Experimental Comparison of Methods for Multi-Label Classification in Different Application Domains

Passent El Kafrawy
Faculty of Science
Menofia University, Egypt

Amr Mausad
Faculty of Science
Menofia University, Egypt

Heba Esmail
Faculty of Science
Menofia University, Egypt

ABSTRACT
Real-world applications have begun to adopt the multi-label paradigm. The multi-label classification implies an extra dimension because each example might be associated with multiple labels (different possible classes), as opposed to a single class or label (binary, multi-class) classification. And with increasing number of possible multi-label applications in most ecosystems, there is little effort in comparing the different multi-label methods in different domains. Hence, there is need for a comprehensive overview of methods and metrics. In this study, we experimentally evaluate 11 methods for multi-label learning using 6 evaluation measures over seven benchmark datasets. The results of the experimental comparison revealed that the best performing method for both the example- based evaluation measures and the label-based evaluation measures are ECC on all measures when using C4.5 tree classifier as a single-label base learner.

Keywords
Multi-Label classification, Multi-Label learning, Data Mining

1. INTRODUCTION
In Machine Learning, and particularly in supervised learning, one of the most important applications is classification, where each example (item) in the problem domain (dataset) is associated with an attribute vector, which represents data possible range of values from its domain. Labels represent concepts dependent on the problem domain that can belong to. When each example is associated with a single class (label) this is known as single label classification, while each example might be associated with multiple labels, this is known as multi-label classification. Although single-label classification is considered the standard task in multi-class problems, multi-label classification is viable in most ecosystems. At the same time, multi-label classification is by no means easy or intuitive.

Multi-label classification problems appear in a wide range of real world situations and applications. A good example is how Gmail has replaced the old "folder" metaphor with labels. Many online news sites, for example the BBC, often link to the same news article from different category headings, i.e. a multi-label association. A multitude of other sources have also, knowingly or not, embraced the multi-label context. Domains such as microbiology or medicine often inherently require a multi-label scheme: a single gene may influence the production of more than one protein, and a patient's symptoms may be linked to multiple ailments. This explains the explosion of interest in multi-label classification in the academic literature over recent years.

According to Tsoumakas and Katakis [1], multi-label classification groups methods into two main categories: problem transformation and algorithm adaptation. The task of multi-label classification with problem transformation, mean to have a multi-label problem transformed into one or more single-label problems. This scheme allows us to addressing the problem of multi-label classification using algorithms that are not designed for the specificities of the task. This is opposed to algorithm adaptation, where a specific classifier is modified to carry out multi-label classification; often highly suited to specific domains or contexts but not as flexible and has a high computational complexity. In this study, extend this categorization of multi-label methods with a third group of methods, namely, ensemble methods. This group of methods consists of methods that use ensembles to make multi-label predictions and their base classifiers belong to either problem transformation or algorithm adaptation methods. Each approach brings benefits but also has disadvantages that it is necessary to know in order to choose the best option.

Although there are a reasonable number of multi-label classification methods proposed in the literature, and with increasing number of possible multi-label applications in different domains there is little effort in comparing the different multi-label methods in different applications. Hence, there is need for a comprehensive overview of methods and metrics. There is a strong need for a wider, extensive, and unbiased experimental comparison of multi-label learning methods. This paper addresses this need and analyses multi-label learning methodologies.

In this paper, we will experimentally evaluate 11 methods for multi-label learning using 6 evaluation measures over 7 bench-mark datasets. The multi-label methods comprise two algorithm adaptation methods, six problem transformation methods and three ensemble methods. The large number of methods, datasets and evaluation measures enables to draw some general conclusions and performs an unbiased assessment of the predictive performance of the multi-label methods.

This paper is organized as follows. Section 2 defines the tasks of multi-label classification. The state-of-the-art methods for multi-label classification used in the experimental evaluation and multi-label evaluation measures are presented in Section 3and Section 4. Section 5 describes the experimental setup, while Section 6 presents and discusses the experimental results. Finally, the conclusions are given in Section 7.

2. MULTI-LABEL CLASSIFICATION
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. Depending on the goal, 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 in 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) of the labels from the set of possible labels. Our learning model is the following standard extension of the binary case.

Assume that an instance x ∈X can be associated with a subset of labels y, which is referred to as the relevant set of labels for x., for which we use subset notation y ⊆ [L] or vector notation

$y \in \{0, 1\}^L$, as dictated by convenience. Assume Y is a given set of predefined binary labels $Y = \{\lambda_1, \ldots, \lambda_L\}$. For a given set of labeled examples $D = \{x_1, x_2, \ldots, x_n\}$ the goal of the learning process is to find a classifier $h : X \rightarrow Y$, which maps an object $x \in X$ to a set of its classification labels $y \in Y$, such that $h(x) \subseteq \{\lambda_1, \ldots, \lambda_L\}$ for all x in X.

Multi-label classification exhibits several challenges not present in the binary case. The labels may be interdependent, so that the presence of a certain label affects the probability of other labels' presence. Thus, exploiting dependencies among the labels could be beneficial for the classifier's predictive performance

3. METHODS FOR MULTI-LABEL CLASSIFICATION

In this section, present the three categories of methods for multi-label learning: algorithm adaptation, problem transformation and ensemble methods, and discuss the advantages and disadvantages of each method.

3.1 Problem Transformation Method

The problem transformation methods are multi-label learning methods that transform the multi-label learning problem into one or more single-label classification or regression problems. For smaller single-label problems, there exists a plethora of machine learning algorithms. Problem transformation methods can be grouped into three main categories: binary relevance [1], label power-set [1,2,3] and pair-wise [4,5] methods.

