.478	0.554	0.215	0.689	0.454	0.547	0.386	0.235	0.263	0.307	4.57	0.100	0.700
balanced-k-means - HMC	**0.067**	0.515	0.688	0.604	0.643	**0.253**	0.704	0.563	0.625	**0.735**	0.341	0.409	0.246	**3.35**	0.066	0.774
agglomerative (complete) - HMC	0.068	0.501	0.699	0.571	0.628	0.250	0.717	0.524	0.605	0.629	0.283	0.344	0.247	3.54	0.071	0.767
agglomerative (single) - HMC	0.068	0.501	0.699	0.571	0.628	0.250	0.717	0.524	0.605	0.629	0.283	0.344	0.247	3.54	0.071	0.767
PCTs - HMC	0.101	**0.559**	**0.746**	**0.703**	**0.723**	0.184	**0.742**	**0.625**	**0.678**	0.675	**0.358**	**0.418**	**0.084**	11.64	**0.055**	**0.835**
bibtex																
no hierarchy (flat MLC)	**0.014**	0.046	0.140	0.046	0.069	0.004	**1.000**	0.057	0.108	0.006	0.006	0.006	0.783	58.60	0.256	0.212
balanced-k-means - HMC	0.015	**0.243**	**0.368**	**0.290**	**0.324**	0.113	0.550	**0.259**	**0.352**	**0.296**	**0.174**	**0.202**	**0.449**	**30.36**	**0.105**	**0.491**
agglomerative (complete) - HMC	**0.014**	0.175	0.289	0.183	0.225	0.103	0.749	0.145	0.243	0.079	0.044	0.052	0.589	45.74	0.190	0.341
agglomerative (single) - HMC	**0.014**	0.175	0.289	0.183	0.225	0.103	0.749	0.145	0.243	0.079	0.044	0.052	0.589	45.74	0.190	0.341
PCTs - HMC	**0.014**	0.197	0.328	0.204	0.251	**0.117**	0.796	0.161	0.268	0.082	0.056	0.062	0.541	36.93	0.152	0.388
bookmarks																
no hierarchy (flat MLC)	**0.009**	0.133	0.133	0.137	0.135	0.129	**0.947**	0.076	0.141	0.018	0.016	0.017	0.817	73.78	0.258	0.213
balanced-k-means - HMC	**0.009**	**0.205**	**0.224**	**0.211**	**0.217**	**0.188**	0.776	**0.139**	**0.236**	**0.299**	**0.071**	**0.097**	**0.651**	**50.46**	**0.169**	**0.370**
agglomerative (complete) - HMC	**0.009**	0.160	0.163	0.165	0.164	0.153	0.875	0.097	0.175	0.103	0.026	0.030	0.729	57.99	0.200	0.302
agglomerative (single) - HMC	**0.009**	0.160	0.163	0.165	0.164	0.153	0.875	0.097	0.175	0.103	0.026	0.030	0.729	57.99	0.200	0.302
PCTs - HMC	**0.009**	0.177	0.185	0.181	0.183	0.167	0.846	0.110	0.195	0.116	0.036	0.044	0.699	56.31	0.193	0.328
delicious																
no hierarchy (flat MLC)	0.019	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.592	691.62	0.172	0.206
balanced-k-means - HMC	**0.018**	**0.118**	**0.429**	**0.132**	**0.201**	**0.007**	**0.621**	**0.120**	**0.201**	**0.162**	**0.049**	**0.062**	**0.386**	**548.01**	**0.121**	**0.336**
agglomerative (complete) - HMC	0.019	0.074	0.354	0.081	0.132	0.003	0.590	0.077	0.136	0.064	0.018	0.022	0.440	558.78	0.131	0.293
agglomerative (single) - HMC	0.019	0.074	0.354	0.081	0.132	0.003	0.590	0.077	0.136	0.064	0.018	0.022	0.440	558.78	0.131	0.293
PCTs - HMC	0.019	0.097	0.376	0.107	0.167	0.002	0.609	0.101	0.173	0.066	0.029	0.034	0.418	553.65	0.128	0.316

6 Conclusions and further work

In this paper, we have investigated the use of label hierarchies in multi-label classification, constructed in a data-driven manner. We consider flat label-sets and construct label hierarchies from the label sets that appear in the annotations of the training data by using clustering approaches based on balanced k-means clustering, agglomerative clustering with single and complete linkage, and clustering performed by PCTs. The hierarchies are then used in conjunction with hierarchical multi-label classification approaches in the hope of achieving better multi-label classification.

