Information Gain Feature Selection for Multi-Label Classification

Rafael B. Pereira[1], Alexandre Plastino[1], Bianca Zadrozny[2], Luiz H. C. Merschmann[3]

[1] Universidade Federal Fluminense, Brazil
rbarros@ic.uff.br, plastino@ic.uff.br
[2] IBM Research, Brazil
biancaz@br.ibm.com
[3] Universidade Federal de Ouro Preto (UFOP), Brazil
luizhenrique@iceb.ufop.br

Abstract. In many important application domains, such as text categorization, biomolecular analysis, scene or video classification and medical diagnosis, instances are naturally associated with more than one class label, giving rise to multi-label classification problems. This fact has led, in recent years, to a substantial amount of research in multi-label classification. And, more specifically, many feature selection methods have been developed to allow the identification of relevant and informative features for multi-label classification. However, most methods proposed for this task rely on the transformation of the multi-label data set into a single-label one. Besides, there is no single work that carries out a comprehensive evaluation of the various multi-label classification techniques coupled with feature selection methods over data sets from different domains. In this work, we perform these experimental evaluations, and also propose an adaptation of the information gain feature selection technique to handle multi-label data directly.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications; I.2.6 [**Artificial Intelligence**]: Learning

Keywords: classification, data mining, feature selection, multi-label classification

1. INTRODUCTION

A large body of research in supervised learning deals with the analysis of single-label data, where instances are associated with a single label from a set of class labels. More specifically, the single-label classification problem can be stated as the process of predicting the class label of new instances described by their feature values. However, in many data mining applications, the instances can be associated with more than one class label. This characterizes the multi-label classification problem, a relevant topic of research, which has become a very common real-world task [Zhang and Zhou 2007].

Classification strategies that deal with multi-label data can be divided into two groups: transformation and adaptation strategies. Transformation strategies convert the multi-label data into single-label data and then use single-label classifiers. Adaptation strategies adapt or extend single-label classifiers to cope with multi-label data directly. In the former group one can find popular methods like Label Powerset and Binary Relevance transformations, and in the latter group some adaptations are: the multi-label k-nearest neighbors (ML-KNN) [Zhang and Zhou 2007], the ML Naive Bayes classifier [Zhang et al. 2009], ML Decision Tree [Clare and King 2001], among others [Tsoumakas et al. 2010].

The performance of a classification method is closely related to the inherent quality of the training data. Redundant and irrelevant features may not only decrease the classifier's accuracy but also make

the process of building the model or running the classification algorithm slower. Feature selection is a data preprocessing step which aims at identifying relevant features for a target data mining task – specifically in this paper, the multi-label classification task.

There is an extensive literature regarding feature selection for single-label classification, which has been summarized in surveys such as in [Dash and Liu 1997; Guyon et al. 2006]. In the last few years, given the increasing popularity of multi-label classification and the challenge of selecting features in this context, there has been significant research specifically on feature selection for multi-label classification. Most methods proposed for this task rely on the transformation of the multi-label data set into a single-label one. This can cause a loss of information from the multi-label data, like label dependence, an important issue in multi-label learning [Spolaôr et al. 2013]. Also, there is no single work that carries out a comprehensive evaluation of the various multi-label classification techniques coupled with feature selection methods over data sets from different domains.

In this work, we perform a comprehensive evaluation of various multi-label feature selection techniques, and propose an adaptation of the information gain metric to handle multi-label data directly. Using data sets from various domains, including large data sets, the proposed algorithm is experimentally compared to well-known transformation-based feature selection techniques coupled with multi-label classifiers. The results show that the proposed algorithm is competitive and more scalable than the other compared techniques.

The remainder of this paper is organized as follows. In Section 2, we revisit the multi-label classification problem. In Section 3, we describe the multi-label feature selection process and current work. In Section 4, we describe our adaptation proposal of a novel multi-label feature selection technique and the experiments that compare it with methods currently used in the literature. Finally, in Section 5, we make our concluding remarks and point to directions for future research.

2. MULTI-LABEL CLASSIFICATION

In the multi-label classification task, each data instance may be associated with multiple labels. Multi-label classification is suitable for many domains, such as text categorization, scene or video classification, medical diagnosis, bioinformatics and microbiology. In all these cases, the task is to assign for each unseen instance a label set whose size is unknown a priori [Zhang and Zhou 2007].

The simplest way to apply a classification algorithm to multi-label data is to transform them into single-label data. Then a traditional classification technique – like k-NN or a decision tree – can be employed to perform the classification task. The advantage of using a transformation technique is allowing the usage of one or more single-label classification algorithms for the learning task, which have been thoroughly studied and perfected over the last decades.

Simple transformation techniques used to convert a multi-label data set into a single-label one consist of selecting among the label subsets of each instance the most frequent label in the data set (select-max), the least frequent label (select-min), a random label (select-random) or simply discard every multi-label example (select-ignore, although this is not useful if all the data set is multi-label) [Boutell et al. 2004; Chen et al. 2007]. Another type of transformation consists of copying each multi-label instance n times, where n is the number of labels assigned to that instance. Each copied instance is then assigned one distinct single label from the original set.