- ***The Binary Relevance method (BR),*** BR also known as *one-against-all* (OAA) transforms any multi-label problem into L binary problems. Each binary classifier is then responsible for predicting the association of a single label (one binary problem for each label). Although conceptually simple and relatively fast, it is widely recognized that BR does not explicitly model label correlations. The main contribution of BR the classifier chains method (CC) proposed by Read et al. [6], which overcomes the label independence assumption of BR. The (CC) fixes a particular order of the BR base classifiers and subsequently adds the outputs of the preceding classifiers as new features. Predictive performance depends on the order of the label indices in label-set vector

- ***The label power-set method (LP),*** is a simple and less common problem transformation method. LP treats all label sets as atomic (single) labels to form a single-label problem in which the set of single labels represents all distinct label sets in the multi-label training data. The meta problem is then solved with a normal multiclass algorithm. Although LC can take into account label correlations directly, but the space of possible label subsets can be very large. To resolve this issue, in [7] introduced the pruned sets method, for multi-label classification is centered on the concept of treating sets of labels as single labels. This allows the classification process to inherently take into account correlations between labels. By pruning these sets, PS focuses only on the most important correlations, which reduces complexity and improves accuracy. Another label power-set method is HOMER [8] is an algorithm for effective and computationally efficient multi-label learning in domains with a large number of labels. The (HOMER) organizes all labels into a tree-shaped hierarchy with a much smaller set of labels at each node. A multi-label classifier is then constructed at each non-leaf node, following the BR approach.

- ***Pair-wise methods:*** A third problem transformation approach to solving the multi-label learning problem is pair-wise or round robin classification with binary classifiers. The basic idea here is to use $\frac{n(n-1)}{2}$ classifiers covering all pairs of labels. Each classifier is trained using the samples of the first label as positive examples and the samples of the second label as negative examples. To combine these classifiers, the pair-wise classification method naturally adopts the majority voting algorithm. Given a test example, each classifier predicts (i.e.,votes for) one of the two labels. After the evaluation of all $\frac{n(n-1)}{2}$ classifiers, the labels are ordered according to their sum of votes. A label ranking algorithm is then used to predict the relevant labels for each example. Besides majority voting in CLR, Park et al. [9] propose a more effective voting algorithm. It computes the class with the highest accumulated voting mass, while avoiding the evaluation of all possible pair-wise classifiers. Mencia et al. [10] adapted the QWeighted approach to multi-label learning (QWML).

3.2 Algorithm adaptation methods

The focus of the algorithm adaptation approach aims to modify existing algorithms so that they can deal with multi-label samples, without requiring any preprocessing. In recent years the number of proposals published in this regard has increased strikingly. So here we just list the more remarkable ones.

- ***Decision trees:*** Multi-Label C4.5 (ML-C4.5) [11] is an adaptation of the well known C4.5 algorithm. The learning process is accomplished by allowing multiple labels in the leaves of the tree, the formula for calculating entropy is modified for solving multi- label problems. The modified entropy sums the entropies for each individual class label. The key property of ML-C4.5 is its computational efficiency:

$$\text{Entropy (E)} = - \sum_{i=1}^{N} (p(c_i)\log p(c_i) + q(c_i)\log q(c_i))$$

where E is the set of examples, $p(c_i)$ is the relative frequency of class label c_i and $q(c_i) = 1\_p(c_i)$.

- ***Neural Network based:*** In principle, traditional back-propagation (BP) neural networks can deal with multi-label classification directly through assigning many ones at output layer. In BP-MLL [12], a new empirical loss function is induced from the ranking loss to characterize correlations between labels of an instance. But it is needed to find an additional threshold function using linear regression. It has been shown that this BP method runs very slowly.

- ***Tree Based Boosting:*** ADABOOST.MH and ADABOOST.MR [13] are two extensions of ADABOOST for multi-label data. While AdaBoost.MH is designed to minimize Hamming loss, ADABOOST.MR is designed to find a hypothesis which ranks the correct labels at the top. Furthermore, ADABOOST.MH can also be combined with an algorithm for producing alternating decision trees [14]. The resulting multi-label models of this combination

can be interpreted by humans.

- ***Lazy Learning****:* There are several methods exists based on lazy learning (i.e. k-Nearest Neighbour (kNN)) algorithm. All these methods are retrieving k-nearest examples as a first step. ML-kNN [15] is the extension of popular kNN to deal with multi-label data. It uses the maximum a posteriori principle in order to determine the label set of the test instance, based on prior and posterior probabilities for the frequency of each label within the k nearest neighbors.

- ***Support vector machines****:* Elisseeff and Weston [16] have proposed a ranking approach for multi-label learning that is based on SVMs. The cost function they use is the average fraction of incorrectly ordered pairs of labels.

3.3 Ensemble methods

The ensemble methods for multi-label classification are developed on top of the common problem transformation or algorithm adaptation methods.

- ***Ensembles of classifier chains (ECC)*** *[6],* are an ensemble multi- label classification technique that uses classifier chains as a base classifier. ECC trains m CC classifiers C1,C2, . . . ,Cm. using a standard bagging scheme, where the binary models of each chain are ordered according to a random seed. Each model is then likely to be unique and can predict different label sets from other models. These predictions are summed per label so that each label receives a number of votes. A threshold is used to select the most popular labels that form the final predicted multi-label set.

- ***The random k-label sets (RAkEL)*** *[2],* constructs an ensemble of LP classifiers. It works as follows: It randomly breaks a large set of labels into a number (n) of subsets of small size (k), called k-label sets. For each of them train a multi-label classifier using the LP method. Thus it takes label correlation into account and also avoids LP's problems. Given a new instance, it query models and average their decisions per label. And also uses thresholding to obtain the final model.

