Available online at www.sciencedirect.com

SciVerse ScienceDirect

Electronic Notes in Theoretical Computer Science

Electronic Notes in Theoretical Computer Science 292 (2013) 135–151

www.elsevier.com/locate/entcs

A Comparison of Multi-label Feature Selection Methods using the Problem Transformation Approach

Newton Spolaôr[1,a,2] Everton Alvares Cherman[a,3]
Maria Carolina Monard[a,4] Huei Diana Lee[b,5]

[a] *Laboratory of Computational Intelligence*
Institute of Mathematics and Computer Science
University of São Paulo
13560-970 São Carlos, SP, Brazil

[b] *Laboratory of Bioinformatics*
Western Paraná State University
85867-900 Foz do Iguaçu, PR, Brazil

Abstract

Feature selection is an important task in machine learning, which can effectively reduce the dataset dimensionality by removing irrelevant and/or redundant features. Although a large body of research deals with feature selection in single-label data, in which measures have been proposed to filter out irrelevant features, this is not the case for multi-label data. This work proposes multi-label feature selection methods which use the filter approach. To this end, two standard multi-label feature selection approaches, which transform the multi-label data into single-label data, are used. Besides these two problem transformation approaches, we use ReliefF and Information Gain to measure the goodness of features. This gives rise to four multi-label feature selection methods. A thorough experimental evaluation of these methods was carried out on 10 benchmark datasets. Results show that ReliefF is able to select fewer features without diminishing the quality of the classifiers constructed using the features selected.

Keywords: Multi-label learning, Feature Ranking, ReliefF, Information Gain

1 Introduction

Feature Selection (FS) plays an important role in machine learning and data mining, and it is often applied as a data pre-processing step. FS aims to find a small number

[1] This research was supported by the Brazilian Research Council FAPESP. The authors would like to thank the anonymous referees for their insightful comments on this paper. We would also like to thank Victor Augusto Moraes Carvalho and Antonio Rafael Sabino Parmezan for their help in additional analysis.

[2] Email: newtonspolaor@gmail.com

[3] Email: echerman@icmc.usp.br

[4] Email: mcmonard@icmc.usp.br

[5] Email: hueidianalee@gmail.com

doi:10.1016/j.entcs.2013.02.010

of features that describes the dataset as well as the original set of features does [14], providing support to tackle the "*curse of dimensionality*" problem when learning from high-dimensional data. Feature selection can effectively reduce data dimensionality by removing irrelevant and/or redundant features, speeding up learning algorithms and sometimes improving their performance. In fact, various studies show that features can be removed without performance deterioration [22,7,31,24].

Feature selection algorithms evaluate the goodness of features in two main ways: individual evaluation and subset evaluation. On the one hand, individual evaluation is computationally less expensive, as this approach assesses individual features and assigns them weights (ranks) according to their degree of class prediction. To this end, several feature importance measures have been proposed. Nevertheless, the individual feature evaluation is incapable of detecting redundant features as they are likely to have similar rankings. On the other hand, the subset evaluation approach can handle both, feature relevance and feature redundancy. However, unlike individual evaluation, in this approach the evaluation measures are defined against a subset of features, thus showing a high computational cost.

For single-label learning, where each example (or instance) in the dataset is associated with only one class, feature selection has been studied for many years [31]. However, few results in feature selection on multi-label learning have been reported.

Unlike single-label learning, each example in multi-label learning is associated with a subset of labels, *i.e.*, each example can simultaneously belong to multiple classes. In addition, these labels are usually correlated. Multi-label learning is an emerging research topic due to the increasing number of applications where examples are annotated using more than one class, such as bioinformatics [9], emotion analysis [1], semantic annotation of media [29,2] and text mining [3].

This work proposes and experimentally evaluates four multi-label feature selection methods, which use the filter approach. In this approach, the goodness of a feature is evaluated irrespective of any particular classifier. We use the standard multi-label feature selection approach, which consists of first transforming the multi-label data into single-label, which is then used to select features.

We propose the use of ReliefF (RF) and Information Gain (IG) as feature evaluation measures for each label, and the use of two problem transformation approaches [25], Binary Relevance (BR) and Label Powerset (LP), to previously transform the multi-label data into single-label data.

The Binary Relevance approach transforms the multi-label dataset into many single-label datasets, one for each individual label in the multi-labels. One disadvantage of this transformation, is that it does not take into account label dependence, an important aspect in multi-label learning [4]. After this transformation, the contribution of each feature according to each individual label in the multi-labels is measured and the average of the score of all features across all labels is considered. Finally, features with an average score greater than a threshold are selected.

The Label Powerset approach directly transforms the multi-label dataset into one single-label dataset by considering each different combination of labels in the training set as a distinct class value of that single-label dataset, in which any single-

label feature selection method can be directly applied. Moreover, this approach implicitly takes into account label dependence.

The combination of the proposed feature evaluation measures and the problem transformation approaches gives rise to the four multi-label feature selection methods proposed in this work: *RF-BR*, *RF-LP*, *IG-BR* and *IG-LP*.

It should be observed that the first method, *RF-BR*, was initially proposed in [20], although it was evaluated in few datasets. The evaluation of *RF-BR* using more datasets was carried out in [21], where it was experimentally compared with another feature selection method which directly measures the feature goodness in the multi-label dataset. Besides these two pieces of work, we are not aware that ReliefF has been used for multi-label feature selection.

The four methods were experimentally evaluated in 10 benchmark datasets. Results suggest that ReliefF using the *BR* and *LP* approaches, *i.e.*, the methods *RF-BR* and *RF-LP*, are able to select less number of features which describe well the datasets.

The rest of this work is organized as follows: Section 2 briefly presents multi-label learning and Section 3 addresses feature selection for multi-label learning, as well as related work. The filter methods proposed are described in Section 4 and their experimental evaluation in Section 5. Section 6 concludes and highlights future work.

2 Multi-label Learning

This section first presents basic concepts and terminology of multi-label learning, followed by the description of two multi-label problem transformation methods, as well as the multi-label *BRkNN* [23] algorithm and the multi-label classifier evaluation measures used in this work.

2.1 Basic Terminology and Concepts

Let D be a dataset composed of N examples $E_i = (\mathbf{x}_i, Y_i)$, $i = 1..N$. Each example E_i is associated with a feature vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iM})$ described by M features X_j, $j = 1..M$, and a subset of labels $Y_i \subseteq L$, where $L = \{y_1, y_2, \ldots y_q\}$ is the set of q labels. Table 1 shows this representation. In this scenario, the multi-label classification task consists in generating a classifier H which, given an unseen instance $E = (\mathbf{x}, ?)$, is capable of accurately predicting its subset of labels Y, *i.e.*, $H(E) \rightarrow Y$.