In particular, we investigate and evaluate the utility of four different data-derived label hierarchies in the context of predictive clustering trees for HMC. The experimental results clearly show that the use of the hierarchy results in improved performance and the more balanced hierarchy offers better representation of the label relationships.

The label hierarchies used in PCTs for HMC greatly improve the performance of PCTs for MTP (as used for MLC): The results show improvement in performance on almost all evaluation measures considered. Multi-branch hierarchy (defined by balanced k-means clustering) outperforms binary hierarchies (defined by agglomerative clustering with single and complete linkage and PCTs) on datasets with higher number of labels (*bibtex*, *bookmarks* and *delicious*). This improvement is especially emphasized on the *delicious* dataset, as a result of the higher label cardinality that this dataset has in comparison to the other evaluated datasets.

The final recommendation considering the performance of the evaluated methods is that we should use data-derived label hierarchies. We should transform the original (flat) multi-label classification problem into hierarchical multi-label one by using more balanced hierarchies, and solve the newly defined hierarchical classification problem by a classifier that can directly handle HMC problems.

We plan to extend this study by using more multi-label classification datasets, in particular more diverse ones. These would include different numbers of possible labels, different numbers of labels per example and different joint distribution properties for the labels (e.g., different degrees of (in)dependence among the labels). This would allow us to draw stronger conclusions on the conditions under which the use of a hierarchy on the label space and the way of its construction improves the performance of the different MLC approaches.

A final direction for further work might be the comparison of hierarchies constructed by humans and hierarchies generated in a data-driven fashion. For HMC problems, we can consider the MLC task defined by the leaves of the provided label hierarchy. We can then construct label hierarchies automatically, as described above, and compare these hierarchies (and their utility) to the originally provided label hierarchy.

Acknowledgements

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).

References

1. Madjarov, G., Kocev, D., Gjorgjevikj, D., Dzeroski, S.: An extensive experimental comparison of methods for multi-label learning. Pattern Recognition **45**(9) (2012) 3084 – 3104
2. Tsoumakas, G., Katakis, I., Vlahavas, I.: Effective and Efficient Multilabel Classification in Domains with Large Number of Labels. In: Proc. of the ECML/PKDD Workshop on Mining Multidimensional Data. (2008) 30–44
3. Kocev, D.: Ensembles for predicting structured outputs. PhD thesis, IPS Jožef Stefan, Ljubljana, Slovenia (2011)
4. Tsoumakas, G., Katakis, I.: Multi Label Classification: An Overview. International Journal of Data Warehouse and Mining **3**(3) (2007) 1–13
5. Mencía, E.L., Park, S.H., Fürnkranz, J.: Efficient voting prediction for pairwise multilabel classification. Neurocomputing **73** (2010) 1164–1176
6. Blockeel, H., Raedt, L.D., Ramon, J.: Top-down induction of clustering trees. In: Proc. of the 15th International Conference on Machine Learning. (1998) 55–63
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. Manning, C.D., Raghavan, P., Schütze, H.: An Introduction to Information Retrieval. Cambridge University Press (2009)
9. Kocev, D., Vens, C., Struyf, J., Džeroski, S.: Tree ensembles for predicting structured outputs. Pattern Recognition **46**(3) (2013) 817–833
10. de Carvalho, A.C.P.L.F., Freitas, A.A.: A tutorial on multi-label classification techniques. In Abraham, A., Hassanien, A.E., Snsel, V., eds.: Foundations of Computational Intelligence (5). Volume 205 of Studies in Computational Intelligence. Springer (2009) 177–195
11. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. In: Data Mining and Knowledge Discovery Handbook. Springer Berlin / Heidelberg (2010) 667–685
12. Silla, CarlosN., J., Freitas, A.: A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery **22** (2011) 31–72
13. Dimitrovski, I., Kocev, D., Loskovska, S., Deroski, S.: Fast and scalable image retrieval using predictive clustering trees. In: Discovery Science. Lecture Notes in Computer Science. Springer Berlin Heidelberg (2013) 33–48
14. Levatić, J., Kocev, D., Džeroski, S.: The use of the label hierarchy in HMC improves performance: a case study in predicting community structure in ecology. In: Proc. of the Workshop on New frontiers in mining complex patterns held in conjunction with ECML/PKDD2013. (2013) 189–201
15. Srivastava, A., Zane-Ulman, B.: Discovering recurring anomalies in text reports regarding complex space systems. In: Proc. of the IEEE Aerospace Conference. (2005) 55–63
16. Katakis, I., Tsoumakas, G., Vlahavas, I.: Multilabel Text Classification for Automated Tag Suggestion. In: Proc. of the ECML/PKDD Discovery Challenge. (2008)