A popular transformation is the Label Powerset (LP) technique, which creates one label for each different subset of labels that exists in the multi-label training data set. Thus, the new set of labels corresponds to the powerset of the original set of labels. After this transformation process, a single-label classification learning algorithm can handle the transformed data set and produce a classifier. This classifier can then be used to assign to new instances one of these new labels, which can be mapped back to the corresponding subset of the original labels [Tsoumakas and Vlahavas 2007].

Label Powerset is recommended only for data sets with a small number of labels, as the number of meta-labels produced in LP is exponential in the number of labels, which is clearly problematic from a classification point of view [Dembczyński et al. 2012]. With the goal of alleviating this problem, the original LP technique has been extended and improved in subsequent work. A few variations are the Pruned Problem Transformation (PPT), proposed in [Read 2008]; Random k-Labelsets (RAKEL) [Tsoumakas and Vlahavas 2007]; and HOMER [Tsoumakas et al. 2008]. In general, these methods construct more than one multi-label classifier, each one dealing with a much smaller set of labels.

Binary Relevance (BR) is a well-known transformation technique that produces a binary classifier for each different label of the original data set. In its simplest implementation, each resulting classifier is capable of predicting if a label is relevant or not for a new instance. So, each classifier handles the data as single-label, since it gives a relevance feedback for just one specific label.

Binary Relevance does not take into account label dependence [Dembczyński et al. 2012], so it may fail to accurately predict label combinations or rank labels [Tsoumakas et al. 2010]. In order to reduce this drawback, several techniques, such as the Classifier Chains (CC) method [Read et al. 2009; Silva et al. 2014; Gonçalves et al. 2013], have been proposed to extend and improve the BR technique.

Regarding algorithm adaptation, most traditional classifiers employed in single-label problems have been adapted to the multi-label paradigm [Tsoumakas et al. 2010]. C4.5 decision-tree induction algorithm was adapted in [Clare and King 2001], by allowing multiple labels in the leaves of the tree. An adaptation of the SVM algorithm has been proposed in [Elisseeff and Weston 2001]. A k-NN adaptation was proposed in [Zhang and Zhou 2007]. A multi-label adaptation of the Naive Bayes algorithm was proposed in [Zhang et al. 2009]. MMAC (Multi-class, Multi-label Associative Classification) follows the paradigm of associative classification which deals with the construction of multi-label classification rule sets using association rule mining [Tsoumakas et al. 2010].

3. MULTI-LABEL FEATURE SELECTION

Feature selection techniques are primarily employed to identify relevant and informative features [Guyon et al. 2006]. Besides, there are other important motivations: the improvement of a classifier predictive accuracy, the reduction and simplification of the data set, the acceleration of the classification task, the simplification of the generated classification model, and others.

In [Chen et al. 2007], the following common simple transformation techniques have been employed to allow the application of traditional feature selection for the multi-label text categorization problem: select-max, select-min, select-random, select-ignore and copy, used to convert a multi-label data set into a single-label one.

In [Trohidis et al. 2008], several multi-label classification strategies were evaluated and compared for the task of automated decision of emotion in a music data set. The Label Powerset transformation was used to produce a single-label data set, and then a common feature selection measure was employed (χ^2 statistic) to select the best features. The work verified that, by using feature selection, the classification result achieved a better Hamming Loss measure than without feature selection, for the evaluated data set and the ML-KNN algorithm as the classifier.

The Label Powerset transformation is also used for feature selection in [Spolaôr et al. 2013], in conjunction with the relief and information gain measures. With this feature selection, it was possible to reduce the size of the data sets without compromising the classification performance. In [Doquire and Verleysen 2011], the Pruned Problem Transformation (PPT) [Read 2008], based on the Label Powerset, was used in the data transformation step before performing the mutual information feature selection on three real-world data sets from different domains.Then the ML-KNN algorithm was employed over the original multi-label data containing only the selected features. When compared with the χ^2 statistic adopted in [Trohidis et al. 2008], in conjunction with the Label Powerset transfor-

mation, and also with a non-feature selection scenario, the mutual information measure allowed the classification phase to achieve a better result in terms of the Hamming Loss and the accuracy of the classifier. In [Tsoumakas and Vlahavas 2007], feature selection was applied to a textual data set to reduce the computational cost of training the RAKEL classifier. The χ^2 statistic was used separately for each label in order to obtain different rankings of all features.

Some text classification work [Yang and Pedersen 1997; Olsson and Oard 2006; Zheng et al. 2004] has employed the Binary Relevance technique before applying single-label feature selection measures, like information gain and χ^2 statistic. For each different label in the original data set, a binary single-label data set is created, and then feature selection is executed for each one. Binary Relevance transformation is also used for feature selection in [Spolaôr et al. 2013], in conjunction with relief and information gain measures. This feature selection strategy is compared with LP transformation using the same measures, with the conclusion that both transformation methods achieved a similar predictive performance in the experiments with data sets from various multi-label domains.