Multi-label learning methods can be organized into two main categories: algorithm adaptation and problem transformation [25]. The first one consists of methods which extend specific learning algorithms in order to handle multi-label data directly. The *BRkNN* algorithm used in this work is in this category. The second category is algorithm independent, allowing the use of any state of the art single-label learning algorithm to carry out multi-label learning. It consists of methods which transform the multi-label classification problem into either several binary classification problems, such as the Binary Relevance approach, or one multi-class

Table 1
Multi-label data.

	X_1	X_2	...	X_M	Y
E_1	x_{11}	x_{12}	...	x_{1M}	Y_1
E_2	x_{21}	x_{22}	...	x_{2M}	Y_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
E_N	x_{N1}	x_{N2}	...	x_{NM}	Y_N

classification problem, such as the Label Powerset approach. Recall that single-label learning is called multi-class classification whenever there are more than two class values, and it is called binary classification when the class values are Yes/No. Both approaches, *BR* and *LP*, are used in this work and are described next.

2.2 Binary Relevance

This approach decomposes the multi-label learning task into q independent binary classification problems, one for each label in L. In other words, the multi-label dataset D is first decomposed into q binary datasets D_{y_j}, $j = 1..q$ which are used to construct q independent binary classifiers. In each binary classification problem, examples associated with the corresponding label are regarded as positive and the other examples are regarded as negative. Finally, to classify a new multi-label instance *BR* outputs the aggregation of the labels positively predicted by the q independent binary classifiers. As *BR* scales linearly with size q of the label set L, it is appropriate for a not very large q. However, it suffers from the deficiency that correlation among the labels is not taken into account.

2.3 Label Powerset

This approach transforms the multi-label learning task into a multi-class learning task. To this end, *LP* considers each unique combination of labels in a multi-label dataset as one class value of the correspondent multi-class dataset. In other words, each $E_i = (\mathbf{x}_i, Y_i)$, $i = 1..N$, is transformed into $E_i = (\mathbf{x}_i, l_i)$ where l_i is the atomic label representing a distinct label subset. In this way, unlike *BR*, *LP* takes into account correlation among the labels. However, as the number of class values of the correspondent multi-class dataset is given by the number of distinct label subsets in D, the main drawback of this approach is that some class values in the multi-class dataset may be associated with a very small number of instances, making the multi-class dataset unbalanced.

2.4 BRkNN

BRkNN is an adaptation of the lazy k Nearest Neighbor (*kNN*) algorithm to classify multi-label examples proposed in [23]. Despite the similarities between the al-

gorithms, *BRkNN* is much faster than *kNN* applied according to the *BR* approach, since only one search for the k nearest neighbors is performed by *BRkNN*.

To improve the predictive performance and to tackle directly the multi-label problem, the extensions *BRkNN-a* and *BRkNN-b* were proposed in [23]. Both extensions are based on a label confidence score, which is estimated for each label from the percentage of the k nearest neighbors, containing this label. *BRkNN-a* classifies an unseen example E using the labels with a confidence score greater than 0.5, *i.e.*, labels included in at least half of the k nearest neighbors of E. If no label satisfies this condition, it outputs the label with the greatest confidence score. On the other hand, *BRkNN-b* classifies E with the $[s]$ (nearest integer of s) labels which have the greatest confidence score, where s is the average size of the label sets of the k nearest neighbors of E. In this work, we use the *BRkNN-b* extension.

Lazy algorithms are useful in the evaluation of feature selection methods, since the classifiers built by lazy algorithms are usually susceptible to irrelevant features.

2.5 Evaluation Measures

Unlike single-label classification where the classification of a new instance has only two possible outcomes, correct or incorrect, multi-label classification should also take into account partially correct classification. To this end, some measures, called *example-based*, were specifically defined for multi-label task, while others, called *labeled-base*, are adaptations from the single-label classification problem. A complete discussion on the performance measures for multi-label classification tasks is out of the scope of this work, and can be found in [25]. In what follows, we briefly describe the six multi-label evaluation measures used in this work.

Four of them, *Hamming Loss*, *Subset Accuracy*, *F-Measure* and *Accuracy*, defined by Equations 1 to 4, are example-based measures, where Δ represents the symmetric difference between two sets; Y_i is the set of true labels and Z_i is the set of predicted labels; $I(\text{true}) = 1$ and $I(\text{false}) = 0$.

$$Hamming\,Loss(H, D) = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \Delta Z_i|}{|L|}. \tag{1}$$

$$Subset\,Accuracy(H, D) = \frac{1}{N}\sum_{i=1}^{N} I(Z_i = Y_i). \tag{2}$$

$$F\text{-}Measure(H, D) = \frac{1}{N}\sum_{i=1}^{N}\frac{2|Y_i \cap Z_i|}{|Z_i| + |Y_i|}. \tag{3}$$

$$Accuracy(H, D) = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|}. \tag{4}$$

The remaining two measures, *Macro F-Measure* (F_a) and *Micro F-Measure* (F_b), defined by Equations 5 and 6 respectively, are label-based measures, where $T_{P_{y_i}}$, $F_{P_{y_i}}$, $T_{N_{y_i}}$ and $F_{N_{y_i}}$ represent, respectively, the number of true/false positives/negatives for a label y_j from the set of labels L.

$$F_a(H,D) = \frac{1}{q}\sum_{j=1}^{q}\frac{2T_{P_{y_j}}}{2T_{P_{y_j}} + F_{P_{y_j}} + F_{N_{y_j}}}. \quad (5)$$

$$F_b(H,D) = \frac{2\sum_{j=1}^{q} T_{P_{y_j}}}{2\sum_{j=1}^{q} T_{P_{y_j}} + \sum_{j=1}^{q} F_{P_{y_j}} + \sum_{j=1}^{q} F_{N_{y_j}}}. \quad (6)$$

All these performance measures have values in the interval [0..1]. For *Hamming Loss*, the smaller the value, the better the multi-label classifier performance is, while for the other measures, greater values indicate better performance.

3 Feature Selection

Feature selection searches the feature space $X = \{X_1, X_2, ..., X_M\}$ in order to find a good subset of features $X' \subseteq X$ which describes the dataset as well as the original set of features X does. This section briefly describes basic feature selection approaches and concepts, as well as related work on FS to support multi-label classification.