- ***Ensembles of Pruned Sets (EPS)*** *[7]* combine pruned sets in an ensemble scheme. PS is particularly suited to an ensemble due to its fast build times and, additionally, the ensemble counters any over-fitting effects of the pruning process and allows the creation of new label sets at classification time.

4. MULTI-LABEL EVALUATION MEASURES

There is no generally accepted procedure for evaluating multi-label classifications. Therefore, several measures from multiclass classification and from information retrieval were adopted and adapted in order to measure multi-label effectively. In experiments, we used various evaluation measures that have been suggested by Tsoumakasetal. [17]. the evaluation measures of predictive performance are divided into two groups: bipartitions-based and rankings-based. The bipartitions-based evaluation measures are calculated based on the comparison of the predicted relevant labels with the ground truth relevant labels. This group of evaluation measures is further divided into example-based and label-based. The ranking-based evaluation measures compare the predicted ranking of the labels with the ground truth ranking.

In the definitions below, y_i denotes the set of true labels of example x_i and $h(x_i)$ denotes the set of predicted labels for the same examples. All definitions refer to the multi- label setting.

4.1 Example based measures

Hamming loss:

evaluates how many times an example-label pair is misclassified, i.e., label not belonging to the example is predicted or a label belonging to the example is not predicted. The smaller the value of hamming loss (h), the better the performance. The performance is perfect when hamming _loss (h) = 0. This metric is defined as

$$\text{Hamming\_loss(h)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{Q}|h(xi)\Delta yi|$$

where Δ stands for the symmetric difference between two sets, N is the number of examples and Q is the total number of possible class labels.

Accuracy:

computes the percentage of correctly predicted labels among all predicted and true labels. Accuracy averaged over all dataset examples is defined as follows:

$$\text{Accuracy(h)} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{h(xi)\cap yi}{h(xi)\cup yi}\right|$$

Accuracy seems to be a more balanced measure and better indicator of an actual algorithm's predictive performance for most standard classification problems than Hamming loss.

Precision:

$$\text{Precision(h)} = \frac{1}{N}\sum_{i=1}^{N}\frac{|h(xi)\cap yi|}{|yi\ |}$$

Recall:

$$\text{Recall(h)} = \frac{1}{N}\sum_{i=1}^{N}\frac{|h(xi)\cap yi|}{|h(xi)|}$$

F1 score:

is the harmonic mean between precision and recall and is defined as

$$F_1 = \frac{1}{N}\sum_{i=1}^{N}\frac{2\times|h(xi)\cap yi|}{|h(xi)|+|yi\ |}$$

F1 is an example based metric and its value is an average over all examples in the dataset. F_1 reaches its best value at 1and worst score at 0.

Subset accuracy**:**

$$\text{Subset\_accuracy(h)} = \frac{1}{N}\sum_{i=1}^{N} I(h(xi) = yi)$$

where I(true)=1 and I(false)=0. It should be noted that subset accuracy is a very strict measure since it requires the predicted set of labels to be an exact match of the true set of labels, and equally penalizes predictions that may be almost correct or totally wrong.

4.2 Label based measures

Macro-precision:

$$\text{Macro\_precision} = = \frac{1}{Q}\sum_{j=1}^{Q}\frac{tpj}{tpj+fpj}$$

where tp_j and fp_j are the number of true positives and false positives for the label λ_j considered as a binary class.

Macro-recall:

$$\text{Macro\_recall} = \frac{1}{Q}\sum_{j=1}^{Q}\frac{tpj}{tpj+fnj}$$

where tp_j and fp_j are defined as for the macro-precision and fnj is the number of false negatives for the label λ_j considered as a binary class.

Macro-F1:
is the harmonic mean between precision and recall, where the average is calculated per label and then averaged across all labels. If p_j and r_j are the precision and recall for all $\lambda_j \in h(x_i)$ from $\lambda_j \in y_i$, the macro-F1 is

$$\text{Macro-F1}= \frac{1}{Q} \sum_{j=1}^{Q} \frac{2 \times pj \times rj}{pj + rj}$$

Micro-precision
(precision averaged over all the example/label pairs) is defined as:

$$\text{Micro\_precision}= \frac{\sum_{j=1}^{Q} tpj}{\sum_{j=1}^{Q} tpj + \sum_{j=1}^{Q} fpj}$$

where tpj, fpj are defined as for macro-precision.

Micro-recall
(recall averaged over all the example/label pairs) is defined as

$$\text{Micro\_recall}= \frac{\sum_{j=1}^{Q} tpj}{\sum_{j=1}^{Q} tpj + \sum_{j=1}^{Q} fnj}$$

where tpj and fnj are defined as for macro-recall.

Micro-F1:
is the harmonic mean between micro-precision and micor-recall. Micro-f1 is defined as:

$$\text{Micro\_f1}= \frac{2 * micro\_precision * micro\_recall}{micro\_precision + micro\_recall}$$

5. EXPERIMENTAL WORK

5.1 Datasets

Multi-label classification problems appear in a wide range of real world situations and applications. The datasets that are included in the experimental setups throughout this work cover the main three application areas in which multi-labeled data is frequently observed: *text categorization*, *multimedia classification* and *bioinformatics*. All datasets were mainly retrieved from the repository" http://mulan.sourceforge.net/datasets.html" of the Mulan Java Library for Multi-Label learning. Table1 summarizes the main properties, which are as following:

Table 1 Datasets used in the experiment: information and statistics

Datasets	Domain	Instances	Attributes	labels	L_{CARD}	L_{DC}
Scene	image	2407	294	6	1.074	15
Emotions	Music	593	72	6	1.869	27
Yeast	*biology*	4417	72	14	4.237	198
Birds	*audio*	645	260	19	1.014	133
Genbase	*biology*	662	1186	27	1.252	32
Medical	*text*	978	1449	45	1.245	94
Enron	*text*	1702	1001	53	3.378	753

Besides the regular classification properties, such as label set and the number of examples, present specific statistical information for multi-label classification. This information includes: (1) Label Cardinality (*L*CARD)—a measure of "multi-labeled-ness" of a dataset introduced by Tsoumakas et al. [17] that quantifies the average number of labels per example in a dataset; and (2) Label Distinct Combinations (*L*DC)—a measure representing the number of distinct combinations of labels found in the dataset.