There are also recently proposed multi-label feature selection techniques that do not require transformation of the data set in order to work – the feature selection is built as an adaptation of techniques suited for the single-label paradigm, or as a wrapper-based technique. In [Zhang et al. 2009], a wrapper technique is used to identify the best feature set. The wrapper feature selection implements a genetic algorithm as the search component. To evaluate this method, the Multi-label Naive Bayes classifier – proposed in the same work – is employed to select the best features. The classification coupled with the feature selection achieved a better result, even when compared with other classifiers.

Common single-label feature selection techniques were adapted to the multi-label paradigm recently. The ReliefF measure was adapted in [Pupo et al. 2013] and in [Spolaôr et al. 2013]. The Mutual Information measure was adapted in [Lee and Kim 2013]. Correlation-based feature selection, capable of handling subset of features, was adapted to the multi-label setting in [Jungjit et al. 2013].

4. COMPARISON BETWEEN MULTI-LABEL FEATURE SELECTION METHODS

4.1 Information Gain feature selection adaptation

In this work, we adapt the information gain measure, based on the entropy concept, to the multi-label feature selection. The entropy is commonly used as measure of feature relevance in filter strategies that evaluate features individually [Yang and Pedersen 1997], and this method has the advantage of being fast. Let $D(A_1, A_2, ..., A_n, C)$, $n \geq 1$, be a data set with $n+1$ attributes, where C is the class attribute. Let m be the number of distinct class values, in a single-label context. The entropy of the class distribution in D, represented by $Entropy(D)$, is defined by Equation 1.

$$Entropy(D) = -\sum_{i=1}^{m} p_i * log_2(p_i), \tag{1}$$

where p_i is the probability that an arbitrary instance in D belongs to class c_i.

The concept defined in Equation 1 is used by the single-label strategy known as Information Gain Attribute Ranking [Yang and Pedersen 1997] to measure the ability of a feature to discriminate between class values.

In [Clare and King 2001], the C4.5 algorithm was adapted for handling multi-label data. This decision tree algorithm allowed multiple labels at the leaves of the tree, by using an adaptation of entropy calculation, described by Equation 2.

$$Entropy.ML(D) = -\sum_{i=1}^{l} p(\lambda_i) * log_2 p(\lambda_i) + q(\lambda_i) * log_2 q(\lambda_i), \tag{2}$$

where $p(\lambda_i)$ is the probability that an arbitrary instance in D belongs to class label λ_i, $q(\lambda_i) = 1-p(\lambda_i)$, and l is the number of labels in the data set. We have adopted this formula to create an information gain feature selection capable of handling multi-label data. By using this as a filter approach, the feature selection can be employed with any multi-label classifier.

The feature selection algorithm works as follows: it receives as input a multi-label data set. Then it computes a multi-label information gain score for each feature using the Entropy.ML measure defined in Equation 2. Next, all the scores are sorted in a ranking. In order to have a list of selected features as an output, it is necessary to inform the number of selected features. This can be either a percentage of the total number of features or a score threshold to split the ranking. In this work we have opted for a percentage of features, in order to compare each technique with equal conditions.

4.2 Experimental Evaluation

In this work we have compared our proposed information gain adaptation (MLInfoGain) with other multi-label feature selection techniques by executing a large number of experiments. For this purpose we have elected commonly used multi-label data sets and classification algorithms. The experiments were executed using the Mulan framework [Tsoumakas et al. 2010]. Mulan is an open-source Java library for learning from multi-label data sets with a variety of state-of-the-art algorithms.

We used in our experiments data sets from various domains available in the Mulan website [Tsoumakas et al. 2010]. Most of the initiatives that compare multi-label learning algorithms experimentally adopt a subset of these available data sets.

The feature selection techniques compared were: Binary Relevance, Copy Transformation, Label Powerset and our proposed Multi-label Information Gain technique. All transformation methods are coupled with the single-label information gain ranking method, in order to achieve an unbiased comparison. The information gain measure requires discrete feature values. Therefore we adopted the recursive entropy minimization heuristic [Fayyad and Irani 1993] to discretize continuous attributes, and a simple unsupervised technique with 10 bins for the multi-label information gain technique.

Each feature selection technique was experimented with nine executions in which we varied the percentage of selected features between 10% and 90%, in increments of 10%. We evaluated the classifiers using 10-fold cross-validation. As an example, Table I shows the results obtained with the BR-KNN classifier, for the Hamming Loss measure and the proposed Multi-label Information Gain technique, compared with the results without feature selection (100%) as a baseline. In bold we mark the results that achieved a value equal or better than the baseline. It is possible to see that most of the feature selection options improve the predictive performance of the classification algorithm, reducing the number of features and achieving a better Hamming Loss score.