3.1 Basic Concepts

Considering the interaction with the learning algorithm, there are three feature selection approaches: filter, wrapper and embedded [14].

The filter approach filters out irrelevant features independently of the learning algorithm. It only uses general characteristics of the dataset to select some features and exclude others. Thus, unlike wrappers explained next, filters may not choose the best features for specific learning algorithms. In addition, filters have the advantage of being fast and simple to implement. Moreover, the filter approach is the one more frequently used in research papers related to multi-label feature selection [22].

The wrapper approach requires a specific learning algorithm to evaluate and to determine which features are selected. Although it tends to find features better suited for the specific learning algorithm, it has a high computational cost since it has to call the learning algorithm for each feature set considered.

The embedded approach is the one used by some specific learning algorithms which incorporate feature selection as part of the training process, such as decision trees, to decide in each stage the feature that has the best ability to discriminate among classes.

Feature selection algorithms based on the filter and embedded approaches may return either a subset of selected features or the weights (measuring feature importance) of all features.

Several feature importance measures have been proposed in the literature to evaluate the goodness of features for classification, such as the Fisher score, Chi-square, ReliefF, Gini Index, Information Gain, CFS [31] and Rough Set [19], to name a few.

The measures related to ReliefF and Information Gain used in this work and described in Section 4, for example, enable the search for features which provide a

better separability among classes and a reduction in the uncertainty, respectively. Despite the evaluation of each feature separately, the measure related to the ReliefF algorithm takes into account the effect of interacting features [6].

3.2 Related work

Feature selection has been an active research topic in supervised, semi-supervised and non-supervised machine learning, with a large number of related publications and comprehensive surveys [12,15,31]. However, most of the research related to supervised feature selection has been mainly to support single-label classification, and few results on multi-label classification have been reported. This was confirmed by a systematic review process related to multi-label feature selection we carried out in [22]. Despite the growing interest on this research topic in recent years, less than 60 related papers were found by the systematic review process. Some of these papers are addressed next.

In [18,30] the wrapper approach is directly addressed in multi-label data using evaluation measures and a metaheuristic to search for the best feature subset, while embedded feature selection in decision tree classifiers are proposed in [5,10].

However, most papers propose a previous transformation of multi-label data to single-label data, *i.e.*, to multi-class data or binary data using respectively the Label Powerset or the Binary Relevance approach. Whenever the *BR* approach is used, features are independently selected in each binary data and the results are combined using, for example, an averaging approach.

After problem transformation, the filter approach is usually applied to the single-label data for which many methods have been proposed. To this end, importance measures which do not consider interaction among the features, such as Information Gain [3,27,7] and Chi-square [24], have been the most frequently used. On the other hand, in [20,21] we propose the use of ReliefF, which takes into account feature interaction.

Currently, methods which perform feature selection considering label correlation have also been proposed. The Chi-square measure is applied according to the *LP* approach in [24]. In [30] an evaluation measure which concerns the ranking quality between output labels is used. The Mutual Information measure is applied in [8] according to a modified *LP* approach [16], which also considers label dependence. The Symmetrical Uncertainty measure is extended in [13] to find relationships between all pairs of features and labels. Furthermore, in [11] it is proposed to simultaneously do feature selection and learn the labels correlation during label ranking.

4 Multi-label FS methods Proposed

In this work, we propose four feature selection methods which use ReliefF and Information Gain as feature importance measures. As both importance measures are originally defined for single-label data, the problem transformation approaches *BR* and *LP* are used to transform the multi-label data to single-label data, more specifically, into binary data or multi-class data respectively. Recall that the *BR*

approach does not take into account the correlation among the labels, while the *LP* approach does. The four methods proposed are:

(i) *RF-BR*: ReliefF based on the *BR* approach;

(ii) *RF-LP*: ReliefF based on the *LP* approach;

(iii) *IG-BR*: Information Gain based on the *BR* approach; and

(iv) *IG-LP*: Information Gain based on the *LP* approach.

The first one, *RF-BR*, was initially proposed in [20], where it was evaluated on few multi-label datasets. In [21] *RF-BR* was evaluated using more datasets and was compared with another method, which uses the new Information Gain measure for multi-label data defined in [5]. This new measure was applied directly in the multi-label dataset as a feature importance measure. Besides these two pieces of work, it is not of our knowledge the use of ReliefF for multi-label feature selection.

ReliefF measures the quality of attributes of single-label data. The main advantage of ReliefF over other strictly univariate measures is that it takes into account the effect of interacting features. The basic idea of ReliefF is to reward an attribute for having different values on a pair of similar examples from different classes, and penalize it for having different values on examples from the same class [6,17]. For each feature, ReliefF outputs a value w, ranging from -1 to 1 with large positive w assigned to important features.

Information Gain is a measure of dependence between one feature and the class label often used in papers related to multi-label feature selection [22,5].

The single-label IG of a feature X_j, $j = 1..M$, calculates the difference between the entropy [6] of the dataset D and the weighted sum of the entropy of each subset $D_v \subseteq D$, where D_v is composed by the examples where X_j has the value v. Therefore, if X_j has 10 distinct values in D, the weighted sum would be applied to 10 different D_v datasets. The IG measure is defined by Equation 7.

$$IG\,(D, X_j) = entropy\,(D) - \sum_{v \in X_j} \frac{|D_v|}{|D|} entropy\,(D_v)\,. \tag{7}$$

A high IG value for feature X_j indicates strong dependence between X_j and the class label.

RF-BR and *IG-BR* initially transform the multi-label dataset into q binary datasets. Afterwards, *RF-BR* using ReliefF and *IG-BR* using IG in the conventional way evaluate the set of features $\{X_1, X_2, ...X_M\}$ on each of the q binary datasets. The q measure values of each feature $X_j, j = 1..M$ are then averaged, and the ones with average values greater than or equal to a ReliefF threshold or to an IG threshold respectively, are the features selected. As both methods use the standard multi-label filter approach, which considers each label separately, label correlation is not considered by these two methods.

RF-LP and *IG-LP*, on the other hand, use the feature importance measure directly calculated from the multi-class dataset, which was generated using the Label

[6] The uncertainty inherent to the data.

Powerset approach. Thus, label correlation is considered by these two methods.