5.2 Procedure

As already mentioned, 11 different multi-label classification methods will be used in this investigation, where six are problem transformation methods (Binary Relevance (BR), Label Powerset (LP), Classifier Chains (CC), Pruned Sets (PS), CalibratedLabelRanking (CLR), and HOMER(HO)), and three are ensemble methods(Random k-labelsets(RA*k*EL), Ensemble of Classifier Chains (ECC) and Ensemble of Pruned Sets (EPS)) and the remaining two are algorithms adaptation methods (Multi-Label k Nearest Neighbours (ML-kNN) and Back-Propagation Multi-Label Learning (BPMLL)). The experimental results were evaluated using Accuracy, Hamming Loss, Micro- F-Measure, Macro- F-Measure, F-Measure and Subset Accuracy to evaluate different MLC methods. The experiments were conducted using the 10-fold cross-validation methodology. Thus, all results presented in paper refer to the mean over 10 different test sets. All multi-label classification methods and supervised learning algorithms used in this work are implementations of the Weka-based [18] package of Java classes for multi-label classification, called Mulan [19]. This package includes implementations of some of the multi-label classification methods most widely applied in the literature. All the algorithms were supplied with Weka's J48 implementation of a C4.5 tree classifier as a single-label base learner. The statistical significance of differences in algorithm results was determined by Friedman test [20].

5.3 Parameter configuration

As for multi-label classification algorithms used in this study, all configurable parameters of the participating algorithms were set to their optimal values as reported in the relevant papers. For BR, LP and CC no parameters were required. For HOMER, its balanced *k* means version with $k = 3$ was set. The PS methods required two parameters *p* and strategy parameter for each dataset. We used p=1 and strategy parameters, A_b, for all datasets with value of b=2 as proposed by [7]. Whereas the Ensemble methods configuration, the number of models in the ECC methods was set to 10 as proposed by [6], For RAKAL the number of models was set to 10 for all datasets and the size of the label-sets K for each datasets was set to half the number of labels. For EPS, at each dataset *p* and strategy, parameters were set to the same values as those used for the PS method. EPS requires additional parameter the number of models was set to 10. For all ensemble methods the majority voting threshold was set to 0.5.

6. RESULTS AND DISCUSSION

In this section, present the results from the experimental evaluation for each type of evaluation measure

6.1 Results on the example-based measures

The example-based evaluation measures include Hamming loss, accuracy, F1 score and subset accuracy. The results are given in tables 2-5 .We analyze the performance of the methods across all six evaluation measures using decision trees as a base classifier for all methods. In respect to the hamming loss, ECC and the RAKEL are the best performing methods according to hamming loss followed by EPS, where they have less mean rank. However hamming loss measures

the percentage of incorrectly predicted labels both positive and negative, thus the low values of the Hamming loss measure do not give an indication of high predictive performance. In respect to accuracy, the ECC is the best performance followed by RAKEL and EPS, where they have largest mean rank and best performance in most cases, while BPMLL performed best in two cases but in other cases given bad result such genbase and medical datasets. Accuracy seems to be a more balanced measure and better indicator of an actual algorithm's predictive performance for most standard classification problems than Hamming loss. In respect to F_1 score measure, also ECC is the best performing in most cases then RAKEL and EPS. Finally in respect to subset accuracy, the ECC is the best performing followed by RAKEL and EPS see (fig 1).

6.2 Results on the label-based measures

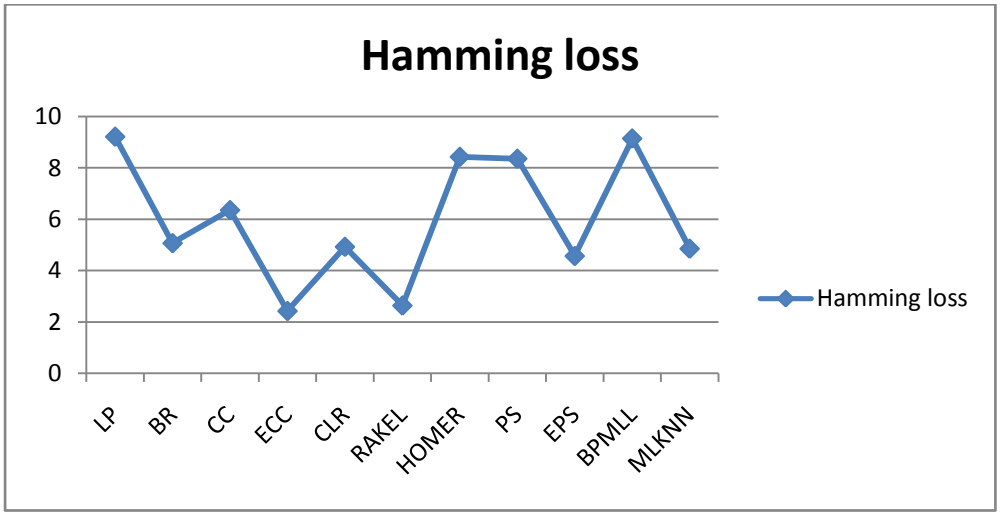
The label-based evaluation measures include micro-F1 and macro-F1. The results are given in table 6 and table 7. As for The label-based evaluation measures, we observed that, in micro-F1 and macro-F1 measures, the ECC or BR are best performance followed by CC and RAKEL. And the BPMLL and ML_MLL are better in some cases but very poor predictive performance in other cases see (fig.2).