Table I. Results achieved with the BR-KNN classifier for the Hamming Loss measure

Data	**Multi-label Information Gain**									**No Sel.**
Set	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
bibtex	**0.0132**	**0.0134**	**0.0135**	**0.0137**	**0.0138**	**0.0139**	**0.0141**	**0.0142**	**0.0143**	0.0143
birds	**0.0438**	**0.0454**	0.0468	0.0459	0.0458	0.0457	0.0459	0.0461	0.0461	0.0454
CAL500	0.1435	**0.1423**	**0.1417**	**0.1412**	**0.1420**	**0.1423**	**0.1422**	**0.1419**	**0.1422**	0.1425
Corel5k	**0.0094**	**0.0094**	**0.0094**	**0.0094**	**0.0094**	**0.0094**	**0.0094**	**0.0094**	**0.0094**	0.0094
emotions	0.2139	0.2128	0.2081	0.2022	0.1949	**0.1918**	**0.1929**	**0.1890**	**0.1901**	0.1934
enron	**0.0580**	0.0596	0.0604	0.0604	0.0581	**0.0576**	**0.0565**	**0.0571**	**0.0568**	0.0580
flags-ml	**0.2655**	**0.2540**	**0.2474**	**0.2595**	**0.2637**	**0.2630**	**0.2681**	**0.2712**	**0.2661**	0.2749
genbase	**0.0038**	**0.0038**	**0.0038**	**0.0038**	**0.0038**	**0.0038**	**0.0038**	**0.0038**	**0.0038**	0.0038
medical	**0.0160**	**0.0169**	**0.0175**	**0.0175**	**0.0176**	**0.0177**	0.0182	0.0184	0.0182	0.0180
scene	0.1559	0.1351	0.1152	0.1084	0.0999	0.0957	0.0935	0.0931	0.0928	0.0920
yeast	0.2137	0.2086	0.1971	0.1963	0.1969	0.1959	0.1953	**0.1942**	0.1964	0.1952

We have employed a large number of classification techniques, from both the transformation paradigm as well as the algorithm adaptation paradigm. The transformation techniques used were: Label Power-

set, Binary Relevance, Classifier Chains, RaKEL and HOMER, coupled with the k-NN, Decision trees (J48) and Naive Bayes single-label classifiers. The algorithm adaptations employed in this experiment were the ML-kNN and the IBLR classifier.

Mulan contains an evaluation framework that calculates a rich variety of performance measures [Tsoumakas et al. 2010]. The following multi-label measures were chosen to evaluate the results: Hamming Loss, Subset 0/1 Loss (counterpart of Subset Accuracy), Example-based Accuracy and Ranking Loss. They were chosen based on their current use in the literature and their diversity, since measures with similar equations are more likely to yield results correlated with each other. Their formulas can be found on related work, like in [Tsoumakas et al. 2010]. Example-based Accuracy values were inverted, so that all measures have the same pattern: the lower the value, the better.

Table II shows the overall result of each feature selection technique coupled with the BR-KNN classifier. Each table section presents the result for a specific performance measure. The first column indicates the data set used. "BR+InfoGain", "Copy+InfoGain" and "LP+InfoGain" stand for a transformation followed by the single-label information gain measure to rank and select features. "MLInfoGain" corresponds to the multi-label information gain technique proposed in this work. "No Sel." is the result without feature selection, and also our baseline. Each cell shows the result of the multi-label measure achieved in each case, varying between 0 and 1, and the lower the value, the better. In parenthesis we show the percentage of selected features that achieved the best value for each technique, and in case of ties we report the smaller percentage. Bold values show the results that achieved a result equal or better than the baseline, and underlined values show the best result achieved in each row. At the end of the table we summarize the results.

With the BR-KNN classifier, the proposed multi-label information gain technique (MLInfoGain) achieved a competitive result, holding the best performance in 22 cases, out of the 44 experiments. The BR+InfoGain also achieved the best result in 22 cases. Only in 8 cases the result without feature selection achieved the best result, indicating that in most cases feature selection is helpful. In 41 cases, the proposed multi-label information gain technique was able to yield a value equal or better than the baseline (without feature selection).

It is worth noting the behaviour for some data sets: the genbase data set is not affected by feature selection, which indicates that it can be drastically reduced without compromising its performance; on the other hand, the scene data set achieves a better performance with most of its features, indicating that it is less suitable for feature selection.

Table III corresponds to a summarized result of the other classifiers performance when coupled with feature selection, similar to the last row of the previous table. It shows the number of times that each feature selection achieved a result better than ($\leq$) the baseline score, considering the evaluated data sets and the four performance measures adopted in this work. The results indicate that most of the time the feature selection was beneficial for the overall classification. For instance, when using the RAKEL + K-NN classifier, the BR + InfoGain feature selection achieved a performance equivalent or better than the result without feature selection 37 times out of 44 results (i.e. 4 measures x 11 data sets). For the Copy + InfoGain feature selection this result was achieved 32 times; for the LP + InfoGain 31 times; and for the proposed MLInfoGain this result occurred 39 times.

4.3 Statistical evaluation

We have employed a Friedman test in order to evaluate if the differences in performance of the multi-label feature selection techniques are statistically significant. A non-parametric test makes no assumption about the data distribution, unlike, for instance, a paired t-test which assumes data normality. We have followed the same procedure described in [Madjarov et al. 2012].