5 Experimental Evaluation

The four feature selection methods proposed in this work were implemented using Mulan [7] [26], a package of Java classes for multi-label classification based on Weka [8] [28]. All the reported results were obtained by Mulan using 10-fold cross-validation with paired folds.

5.1 Datasets and Setup

The experiments were carried out using 10 benchmark multi-label datasets obtained from the Mulan's repository [9].

Table 2 shows, for each dataset, the number of examples (N); the number of features (M), where d indicates that the feature values are discrete and n indicates that the feature values are numeric; the number of labels ($|L|$); the Label Cardinality (LC), which is the average number of single-labels associated with each example defined by Equation 8; the Label Density (LD), which is the normalized cardinality defined by Equation 9; and the number of Distinct Combinations (DC) of labels.

$$LC(D) = \frac{1}{|D|}\sum_{i=1}^{|D|}|Y_i|. \tag{8}$$

$$LD(D) = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i|}{|L|}. \tag{9}$$

Table 2
Description of the datasets used in the experiments

Dataset	N	M	|L|	LC	LD	DC
1-*bibtex*	7395	1836 *d*	159	2.40	0.02	2856
2-*cal500*	502	68 *n*	174	26.04	0.15	502
3-*corel16k001*	13766	500 *d*	153	2.86	0.02	4803
4-*corel5k*	5000	499 *d*	374	3.52	0.01	3175
5-*emotions*	593	72 *n*	6	1.87	0.31	27
6-*enron*	1702	1001 *d*	53	3.38	0.06	753
7-*genbase*	662	1186 *d*	27	1.25	0.05	32
8-*medical*	978	1449 *d*	45	1.25	0.03	94
9-*scene*	2407	294 *n*	6	1.07	0.18	15
10-*yeast*	2417	103 *n*	14	4.24	0.30	198

The specific versions of the *BR* and *LP* problem transformation approaches used in this work, as well as the algorithm *BRkNN-b* described in Section 2.4, are the ones available in Mulan. *BRkNN-b* was executed with k=5. Weka provides the implementation of ReliefF and Information Gain, which are used by the proposed

[7] http://mulan.sourceforge.net
[8] http://www.cs.waikato.ac.nz/ml/weka/
[9] http://mulan.sourceforge.net/datasets.html

methods as feature importance measures, and were executed with default parameters. The threshold value for both measures was set as 0.01, which can be considered a conservative threshold.

Initially, for each dataset, the classifier constructed by *BRkNN-b* using all features was evaluated using the multi-label measures described in Section 2.5. These results were used as a *baseline* to evaluate the goodness of the feature selection methods proposed. For each dataset, the four feature selection methods were executed and the classifier constructed with the selected features was evaluated.

Furthermore, for each dataset D, the *Feature Reduction* measure defined by Equation 10 evaluates the average reduction of features obtained by each feature selection method.

$$Feature\,Reduction(D, X') = 100 - \frac{100 \times |X'|}{M}. \tag{10}$$

were $X' \subseteq X$ is the subset of features selected from a dataset D with M examples.

5.2 Results and Discussion

Table 3 shows, for each dataset, the average *Feature Reduction* and its standard deviation in brackets. The five cases where all features showed a lower quality than the threshold used by the feature selection methods, are denoted by −.

Table 3
Average Feature Reduction and standard deviation.

Dataset	*RF-BR*	*RF-LP*	*IG-BR*	*IG-LP*
1-*bibtex*	78.31(0.31)	38.13 (0.68)	-	0.00 (0.00)
2-*cal500*	8.82(0.98)	0.00 (0.00)	-	0.00 (0.00)
3-*corel16k001*	70.10(0.67)	59.66 (0.82)	-	1.04 (0.13)
4-*corel5k*	43.99(1.62)	21.46 (0.76)	-	0.20 (0.21)
5-*emotions*	23.89(1.94)	16.11 (1.17)	18.61 (1.17)	70.83 (2.85)
6-*enron*	1.27(0.30)	0.12 (0.04)	99.55 (0.08)	0.00 (0.00)
7-*genbase*	95.51(0.21)	96.85 (0.07)	97.64 (0.07)	93.43 (0.15)
8-*medical*	86.62(1.06)	94.49 (0.27)	99.52 (0.00)	49.68 (0.90)
9-*scene*	19.15(0.58)	20.54 (0.65)	3.13 (0.75)	4.25 (1.13)
10-*yeast*	40.58(2.74)	6.41 (1.46)	89.22 (1.74)	-

As can be observed, aside from these 5 cases, which were all obtained by Information Gain as a feature importance measure, the average *Feature Reduction* shows a high variation. It goes from 0.00% (all features were considered important by the feature selection method) up to 99.55% for dataset *6-enron* (only 0.45% of the features were considered important by the *IG-BR* method). Moreover, for some datasets such as *6-enron*, there is a high *Feature Reduction* variation, going from 99.55% for *IG-BR*, down to 0.00% for *IG-LP*.

Next, for each dataset, and only using the features selected by each of the four feature selection methods, the correspondent *BRkNN-b* classifier was constructed and evaluated.

Table 4 shows the average multi-label evaluation measures (Section 2.5) and the correspondent standard deviation of these classifiers, as well as their correspondent *baselines*, given by the classifier constructed using all features. Considering the

standard deviation, light gray cells highlight the measures which showed a degradation compared to the correspondent *baseline* measure. As before, cases where all features showed a lower quality than the threshold used by the feature selection methods are denoted by −.

Table 4
Average evaluation measures (and standard deviation) using the *BRkNN-b* classifier.

BRkNN-b

Hamming-Loss

	baseline	*RF-BR*	*RF-LP*	*IG-BR*	*IG-LP*
1	0.02 (0.00)	0.02 (0.00)	0.02 (0.00)	-	0.02 (0.00)
2	0.17 (0.00)	0.17 (0.01)	0.17 (0.00)	-	0.17 (0.00)
3	0.03 (0.00)	0.03 (0.00)	0.03 (0.00)	-	0.03 (0.00)
4	0.01 (0.00)	0.02 (0.00)	0.02 (0.00)	-	0.02 (0.00)
5	0.22 (0.04)	0.22 (0.03)	0.22 (0.03)	0.23 (0.03)	0.24 (0.02)
6	0.06 (0.00)	0.07 (0.00)	0.07 (0.00)	0.07 (0.00)	0.06 (0.00)
7	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
8	0.02 (0.00)	0.01 (0.00)	0.02 (0.00)	0.02 (0.00)	0.02 (0.00)
9	0.11 (0.01)	0.12 (0.01)	0.11 (0.01)	0.12 (0.01)	0.12 (0.01)
10	0.22 (0.01)	0.24 (0.01)	0.22 (0.01)	0.24 (0.01)	-