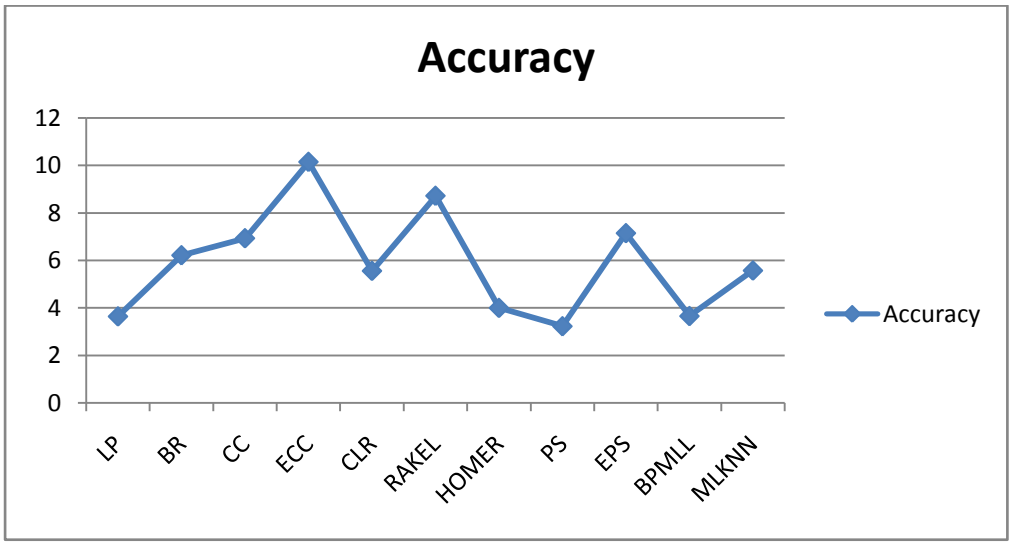
7. CONCLUSIONS

In this study, present an extensive experimental evaluation of multi-label classification methods. The topic of multi-label classification has lately received significant research effort. This has resulted in a variety of methods for addressing the task of multi-label learning. However, a wider experimental comparison of these methods is still lacking in the literature. We evaluated the most popular methods for multi-label classification using a wide range of evaluation measures on a variety of datasets.

The results of the experimental comparison revealed that the best performing method for evaluation measures when using decision trees as a base classifier for all methods. As for the example- based evaluation measures, the ECC is the best performance in all measures followed by RAKEL and EPS. As for the label-based evaluation measures, the best performing methods ECC and BR followed by CC and RAKEL.

From the experimental results, we observed that, all ensemble methods had provided the best results for almost all evaluation metrics. While Algorithm adaptation methods performed best on datasets with small labels and not had provided best results on datasets with high labels. This result may be an indication that the use of ensemble methods can be best choice.