The feature selection techniques were ranked according to their performance for each classification algorithm and data set. The best performing technique was ranked first, the second best was ranked

Table II. Best results achieved with the BR-KNN classifier

HAMMING LOSS					
Data Set	**BR+InfoGain**	**Copy+InfoGain**	**LP+InfoGain**	**MLInfoGain**	**No Sel.**
bibtex	**0.0128 (10%)**	**0.0132 (10%)**	**0.0137 (20%)**	**0.0132 (10%)**	0.0143
birds	**0.0447 (30%)**	0.0458 (90%)	0.0456 (80%)	**0.0438 (10%)**	0.0454
CAL500	**0.1411 (80%)**	**0.1416 (40%)**	**0.1410 (30%)**	**0.1412 (40%)**	0.1425
Corel5k	**0.0094 (10%)**	**0.0094 (10%)**	**0.0094 (10%)**	**0.0094 (10%)**	0.0094
emotions	**0.1917 (90%)**	**0.1910 (80%)**	0.1951 (90%)	**0.1890 (80%)**	0.1934
enron	**0.0525 (10%)**	**0.0579 (10%)**	**0.0523 (10%)**	**0.0565 (70%)**	0.0580
flagsml	**0.2510 (20%)**	**0.2570 (20%)**	**0.2540 (20%)**	**0.2474 (30%)**	0.2749
genbase	**0.0038 (10%)**	**0.0038 (10%)**	**0.0038 (10%)**	**0.0038 (10%)**	0.0038
medical	**0.0139 (10%)**	**0.0160 (10%)**	**0.0162 (10%)**	**0.0160 (10%)**	0.0180
scene	0.0958 (90%)	0.0932 (90%)	0.0947 (90%)	0.0928 (90%)	0.0920
yeast	**0.1924 (70%)**	0.1971 (50%)	**0.1945 (90%)**	**0.1942 (80%)**	0.1952
SUBSET 0/1 LOSS					
Data Set	**BR+InfoGain**	**Copy+InfoGain**	**LP+InfoGain**	**MLInfoGain**	**No Sel.**
bibtex	**0.8817 (10%)**	**0.9120 (10%)**	**0.9516 (30%)**	**0.9118 (10%)**	0.9754
birds	**0.4945 (50%)**	0.5084 (70%)	0.5069 (70%)	**0.4852 (20%)**	0.5039
CAL500	**1.0000 (10%)**	**1.0000 (10%)**	**1.0000 (10%)**	**1.0000 (10%)**	1.0000
Corel5k	**0.9992 (50%)**	**0.9994 (70%)**	**0.9992 (90%)**	**0.9994 (30%)**	1.0000
emotions	**0.6985 (30%)**	**0.6883 (70%)**	**0.7035 (90%)**	**0.6732 (80%)**	0.7085
enron	**0.8908 (10%)**	**0.8837 (40%)**	**0.8996 (40%)**	**0.8866 (40%)**	0.9195
flagsml	**0.8084 (20%)**	**0.8450 (20%)**	**0.8087 (20%)**	**0.8034 (30%)**	0.8547
genbase	**0.0785 (10%)**	**0.0785 (10%)**	**0.0785 (10%)**	**0.0785 (10%)**	0.0785
medical	**0.4530 (10%)**	**0.5471 (10%)**	**0.5471 (10%)**	**0.5359 (10%)**	0.5982
scene	0.4130 (90%)	0.4088 (90%)	0.4088 (90%)	**0.4005 (80%)**	0.4038
yeast	**0.7985 (90%)**	**0.8014 (90%)**	0.8056 (90%)	**0.7964 (80%)**	0.8018
EXAMPLE-BASED ACCURACY (INVERTED)					
Data Set	**BR+InfoGain**	**Copy+InfoGain**	**LP+InfoGain**	**MLInfoGain**	**No Sel.**
bibtex	**0.7894 (10%)**	**0.8369 (10%)**	**0.8848 (30%)**	**0.8369 (10%)**	0.9289
birds	**0.4443 (30%)**	0.4560 (90%)	0.4535 (80%)	**0.4282 (10%)**	0.4482
CAL500	**0.8094 (80%)**	**0.8107 (70%)**	**0.8099 (60%)**	**0.8106 (40%)**	0.8144
Corel5k	**0.9915 (80%)**	**0.9928 (70%)**	**0.9941 (80%)**	**0.9925 (70%)**	0.9975
emotions	**0.4702 (70%)**	**0.4686 (80%)**	0.4871 (50%)	**0.4643 (80%)**	0.4851
enron	**0.6530 (10%)**	**0.7314 (20%)**	**0.7000 (10%)**	**0.7162 (70%)**	0.7973
flagsml	**0.3953 (20%)**	**0.3945 (20%)**	**0.3903 (20%)**	**0.3824 (30%)**	0.4364
genbase	**0.0463 (10%)**	**0.0463 (10%)**	**0.0463 (10%)**	**0.0463 (10%)**	0.0463
medical	**0.3815 (10%)**	**0.4799 (10%)**	**0.4828 (10%)**	**0.4718 (10%)**	0.5437
scene	0.3881 (90%)	0.3831 (90%)	0.3837 (90%)	**0.3750 (80%)**	0.3802
yeast	**0.4975 (90%)**	0.5037 (90%)	0.5002 (90%)	**0.4965 (80%)**	0.4998
RANKING LOSS					
Data Set	**BR+InfoGain**	**Copy+InfoGain**	**LP+InfoGain**	**MLInfoGain**	**No Sel.**
bibtex	**0.1342 (10%)**	**0.1807 (10%)**	**0.2296 (30%)**	**0.1805 (10%)**	0.2830
birds	**0.0861 (70%)**	0.0889 (90%)	0.0878 (40%)	0.0872 (60%)	0.0864
CAL500	**0.2301 (70%)**	**0.2301 (30%)**	**0.2295 (40%)**	**0.2310 (90%)**	0.2310
Corel5k	**0.1887 (10%)**	**0.1997 (10%)**	**0.2254 (10%)**	**0.1983 (10%)**	0.3243
emotions	0.1624 (70%)	0.1623 (80%)	**0.1599 (90%)**	**0.1584 (60%)**	0.1610
enron	**0.1165 (10%)**	**0.1096 (10%)**	**0.1260 (10%)**	**0.1087 (10%)**	0.1655
flagsml	**0.1815 (50%)**	**0.1855 (20%)**	**0.1816 (50%)**	**0.1891 (40%)**	0.1978
genbase	**0.0052 (10%)**	**0.0052 (10%)**	**0.0052 (10%)**	**0.0052 (10%)**	0.0052
medical	**0.0350 (10%)**	**0.0438 (10%)**	**0.0445 (10%)**	**0.0437 (10%)**	0.0475
scene	0.0925 (90%)	0.0902 (90%)	0.0927 (90%)	0.0905 (90%)	0.0889
yeast	**0.1757 (90%)**	**0.1766 (90%)**	0.1797 (90%)	**0.1755 (80%)**	0.1778
Best values (underlined)	22	7	10	22	8
$\leq$ baseline score (bold)	39	33	31	41	