Subset-Accuracy

	baseline	*RF-BR*	*RF-LP*	*IG-BR*	*IG-LP*
1	0.04 (0.01)	0.07 (0.01)	0.06 (0.01)	-	0.04 (0.01)
2	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	-	0.00 (0.00)
3	0.01 (0.00)	0.00 (0.00)	0.00 (0.00)	-	0.01 (0.00)
4	0.01 (0.00)	0.01 (0.00)	0.00 (0.00)	-	0.00 (0.00)
5	0.26 (0.07)	0.26 (0.07)	0.28 (0.08)	0.25 (0.06)	0.25 (0.06)
6	0.10 (0.03)	0.04 (0.01)	0.10 (0.03)	0.04 (0.02)	0.10 (0.03)
7	0.93 (0.03)	0.94 (0.03)	0.94 (0.02)	0.92 (0.04)	0.93 (0.02)
8	0.52 (0.08)	0.64 (0.05)	0.57 (0.09)	0.52 (0.06)	0.51 (0.07)
9	0.64 (0.03)	0.64 (0.03)	0.64 (0.03)	0.64 (0.03)	0.64 (0.04)
10	0.14 (0.03)	0.12 (0.03)	0.15 (0.03)	0.10 (0.02)	-

F-Measure

	baseline	*RF-BR*	*RF-LP*	*IG-BR*	*IG-LP*
1	0.22 (0.01)	0.27 (0.01)	0.25 (0.01)	-	0.22 (0.01)
2	0.40 (0.01)	0.40 (0.01)	0.40 (0.01)	-	0.40 (0.01)
3	0.15 (0.01)	0.19 (0.00)	0.18 (0.01)	-	0.13 (0.01)
4	0.16 (0.01)	0.12 (0.01)	0.08 (0.01)	-	0.16 (0.01)
5	0.63 (0.08)	0.62 (0.06)	0.63 (0.06)	0.61 (0.07)	0.59 (0.04)
6	0.40 (0.03)	0.43 (0.02)	0.40 (0.03)	0.45 (0.02)	0.40 (0.03)
7	0.97 (0.02)	0.98 (0.01)	0.98 (0.01)	0.97 (0.02)	0.97 (0.02)
8	0.65 (0.07)	0.73 (0.04)	0.70 (0.06)	0.60 (0.06)	0.64 (0.06)
9	0.68 (0.03)	0.67 (0.03)	0.68 (0.03)	0.67 (0.03)	0.67 (0.03)
10	0.61 (0.02)	0.59 (0.02)	0.62 (0.03)	0.58 (0.02)	-

Accuracy

	baseline	*RF-BR*	*RF-LP*	*IG-BR*	*IG-LP*
1	0.16 (0.01)	0.21 (0.01)	0.19 (0.01)	-	0.16 (0.01)
2	0.26 (0.01)	0.26 (0.01)	0.26 (0.01)	-	0.26 (0.01)
3	0.10 (0.00)	0.13 (0.00)	0.12 (0.00)	-	0.09 (0.01)
4	0.10 (0.01)	0.08 (0.01)	0.05 (0.01)	-	0.10 (0.01)
5	0.54 (0.08)	0.53 (0.06)	0.54 (0.06)	0.52 (0.07)	0.50 (0.04)
6	0.31 (0.03)	0.32 (0.02)	0.31 (0.03)	0.34 (0.02)	0.31 (0.03)
7	0.96 (0.02)	0.97 (0.02)	0.97 (0.02)	0.96 (0.02)	0.96 (0.02)
8	0.61 (0.07)	0.71 (0.04)	0.67 (0.07)	0.58 (0.06)	0.61 (0.06)
9	0.67 (0.03)	0.66 (0.03)	0.67 (0.03)	0.66 (0.03)	0.66 (0.03)
10	0.50 (0.02)	0.48 (0.02)	0.51 (0.03)	0.46 (0.02)	-

F_a

	baseline	*RF-BR*	*RF-LP*	*IG-BR*	*IG-LP*
1	0.13 (0.01)	0.16 (0.01)	0.15 (0.01)	-	0.13 (0.01)
2	0.16 (0.01)	0.16 (0.01)	0.16 (0.01)	-	0.16 (0.01)
3	0.05 (0.00)	0.03 (0.00)	0.03 (0.00)	-	0.05 (0.01)
4	0.02 (0.00)	0.02 (0.00)	0.01 (0.00)	-	0.02 (0.00)
5	0.61 (0.08)	0.60 (0.06)	0.61 (0.06)	0.59 (0.06)	0.56 (0.03)
6	0.14 (0.02)	0.12 (0.02)	0.14 (0.02)	0.12 (0.02)	0.14 (0.02)
7	0.77 (0.13)	0.73 (0.14)	0.78 (0.13)	0.74 (0.13)	0.77 (0.13)
8	0.37 (0.06)	0.44 (0.04)	0.41 (0.07)	0.30 (0.03)	0.36 (0.05)
9	0.68 (0.04)	0.67 (0.03)	0.68 (0.03)	0.67 (0.03)	0.67 (0.03)
10	0.42 (0.02)	0.40 (0.02)	0.43 (0.02)	0.39 (0.02)	-

F_b

	baseline	*RF-BR*	*RF-LP*	*IG-BR*	*IG-LP*
1	0.23 (0.01)	0.27 (0.01)	0.25 (0.01)	-	0.23 (0.01)
2	0.40 (0.01)	0.40 (0.01)	0.40 (0.01)	-	0.40 (0.01)
3	0.15 (0.01)	0.20 (0.00)	0.18 (0.01)	-	0.13 (0.01)
4	0.16 (0.01)	0.12 (0.01)	0.08 (0.01)	-	0.16 (0.01)
5	0.64 (0.07)	0.64 (0.06)	0.64 (0.05)	0.62 (0.06)	0.60 (0.03)
6	0.40 (0.02)	0.44 (0.02)	0.40 (0.02)	0.47 (0.01)	0.40 (0.02)
7	0.96 (0.02)	0.96 (0.02)	0.97 (0.02)	0.96 (0.02)	0.96 (0.02)
8	0.64 (0.07)	0.73 (0.04)	0.69 (0.06)	0.61 (0.06)	0.64 (0.06)
9	0.67 (0.03)	0.66 (0.03)	0.67 (0.02)	0.66 (0.03)	0.66 (0.03)
10	0.63 (0.02)	0.61 (0.02)	0.63 (0.02)	0.60 (0.02)	-

From the total of 210 performance measure values tabulated in Table 4 (4 FS methods × 6 performance measures × 10 datasets − 30 empty subset of features), only 20 of them (less than 10%), show a degradation compared to the correspondent *baseline* measure. This can be considered a very good result. In fact, the majority of these 20 cases refers to the *Hamming Loss* and *Subset Accuracy* measures (6 cases each). Recall that *Hamming Loss*, defined by Equation 1, is the relative frequency of correct labels not predicted and correct labels predicted. *Subset Accuracy*, defined by Equation 2, is a very strict evaluation measure as it requires the exact match of the predicted and the true multi-label to maximize its value.