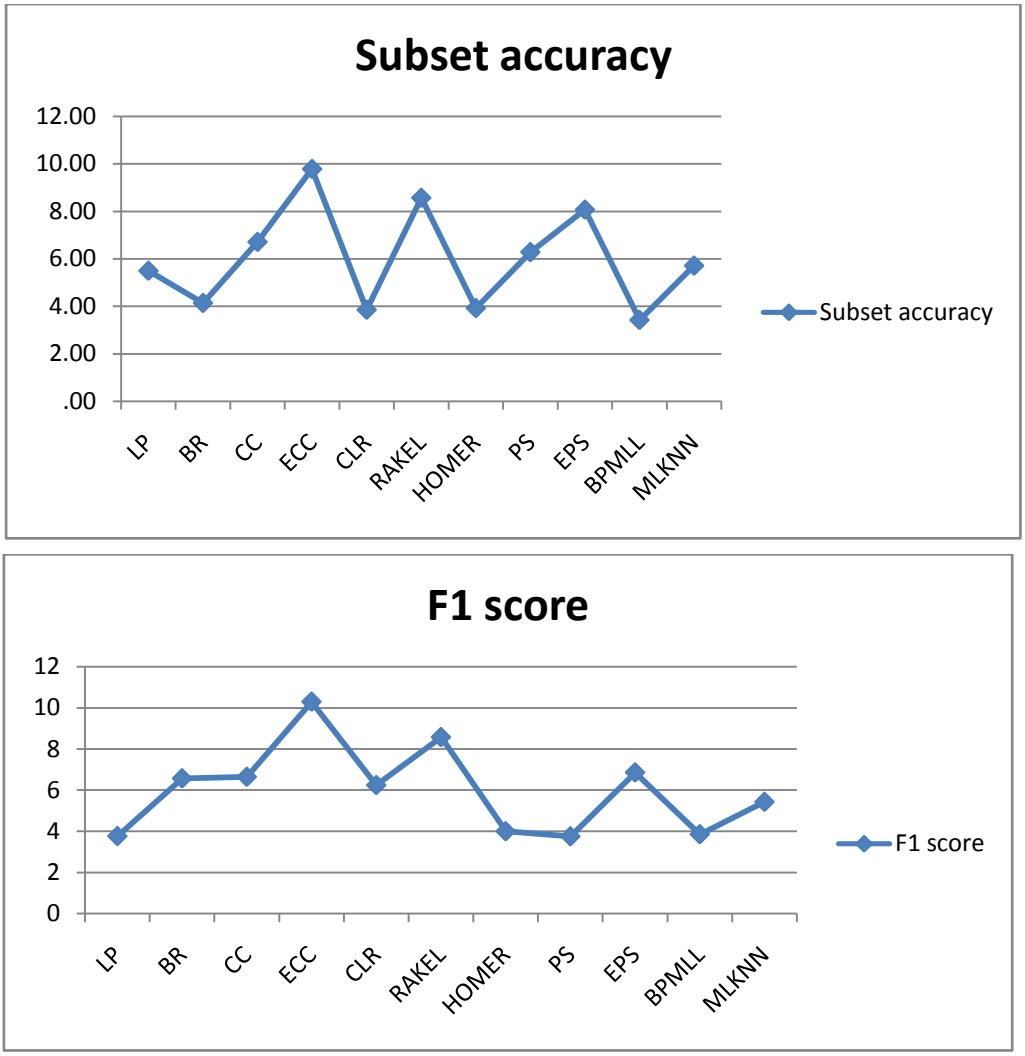
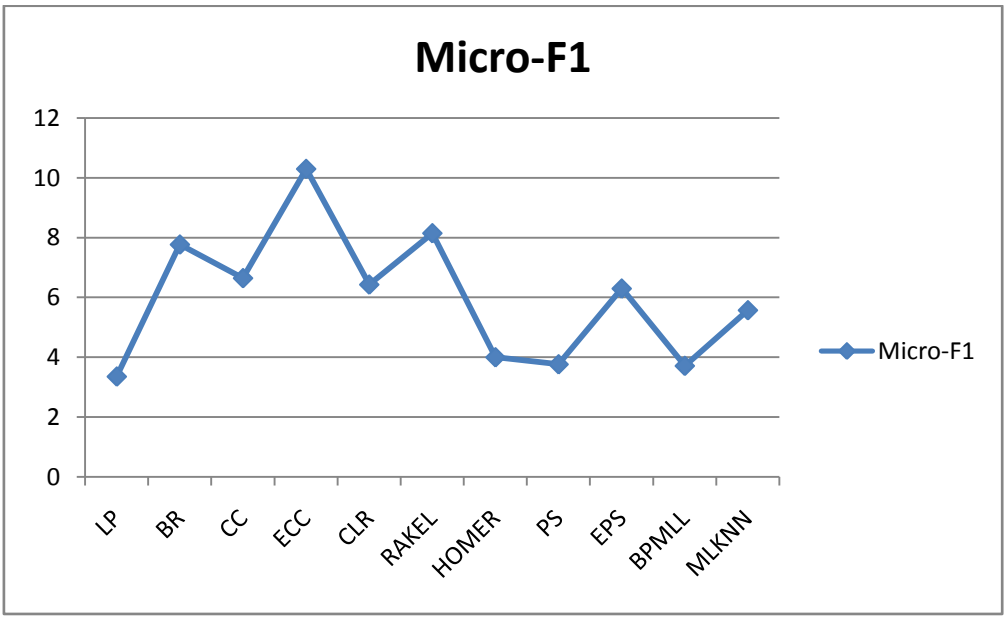


Figure.1. Mean rank of each algorithm over all datasets for the example-based evaluation measures using a Friedman test at *p* = 0.05.



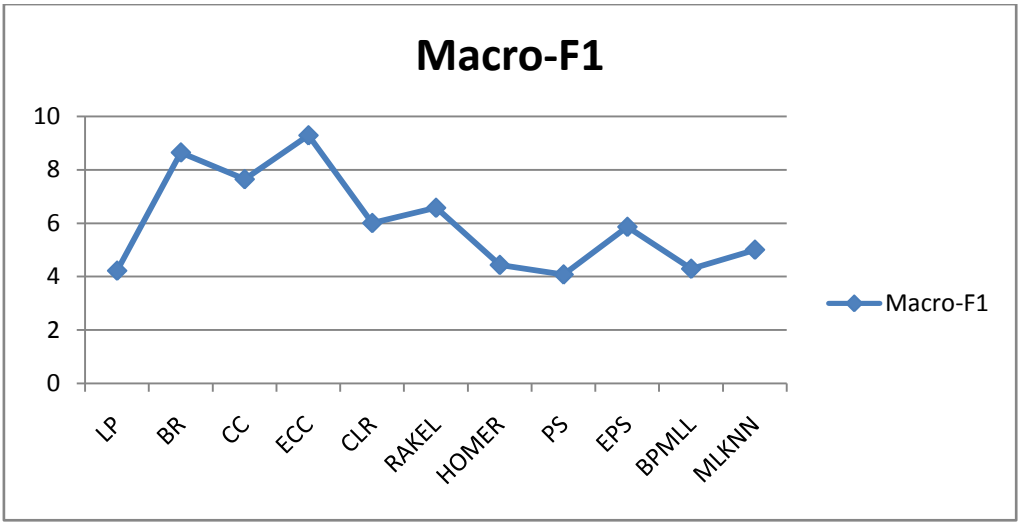


Figure.2. Mean rank of each algorithm over all datasets for the label-based evaluation measures using a Friedman test at p = 0.05.