second, and so on. In case of ties, the ranks were averaged. From the average ranks of the techniques, the Friedman statistic was calculated, and then at a significance level of 5%, the hypothesis that techniques performed equally in mean ranking was rejected.

Then a post-hoc Nemenyi test was used to compare the feature selection techniques to each other. The performance of two techniques is considered significantly different if their average ranks differ by more than a critical distance value. Figure 1 shows the results from the Nemenyi post-hoc test for the four different measures used in the experiments for the BRKNN classifier. Each diagram presents an enumerated axis with the average ranks of each technique. The best ranking ones are at the right-most

Table III. Number of times that each feature selection achieved a result better than ($\leq$) the baseline score

Classifier	BR+InfoGain	Copy+InfoGain	LP+InfoGain	MLInfoGain
BR + DecisionTree	42	40	42	42
BR + NaiveBayes	37	30	31	37
BRKNN	39	33	31	41
CC + DecisionTree	43	38	40	40
CC + K-NN	38	37	35	43
CC + NaiveBayes	32	29	30	36
HOMER + K-NN	41	40	42	43
IBLR_ML	38	34	31	37
LP + DecisionTree	42	38	38	41
LP + K-NN	39	39	36	43
LP + NaiveBayes	32	27	30	30
ML-KNN	39	28	27	37
PPT + K-NN	38	33	33	38
RAKEL + K-NN	37	32	31	39
RK + DecisionTree	37	32	33	34
RK + NaiveBayes	37	28	29	34

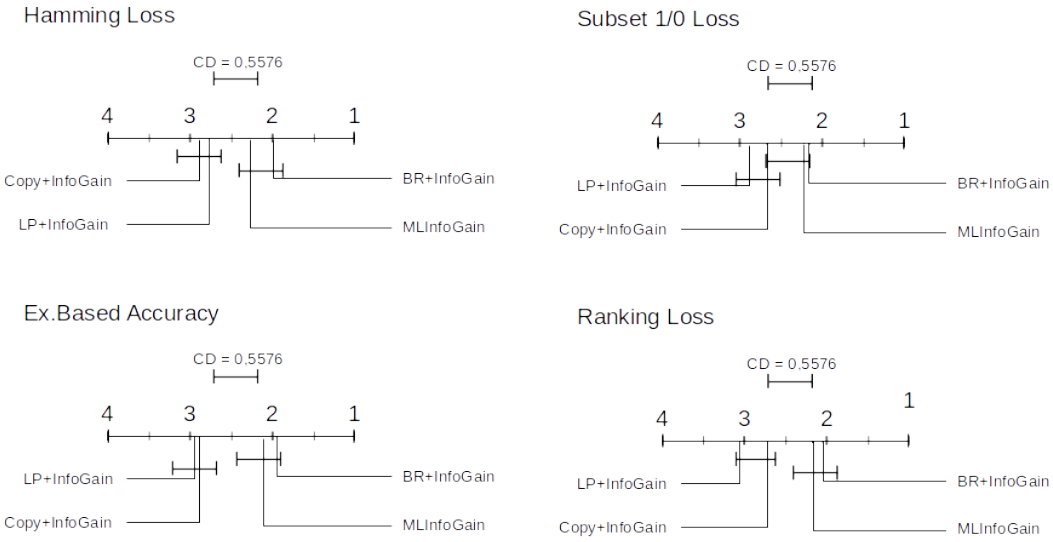

Fig. 1. Critical diagram for each measure in the BRKNN classifier from the Nemenyi post-hoc test at 0.05 significance

side of the diagram. The lines for the average ranks of the algorithms that do not differ significantly (at the significance level of $p = 0.05$) are connected with a line.