To support the experimental comparison, spider graphs based on all the performance measures used in this work, where greater values indicate better performance, were generated using the R framework [10] . For each dataset, except dataset 2-*cal500*, which shows similar results across the performance measures, Figure 1 shows the performance of the four feature selection methods proposed, as well as their *baseline* (dashed line). Thus, better results are the ones plotted far away from the center and the dashed line.

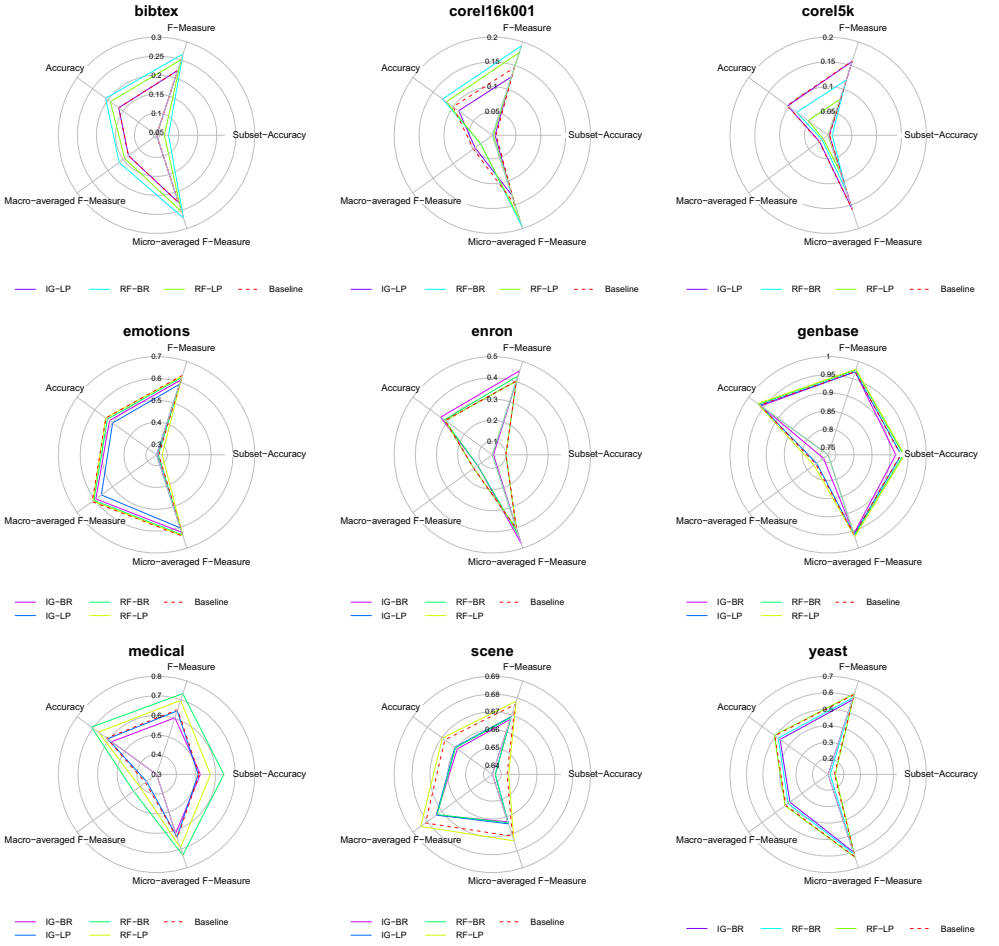

Fig. 1. Graphical evaluation of the datasets using *BRkNN-b* as multi-label learning algorithm.

As can be observed, *RF-LP* shows in general good performance, followed by *RF-BR* and *IG-LP*. On the other hand, *IG-BR* is the one that failed most (4 cases) in finding important features, followed by *IG-LP* (1 case). Furthermore, the few measure values worse than the *baseline* are concentrated in three datasets: 3-*corel16k001*, 4-*corel5k* and 6-*enron*. Observe that these three datasets have a high number of different combinations of labels in common.

[10] http://www.r-project.org

In general, the experimental results suggest a relative superiority of the methods which use ReliefF as importance measure, compared with the ones that use Information Gain. This could be due to the fact that ReliefF considers the interaction among features. Moreover, there is little difference between the measures obtained by the feature selection methods built using the *LP* or *BR* approaches and the same importance measure, *i.e.*, ReliefF or Information Gain. However, it was expected that methods using the *LP* approach would show better results than the ones using the *BR* approach, as *LP* takes into account label interaction. Indeed, it is expected that methods that take the interaction among labels into consideration should lead to better results [4].

Nevertheless, there are two aspects that should be jointly considered when feature selection methods are evaluated: the reduction in the number of features *versus* the performance measure values of the classifier generated with the features selected. This sort of evaluation is better carried out by a graphical analysis. In what follows, this analysis is illustrated in datasets *8-medical* and *1-bibtex*. Graphs for all the datasets used in this work can be found at `http://www.labic.icmc.usp.br/pub/mcmonard/ExperimentalResultsCLEI2012.pdf`

The following figures show in the x-axis the values of the performance measure, as well as their correspondent *baseline*. The y-axis shows the percentage of selected features, which is given by

$$Selected\,Features = 1 - Feature\,Reduction$$

where *Feature Reduction* is defined by Equation 10.

Note that for *Hamming Loss* results nearer to the left-hand bottom corner of the figure are the best, as they show a low *Hamming Loss* value obtained with less features. On the other hand, for the other performance measures the best results are the ones nearer to the right-hand bottom corner, as they show a high measure value obtained with less features.