Table. 2. The performance of the multi-label learning approaches in terms of the Hamming loss measure

Datasets	LP	BR	CC	ECC	CLR	RAKEL	HOMER	PS	EPS	BPMLL	MLKNN
Scene	0.1437	0.1368	0.1444	0.0920	0.1383	0.1012	0.1418	0.1425	0.0974	0.2633	**0.0862**
Emotions	0.2772	0.2474	0.2550	0.1954	0.2423	0.2181	0.2671	0.2727	0.2097	0.2043	**0.1951**
Yeast	0.2769	0.2454	0.2682	0.2041	0.2202	0.2030	0.2753	0.2799	0.2106	0.2239	**0.1933**
Birds	0.0735	0.0561	0.0562	0.0492	0.0506	**0.0489**	0.0788	0.0716	0.0508	0.0945	0.0543
Genbase	0.0019	**0.0011**	**0.0011**	0.0012	0.0013	0.0013	0.0021	0.0019	0.0020	0.4879	0.0048
Medical	0.0135	0.0103	0.0102	0.0098	0.0131	**0.0097**	0.0128	0.0127	0.0114	0.6406	0.0151
Enron	0.0708	0.0508	0.0524	0.0485	0.0471	**0.045**	0.0625	0.0635	0.0498	0.2788	0.0523

Table 3. The performance of the multi-label learning approaches in terms of the Accuracy measure

Datasets	**LP**	**BR**	**CC**	**ECC**	**CLR**	**RAKEL**	**HOMER**	**PS**	**EPS**	**BPMLL**	**MLKNN**
Scene	0.5893	0.5353	0.5866	**0.6702**	0.5265	0.6247	0.5936	0.5924	0.6447	0.3598	0.6670
Emotions	0.4376	0.4623	0.4703	0.5585	0.4771	0.5091	0.4522	0.4444	0.5264	**0.5665**	0.5326
Yeast	0.4144	0.4395	0.4280	0.5221	0.4685	0.5046	0.4032	0.4029	0.4958	**0.5232**	0.5162
Birds	0.4666	0.5295	0.5222	**0.5572**	0.5280	0.5452	0.4600	0.4516	0.5139	0.4321	0.4814
Genbase	0.9826	**0.9862**	**0.9862**	0.9847	0.9857	0.9845	0.9794	0.9826	0.9804	0.0362	0.9416
Medical	0.7358	0.7465	0.7581	0.7773	0.6202	**0.7774**	0.7453	0.7476	0.7516	0.0328	0.5813
Enron	0.3460	0.4129	0.4233	**0.4621**	0.4209	0.428	0.3817	0.3667	0.4226	0.1884	0.3316

Table 4. The performance of the multi-label learning approaches in terms of the F1 score measure

Datasets	LP	BR	CC	ECC	CLR	RAKEL	HOMER	PS	EPS	BPMLL	MLKNN
Scene	0.6037	0.5732	0.6032	**0.6847**	0.5732	0.6464	0.6100	0.6061	0.6588	0.4817	0.6811
Emotions	0.5178	0.5566	0.5482	0.6352	0.5712	0.5943	0.5407	0.5263	0.6087	**0.6586**	0.6138
Yeast	0.5171	0.5635	0.5279	0.6275	0.5841	0.6117	0.5167	0.5046	0.6034	**0.6350**	0.6204
Birds	0.4963	0.5611	0.5510	**0.5857**	0.5523	0.5706	0.4892	0.4782	0.5327	0.4368	0.4961
Genbase	0.9858	**0.9904**	**0.9904**	0.9894	0.9899	0.9888	0.9830	0.9858	0.9844	0.0675	0.9511
Medical	0.7600	0.7771	0.7850	**0.8063**	0.6496	0.8037	0.7703	0.7710	0.7762	0.0629	0.6065
Enron	0.4444	0.5257	0.5299	**0.5713**	0.5324	0.564	0.4932	0.4624	0.5293	0.2967	0.4288

Table 5. The performance of the multi-label learning approaches in terms of the Subset accuracy measure

Datasets	LP	BR	CC	ECC	CLR	RAKEL	HOMER	PS	EPS	BPMLL	MLKNN
Scene	0.5472	0.4266	0.5376	**0.6273**	0.3988	0.5604	0.5455	0.5521	0.6028	0.0565	0.6248
Emotions	0.2073	0.1838	0.2478	**0.3235**	0.1838	0.2478	0.1955	0.2055	0.2866	0.2882	0.2831
Yeast	0.1357	0.0683	0.1531	0.1800	0.0972	0.1671	0.0778	0.1282	0.1605	0.1390	**0.1874**
Birds	0.3820	0.4469	0.4501	**0.4841**	0.4626	0.4749	0.3914	0.3850	0.4687	0.4254	0.4441
Genbase	**0.9728**	0.9713	0.9713	0.9683	0.9698	0.9698	0.9683	**0.9728**	0.9668	0.0000	0.9110
Medical	0.6635	0.6553	0.6778	0.6891	0.5315	**0.6993**	0.6716	0.6788	0.6788	0.0000	0.5060
Enron	0.1116	0.1028	0.1269	**0.1445**	0.1005	0.136	0.1010	0.1316	0.1386	0.0012	0.0740

Table 6. The performance of the multi-label learning approaches in terms of the micro-F1 measure