The diagrams show that for most measures the MLInfoGain feature selection technique is significantly better than the Copy+InfoGain and LP+InfoGain techniques. However, when comparing MLInfoGain and BR+InfoGain techniques, the diagrams reveal no significant difference.

4.4 Experiments on large multi-label data sets

Most multi-label classification methods either do not scale or have unsatisfactory performance [Tang et al. 2009]. In this section, we report experiments on larger multi-label data sets. We have chosen 11 independently compiled data sets from the Yahoo! directory [Tang et al. 2009] each one with more than 5.000 instances and 30.000 features, being suitable for our scalability experiments.

For these experiments, we employed the BR-KNN classifier, implemented using a single search for k nearest neighbors but at the same time making independent predictions for each label [Sorower 2010]. While BR followed by k-NN has a computational complexity of L times the cost of computing the k nearest instances, where L is the number of labels in the data set, this adaptation runs much more

faster, and is more scalable than the other classification algorithms used in the experiments.

Table IV shows the result of the experiment with larger data sets executed in a similar fashion as the previous one. We used BR+InfoGain and the proposed MLInfoGain techniques with the 10% parameter of selected features. Each row shows the result on a Yahoo data set. Columns "HLoss", "SLoss", "EbAcc" and "RLoss" show the result of the Hamming Loss, Subset 0/1 Loss, Example-based Accuracy (inverted) and Ranking Loss, respectively. Column "Time(s)" shows the total execution time of the experiment (feature selection time + classification time), in seconds. The computer used in the experiments was a AMD FX 8210 8-Core 3.1 Ghz with 8 Gb of RAM and a 64 bit OS.

Table IV. Result of experiments on large data sets with BR-KNN classifier

Data Set	BR+InfoGain 10%					MLInfoGain 10%				
	HLoss	SLoss	EbAcc	RLoss	Time(s)	HLoss	SLoss	EbAcc	RLoss	Time(s)
Arts	0.0595	0.8991	0.8770	0.1941	53,692	0.0617	0.9280	0.9128	0.2093	0,686
Business	0.0267	0.4464	0.3000	0.0745	93,634	0.0270	0.4497	0.3026	0.0767	1,015
Computers	0.0360	0.6497	0.5900	0.1509	186,670	0.0368	0.6439	0.5812	0.1604	1,869
Education	0.0413	0.8771	0.8578	0.1658	142,035	0.0427	0.9192	0.9035	0.1854	1,487
Entertainment	0.0578	0.7621	0.7390	0.1778	125,560	0.0578	0.8113	0.7944	0.1933	1,726
Health	0.0430	0.6890	0.6141	0.1292	110,008	0.0456	0.7299	0.6295	0.1342	1,174
Recreation	0.0559	0.8262	0.8117	0.1990	122,812	0.0584	0.8757	0.8624	0.2328	1,647
Reference	0.0317	0.6342	0.6002	0.2009	133,902	0.0326	0.6839	0.6546	0.2107	1,344
Science	0.0343	0.9054	0.8940	0.2100	120,105	0.0350	0.9456	0.9379	0.2264	1,069
Social	0.0254	0.6204	0.5937	0.1277	334,846	0.0276	0.6849	0.6579	0.1341	3,080
Society	0.0537	0.7762	0.7207	0.1898	215,605	0.0547	0.8075	0.7620	0.1998	2,442

The same non-parametric statistical test used before shows no significant difference between both techniques for the multi-label measures. However, the computational time of BR+InfoGain is much more larger than the MLInfoGain. It takes roughly 100 times more to execute the same experiment with the BR approach. Higher computational time also occurs to the Copy and LP transformation, and they are not reported in this work due to their low performance. It is worth noting that running the same experiment without feature selection is faster than using the BR approach, but slower than using the MLInfoGain. Even though the classification task is accelerated by reducing the number of features in both cases, the feature selection algorithm also counts for the overall time performance.

5. CONCLUSIONS

In this work we have presented an experimental evaluation of various multi-label feature selection methods coupled with different classification techniques and data sets. We have also proposed an adaptation of the information gain feature selection technique to handle multi-label data directly, and performed experimental evaluations to compare it with transformation-based techniques.

Experimental results on a large number of multi-label classification techniques indicate that the proposed multi-label information gain feature selection adaptation achieves a competitive performance against other techniques and outperforms the baseline on most cases. For larger data sets, the proposed technique scales much better than the other feature selection methods.

As future work, we plan to perform a comparative analysis with other multi-label adaptations of feature selection techniques, like multi-label ReliefF [Pupo et al. 2013; Spolaôr et al. 2013] and mutual information [Lee and Kim 2013].

REFERENCES

Boutell, M. R., Luo, J., Shen, X., and Brown, C. M. Learning multi-label scene classification. *Pattern recognition* 37 (9): 1757–1771, 2004.