Figures 2 and 3 show the results for *Hamming Loss*. It can be observed that for dataset *8-medical* — Figure 2 — the best results are obtained by *IG-BR*, followed by *RF-LP* and *RF-BR*. For dataset *1-bibtex* — Figure 3 — the best results are obtained by *RF-BR* followed by *RF-LP*.

Figures 4 and 5 show the results for *F-measure*. For dataset *8-medical* — Figure 4 — the best results are obtained by *RF-BR*, followed by *RF-LP*. Observe that although *IG-BR* uses less features, its *F-measure* value degrades. For dataset *1-bibtex* — Figure 5 — the best results are obtained by *RF-BR*.

For both datasets, similar results are obtained by the other measures. In conclusion, this kind of analysis allowed us to identify the best choice among several feature selection methods for a given dataset.

6 Conclusion

This work proposes and analyses four feature selection methods for multi-label learning, which use the filter approach to select features. To this end, two standard

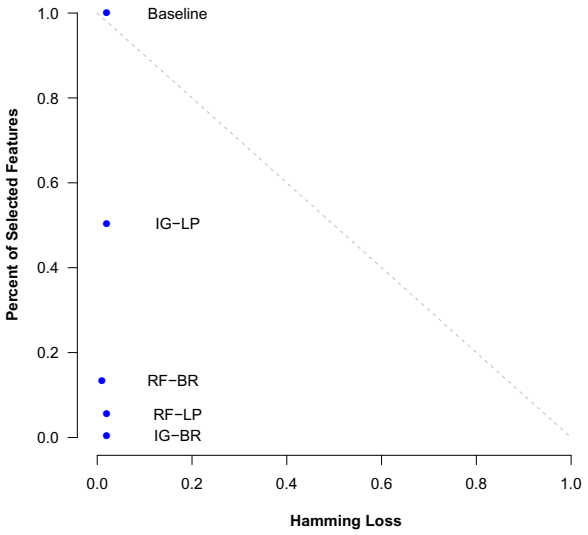

Fig. 2. *8-medical - Hamming Loss*

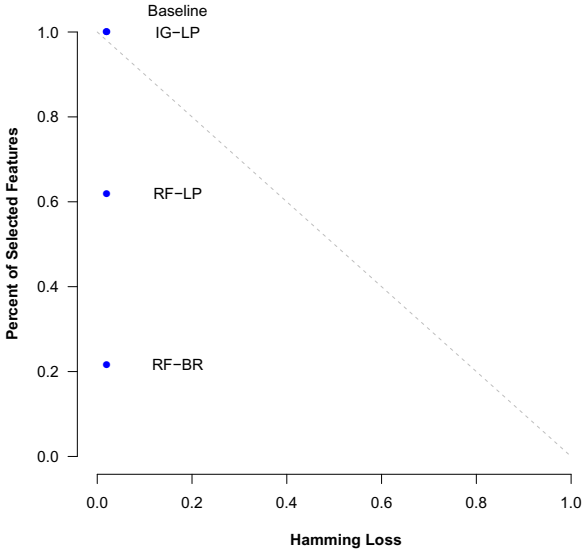

Fig. 3. *1-bibtex - Hamming Loss*

multi-label feature selection approaches are used. The first one transforms the multi-label data to single-label data using the multi-label Binary Relevance problem transformation approach. Afterwards, the contribution of each feature according to each individual label in the multi-labels is measured and the average of the score of

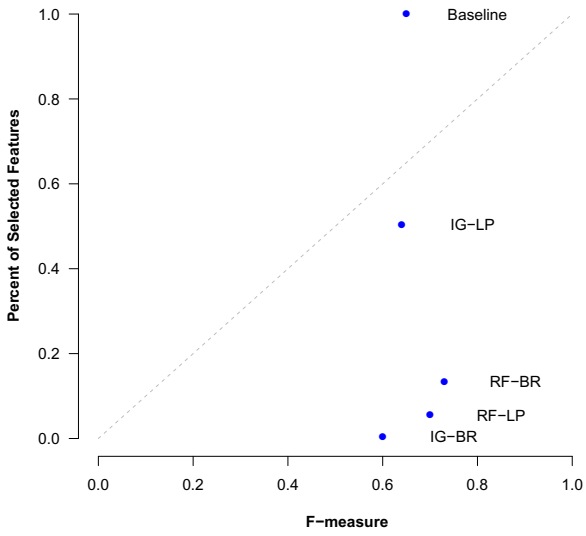

Fig. 4. *8-medical - F-measure*

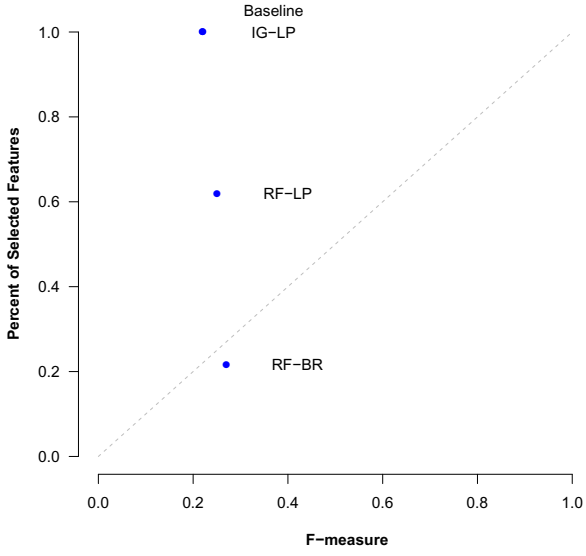

Fig. 5. *1-bibtex - F-measure*

all features across all labels is considered. Finally, features with an average score greater than a threshold are selected. The second method transforms the multi-label data into single-label data using the Label Powerset problem transformation approach, in which any single-label feature selection method can directly be applied.

To measure the goodness of the features we use ReliefF and Information Gain, giving rise to the four feature selection methods: *RF-BR*, *RF-LP*, *IG-BR* and *IG-LP*.

The methods were experimentally evaluated using 10 multi-label benchmark datasets. To this end, we use the *BRkNN-b* multi-label learning algorithm, using the selected features to construct the classifiers for each dataset. In addition, the classifiers constructed by *BRkNN-b* using all features are considered as *baseline* for each dataset.

Results show that the methods which use ReliefF as a feature evaluation measure, select more often smaller number of features than the ones that use Information Gain, with no degradation on the correspondent classifiers. This could be due to the fact that ReliefF considers interactions among features.