Datasets	LP	BR	CC	ECC	CLR	RAKEL	HOMER	PS	EPS	BPMLL	MLKNN
Scene	0.5982	0.6194	0.6001	0.7250	0.6276	0.6979	0.6060	0.6012	0.7042	0.5197	**0.7343**
Emotions	0.5499	0.6020	0.5878	0.6805	0.6276	0.6404	0.5668	0.5608	0.6540	**0.6896**	0.6598
Yeast	0.5411	0.5857	0.5499	**0.6503**	0.6158	0.6369	0.5448	0.5317	0.6299	0.6499	0.6471
Birds	0.3258	**0.3747**	0.3514	0.3608	0.3127	0.3344	0.2584	0.2634	0.2495	0.0730	0.1934
Genbase	0.9801	**0.9880**	**0.9880**	0.9869	0.9862	0.9864	0.9782	0.9801	0.9789	0.0932	0.9462
Medical	0.7529	0.8091	0.8115	**0.8226**	0.7136	0.8187	0.7626	0.7643	0.7846	0.0636	0.6800
Enron	0.4364	0.5481	0.5363	**0.5731**	0.5672	0.548	0.4920	0.4528	0.5313	0.2798	0.4778

Table 7. The performance of the multi-label learning approaches in terms of the macro-F1 measure

Datasets	LP	BR	CC	ECC	CLR	RAKEL	HOMER	PS	EPS	BPMLL	MLKNN
Scene	0.6092	0.6285	0.6126	0.7318	0.6442	0.7073	0.6164	0.6140	0.7097	0.5410	**0.7355**
Emotions	0.5377	0.5868	0.5760	0.6659	0.6167	0.6225	0.5523	0.5461	0.6289	**0.6783**	0.6243
Yeast	0.3866	0.3920	0.3966	0.4039	0.3834	0.3852	0.3894	0.3782	0.3763	**0.4358**	0.3853
Birds	0.3709	**0.3856**	0.3585	0.3572	0.3494	0.3744	0.2796	0.2992	0.3489	0.1338	0.3183
Genbase	0.9464	**0.9572**	**0.9572**	0.9533	0.9561	0.9536	0.9459	0.9464	0.9349	0.2420	0.8403
Medical	0.6886	0.7566	0.7578	**0.7590**	0.6838	0.7549	0.7078	0.7083	0.7386	0.2858	0.6574
Enron	0.2194	**0.3171**	0.3086	0.3128	0.3052	0.115	0.2759	0.2760	0.2902	0.2625	0.2569

8. REFERENCES

[1] G. Tsoumakas, I. Katakis, Multi label classification: an overview, International Journal of Data Warehouse and Mining 3 (3) (2007) 1–13.

[2] G. Tsoumakas, I. Vlahavas, Random k-labelsets: an ensemble method for multi-label classification, in: Proceedings of the 18th European conference on Machine Learning, 2007, pp. 406–417.

[3] J. Read, B. Pfahringer, G. Holmes, Multi-label classification using ensembles of pruned sets, in: Proceedings of the 8th IEEE International Conference on Data Mining, 2008, pp. 995–1000.

[4] J.F ¨ urnkranz, Round robin classification, Journal of Machine Learning Research 2(2002)721–747.

[5] T.-F. Wu, C.-J. Lin, R.C. Weng, Probability estimates for multi-class classification by pairwise coupling, Journal of Machine Learning Research 5 (2004) 975–1005.

[6] J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains for multi-label classification, in: Proceedings of the 20th European Conference on Machine Learning, 2009, pp. 254–269.

[7] J. Read, B. Pfahringer and G. Holmes, "Multi-label classification using ensembles of pruned sets", Proc 8th IEEE International Conference on Data Mining, Pisa, Italy, pages 995-1000. IEEE Computer Society, 2008.

[8] G. Tsoumakas, I. Katakis, I. Vlahavas, Effective and efficient multi-label classification in domains with large number of labels, in: Proceedings of the ECML/PKDD Workshop on Mining Multidimensional Data, 2008, pp. 30–44.

[9] S.-H.Park,J.F ¨ urnkranz, Efficient pairwise classification, in: Proceedings of the 18th European Conference on Machine Learning,2007,pp.658–665.

[10] E.L.Mencı´a, S.-H.Park,J.F ¨ urnkranz, Efficient voting prediction for pairwise multi-label classification,Neurocomputing73(2010)1164–1176.

[11] A. Clare, R.D. King, Knowledge discovery in multi-label phenotype data, in: Proceedings of the 5th European Conference on PKDD, 2001, pp. 42–53.

[12] M.L. Zhang, Z.H. Zhou, Multi-label neural networks with applications to functional genomics and text categorization, IEEE Transactions on Knowledge and Data Engineering 18 (10) (2006) 1338–1351.

[13] R.E. Schapire, Y. Singer, Boostexter: a boosting-based system for text categorization, Machine Learning 39 (2000) 135–168.

[14] F.DeComite, R.Gilleron, M. Tommasi, Learning multi-label alternating decision trees from texts and data, in: Proceedings of the 3rd international conference on Machine learning and data mining in pattern recognition, 2003, pp.35–49.

[15] M.L. Zhang, Z.H. Zhou, ML-kNN: a lazy learning approach to multi-label learning, Pattern Recognition 40 (7) (2007) 2038–2048.

[16] A. Elisseeff, J. Weston, A Kernel method for multi-labelled classification, in: Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval, 2005, pp. 274–281.

[17] G. Tsoumakas, I. Katakis, I. Vlahavas, Mining multi-label data, in: Data Mining and Knowledge Discovery Handbook, Springer, Berlin/Heidelberg, 2010, pp. 667–685.

[18] I. H.Witten, ans E. Frank, "Data Mining: Practical Machine Learning tools and techniques", Morgan Kaufmann, 2005.

[19] G. Tsoumakas, R. Friberg, E. Spyromitros-Xiou, I, Kataks, and J. Vilcek, "Mulan software - java classes for multi-label classification Available at: http://mlkd.csd.auth.gr/multilabel.html#Software

[20] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, Annals of Mathematical Statistics 11 (1940) 86–92.