Chen, W., Yan, J., Zhang, B., Chen, Z., and Yang, Q. Document transformation for multi-label feature selection in text categorization. In *Proc. of the 7th IEEE International Conference on Data Mining*. pp. 451–456, 2007.

Clare, A. and King, R. D. Knowledge discovery in multi-label phenotype data. In *Proc. of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*. pp. 42–53, 2001.

Dash, M. and Liu, H. Feature selection for classification. *Intelligent Data Analysis* vol. 1, pp. 131–156, 1997.

Dembczyński, K., Waegeman, W., Cheng, W., and Hüllermeier, E. On label dependence and loss minimization in multi-label classification. *Machine Learning* 88 (1-2): 5–45, 2012.

Doquire, G. and Verleysen, M. Feature selection for multi-label classification problems. In *Proc. of the 11th Conf. on Artificial neural networks on Advances in computational intelligence*. Springer-Verlag, Spain, pp. 9–16, 2011.

Elisseeff, A. and Weston, J. A kernel method for multi-labelled classification. In *Advances in Neural Information Processing Systems 14*. Vol. 14. pp. 681–687, 2001.

Fayyad, U. M. and Irani, K. B. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proc. of the 13th International Joint Conference on Artificial Intelligence (IJCAI'93)*. pp. 1022–1029, 1993.

Gonçalves, E., Plastino, A., and Freitas, A. A genetic algorithm for optimizing the label ordering in multi-label classifier chains. In *2013 IEEE 25th International Conference on Tools with Artificial Intelligence, Herndon, VA, USA, November 4-6, 2013*. pp. 469–476, 2013.

Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L., editors. *Feature Extraction, Foundations and Applications*. Springer, 2006.

Jungjit, S., Michaelis, M., Freitas, A. A., and Cinatl, J. Two extensions to multi-label correlation-based feature selection: a case study in bioinformatics. In *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*. IEEE, pp. 1519–1524, 2013.

Lee, J. and Kim, D.-W. Feature selection for multi-label classification using multivariate mutual information. *Pattern Recognition Letters* 34 (3): 349–357, 2013.

Madjarov, G., Kocev, D., Gjorgjevikj, D., and Dvzeroski, S. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 2012.

Olsson, J. and Oard, D. W. Combining feature selectors for text classification. In *Proc. of the 15th ACM international conference on Information and knowledge management*. ACM, pp. 798–799, 2006.

Pupo, O. G. R., Morell, C., and Soto, S. V. Relieff-ml: An extension of relieff algorithm to multi-label learning. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Springer, pp. 528–535, 2013.

Read, J. A pruned problem transformation method for multi-label classification. In *Proc. NZ Computer Science Research Student*. pp. 143–150, 2008.

Read, J., Pfahringer, B., Holmes, G., and Frank, E. Classifier chains for multi-label classification. In *Proc. of the European Conf. on Machine Learning and Knowledge Discovery in Databases*. Bled, Slovenia, pp. 254–269, 2009.

Silva, P., Gonçalves, E., Plastino, A., and Freitas, A. Distinct chains for different instances: An effective strategy for multi-label classifier chains. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part II*. pp. 453–468, 2014.

Sorower, M. S. A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis* vol. 63, 2010.

Spolaôr, N., Cherman, E. A., Monard, M. C., and Lee, H. D. A comparison of multi-label feature selection methods using the problem transformation approach. *Electronic Notes in Theoretical Computer Science* vol. 292, pp. 135–151, 2013.

Spolaôr, N., Cherman, E. A., Monard, M. C., and Lee, H. D. Relieff for multi-label feature selection. In *Intelligent Systems (BRACIS), 2013 Brazilian Conference on*. IEEE, pp. 6–11, 2013.

Tang, L., Rajan, S., and Narayanan, V. K. Large scale multi-label classification via metalabeler. In *Proc. of the 18th international conference on World wide web*. ACM, pp. 211–220, 2009.

Trohidis, K., Tsoumakas, G., Kalliris, G., and Vlahavas, I. P. Multi-label classification of music into emotions. In *ISMIR* (2009-12-28), J. P. Bello, E. Chew, and D. Turnbull (Eds.). pp. 325–330, 2008.

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

Tsoumakas, G., Katakis, I., and Vlahavas, I. Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach (Eds.). Springer US, pp. 667–685, 2010.

Tsoumakas, G. and Vlahavas, I. Random k-labelsets: An ensemble method for multilabel classification. In *Proc. of the 18th European conference on Machine Learning*. Warsaw, Poland, pp. 406–417, 2007.

Yang, Y. and Pedersen, J. O. A comparative study on feature selection in text categorization. In *Proc. of the 14th International Conference on Machine Learning*. pp. 412–420, 1997.

Zhang, M.-L., Peña, J. M., and Robles, V. Feature selection for multi-label naive bayes classification. *Information Sciences* 179 (19): 3218–3229, 2009.

Zhang, M.-L. and Zhou, Z.-H. Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition* 40 (7): 2038–2048, 2007.

Zheng, Z., Wu, X., and Srihari, R. Feature selection for text categorization on imbalanced data. *ACM SIGKDD Explorations Newsletter* 6 (1): 80–89, 2004.