Although *BR* and *LP* are classic problem transformation methods in multi-label learning, and RF and IG are classic measures in feature selection, to the best of our knowledge they have not been combined before as proposed in this work. Therefore, the proposed multi-label feature selection methods could be useful for future comparisons with novel FS methods to support multi-label learning.

As future work, we plan to broaden the experimental evaluation of ReliefF using synthetic datasets. Furthermore, the analysis of potential ReliefF extensions for feature selection in order to tackle the multi-label feature selection problem directly, *i.e.*, without any problem transformation, will also be considered.

References

[1] Bhowmick, P. K., A. Basu, P. Mitra and A. Prasad, *Multi-label text classification approach for sentence level news emotion analysis*, in: *International Conference on Pattern Recognition and Machine Intelligence*, 2009, pp. 261–266.

[2] Boutell, M. R., J. Luo, X. Shen and C. M. Brown, *Learning multi-label scene classification*, Pattern Recognition **37** (2004), pp. 1757–1771.

[3] Chen, W., J. Yan, B. Zhang, Z. Chen and Q. Yang, *Document transformation for multi-label feature selection in text categorization*, in: *IEEE International Conference on Data Mining*, 2007, pp. 451–456.

[4] Cherman, E. A., J. Metz and M. C. Monard, *Incorporating label dependency into the binary relevance framework for multi-label classification*, Expert Systems with Applications **39** (2012), pp. 1647–1655.

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

[6] Demšar, J., *Algorithms for subsetting attribute values with Relief*, Machine Learning **78** (2010), pp. 421–428.

[7] Dendamrongvit, S., P. Vateekul and M. Kubat, *Irrelevant attributes and imbalanced classes in multi-label text-categorization domains.*, Intelligent Data Analysis **15** (2011), pp. 843–859.

[8] Doquire, G. and M. Verleysen, *Feature selection for multi-label classification problems*, in: J. Cabestany, I. Rojas and G. Joya, editors, *Advances in Computational Intelligence*, Springer-Verlag/Heidelberg, 2011 pp. 9–16.

[9] Elisseef, A. and J. Weston, *Kernel methods for multi-labelled classification and categorical regression problems*, Advances in Neural Information Processing Systems **14** (2002), pp. 681–687.

[10] Esuli, A., T. Fagni and F. Sebastiani, *Boosting multi-label hierarchical text categorization*, Information Retrieval **11** (2008), pp. 287–313.

[11] Gu, Q., Z. Li and J. Han, *Correlated multi-label feature selection*, in: *ACM International Conference on Information and Knowledge Management*, 2011, pp. 1087–1096.

[12] Guyon, I. and A. Elisseeff, *An introduction to variable and feature selection*, Journal of Machine Learning Research **3** (2003), pp. 1157–1182.

[13] Lastra, G., O. Luaces, J. R. Quevedo and A. Bahamonde, *Graphical feature selection for multilabel classification tasks*, in: *International conference on Advances in intelligent data analysis*, 2011, pp. 246–257.

[14] Liu, H. and H. Motoda, "Computational Methods of Feature Selection," Chapman & Hall/CRC, 2008.

[15] Liu, H. and L. Yu, *Toward integrating feature selection algorithms for classification and clustering*, IEEE Transactions on Knowledge and Data Engineering **17** (2005), pp. 491–502.

[16] Read, J., *A pruned problem transformation method for multi-label classification*, in: *New Zealand Computer Science Research Student Conference*, 2008, pp. 143–150.

[17] Robnik-Sikonja, M. and I. Kononenko, *Theoretical and empirical analysis of ReliefF and RReliefF*, Machine Learning **53** (2003), pp. 23–69.

[18] Shao, H., G. Li, G. Liu and Y. Wang, *Symptom selection for multi-label data of inquiry diagnosis in traditional chinese medicine*, Science China Information Sciences (2012), pp. 1–13.

[19] Slezak, D. and W. Ziarko, *Attribute reduction in the bayesian version of variable precision rough set model*, Electronic Notes in Theoretical Computer Science **82** (2003), pp. 263–273.

[20] Spolaôr, N., E. A. Cherman and M. C. Monard, *Using ReliefF for multi-label feature selection (in portuguese)*, in: *Conferencia Latinoamericana de Informática*, 2011, pp. 960–975.

[21] Spolaôr, N., E. A. Cherman, M. C. Monard and H. D. Lee, *Filter approach feature selection methods to support multi-label learning based on ReliefF and Information Gain (in print)*, in: *Brazilian Symposium on Artificial Intelligence*, 2012, pp. 1–10.

[22] Spolaôr, N., M. C. Monard and H. D. Lee, *A systematic review to identify feature selection publications in multi-labeled data*, ICMC Technical Report Nº 374. 31 pg. (2012), University of São Paulo.

[23] Spyromitros, E., G. Tsoumakas and I. Vlahavas, *An empirical study of lazy multilabel classification algorithms*, in: *Hellenic conference on Artificial Intelligence* (2008), pp. 401–406.

[24] Trohidis, K., G. Tsoumakas, G. Kalliris and I. Vlahavas, *Multi-label classification of music into emotions*, in: *International Conference on Music Information Retrieval*, 2008, pp. 1–6.

[25] Tsoumakas, G., I. Katakis and I. Vlahavas, *Mining multi-label data*, Data Mining and Knowledge Discovery Handbook (2009), pp. 1–19.

[26] Tsoumakas, G., E. Spyromitros-Xioufis, J. Vilcek and I. Vlahavas, *Mulan: A java library for multi-label learning*, Journal of Machine Learning Research **12** (2011), pp. 2411–2414.

[27] Wei, Q., Z. Yang, Z. Junping and Y. Wang, *Semi-supervised multi-label learning algorithm using dependency among labels*, in: *International Conference on Machine Learning and Computing*, 2009, pp. 112–116.

[28] Witten, I. H. and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques," Morgan Kaufmann, 2011.

[29] Worring, M., C. Snoek, O. de Rooij, G. Nguyen and A. Smeulders, *The mediamill semantic video search engine*, in: *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007, pp. 1213–1216.

[30] Zhang, M.-L., J. M. Peña and V. Robles, *Feature selection for multi-label Naïve Bayes classification*, Information Sciences **179** (2009), pp. 3218–3229.

[31] Zhao, Z., F. Morstatter, S. Sharma, S. Alelyani, A. Anand and H. Liu, *Advancing feature selection research - ASU feature selection repository*, Technical Report (2011), Arizona State University.