Using label powerset for multi-label feature selection: an experimental comparison

Newton Spolaôr[1], Huei Diana Lee[1,2,*], Weber Shoity Resende Takaki[1,2], Antonio Rafael Sabino Parmezan[1], Feng Chung Wu[1,2]

[1]Laboratory of Bioinformatics, Western Paraná State University.
Presidente Tancredo Neves Avenue, 6731, ZIP code 85867-900, Foz do Iguaçu, Brazil

[2]Service of Coloproctology, Faculty of Medical Sciences, University of Campinas.
Tessália Vieira de Camargo Street, 126, ZIP code 13083-887, Campinas, Brazil

{newtonspolaor, wufengchung}@gmail.com, huei.lee@unioeste.br

Abstract Feature selection is a task potentially useful to improve data mining, as it can find a subset of the data features that are relevant to the class and support pattern extraction. Many data mining applications, such as image annotation, deal with multi-label data, in which each instance is associated with one or more labels. Feature selection has successfully supported learning by using approaches such as label powerset, which transforms the multi-label data into a format compatible with traditional algorithms. In this work, we combined this approach with 3 algorithms – Correlation-based feature selection, ReliefF and Information Gain. To compare the resulting multi-label feature selection methods, we evaluate the performance of multi-label classifiers built from the features selected in 5 benchmark datasets. As a result, ReliefF highlighted in terms of F-measure, Hamming Loss and Accuracy. Moreover, all methods outperformed a baseline, suggesting their competitiveness.

1 Introduction

The Knowledge Discovery in Databases (KDD) process has been useful to gain understanding of the data and assist decision making [1]. To do so, pre-processing tasks such as Feature Selection (FS) are typically applied before other KDD steps. In particular, FS can be defined as a process of searching for a subset of important features in terms of an importance measure or criterion that reflects relevance and/or non redundancy of features [2]. Afterwards, it removes the remaining fea-

tures. This process leads to a potential reduction of "curse of dimensionality" effects that impair the learning from data.

FS algorithms have traditionally been applied into single-label datasets, in which each instance (or example) is associated with a unique label (target concept). However, the data inherent to some emerging applications, such as emotion analysis and image annotation, includes instances associated with two or more labels, leading to multi-label datasets [3, 4]. Besides dealing with multiple labels simultaneously, feature selection algorithms for multi-label data should take into account the dependency among labels to provide better support for data mining. In fact, considering label dependence for multi-label FS has led to good results, as some recent surveys indicate [5, 6].

In this work, we aim to compare feature selection methods able to explore label dependence. In particular, we choose three traditional algorithms: Correlation-based Feature Selection (CFS), ReliefF (RF) and Information Gain (IG). To make their application within the multi-label scenario possible, we combine the algorithms with the Label Powerset (LP) approach [4], which preserves relations among labels. Afterwards, an experimental evaluation of these methods in five benchmark datasets assesses their ability to assist multi-label learning. To the best of our knowledge, these methods have not been compared.

The rest of this work is organized as follows. Sections 2 and 3 describe multi-label learning and feature selection methods, respectively. Section 4 presents the experimental settings used to obtain the results discussed in Section 5. Section 6 concludes this paper.

2 Multi-label learning

Let D be a dataset composed of N instances $E_i = (x_i, Y_i)$, $i = 1...N$. Each instance E_i in turn is associated with a feature vector $x_i = (x_{i1}, x_{i2}, ..., x_{iM})$ described by M features X_j , $j = 1...M$, and its multi-label Y_i, which consists of a subset of labels $Y_i \subseteq L$, where $L = y_1, y_2, ..., y_q$ 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 that, given a new instance $E = (x, ?)$, is capable of accurately predicting its multi-label Y, *i.e.*, $H(E) \rightarrow Y$.

Table 1 Multi-label data definition.

	X_1	X_2	X_M	Y
$E1$	x_{11}	x_{12}	x_{1M}	Y_1
$E2$	x_{21}	x_{22}	x_{2M}	Y_2
...	...	...	...	...
EM	x_{N1}	x_{N2}	x_{NM}	Y_N

The main difference between multi-label and single-label learning is that the former deals with a set of labels often correlated, whereas the latter considers possible values of the class (labels) that are mutually exclusive. In single-label classification, each instance E_i is associated with only one class value, which in turn is the label y_i contained in the set of labels L, *i.e.*, $y_i \in L$, with $|L| > 1$. If there are two or more possible class values ($|L| > 2$), the problem is named multi-class classification. If the class value is *yes* or *no*, the problem is named binary classification.

2.1 Learning methods

Multi-label learning methods can be organized into two main categories [3, 4]:

- Problem Transformation: transforms the multi-label dataset into one or more single-label datasets. After processing, traditional classification algorithms are used to solve the single-label problem(s) separately. This category is illustrated by the Label Powerset (LP) approach;
- Algorithm Adaptation: learns from multi-label datasets directly, *i.e.*, without transforming them, after adapting specific learning algorithms. A method that exemplifies this category consists in Binary Relevance k Nearest Neighbor (BRkNN) [7].

LP transforms a multi-label dataset into a single-label (multiclass) one by mapping each distinct multi-label Y into a single class value. Although LP can lead to the imbalance problem in multi-class data if the number of distinct multi-labels is high, it partially considers the label dependence by preserving label relations.

In this work, we use LP to convert the multi-label data into a format compatible with the input expected by the single-label feature selection algorithms chosen.

The BRkNN classification method adapts the lazy k nearest neighbor algorithm to efficiently deal with multi-label data, searching for the neighbors only once. This method was extended in [7] to take into account a label confidence value, estimated for each label according to the percentage of the k neighbors that contains this label. In particular, the *BRkNN-b* extension uses a more sophisticated strategy to specify this percentage, which considers the average size of the multi-labels of the neighbors.

We apply *BRkNN-b* to build classifiers from the data described only by the features chosen by each feature selection method, as lazy algorithms are susceptible to irrelevant features. The better the classifier, the better the FS method is. Section 4 indicates the measures considered by us to evaluate the learning performance.

3 Feature selection

Regardless of the multi-label learning approach, any FS method addresses a few relevant issues, such as the importance measure and the interaction with the learning algorithm. In particular, three approaches determine different interactions between a FS method and the learning algorithm: wrapper, embedded and filter [2]. The first two approaches strongly interact with the learning algorithm. On the other hand, filters use general properties of the dataset to remove unimportant features from it, regardless of the learning algorithm. Thus, the features chosen using the filter approach may not be the best ones for a specific learning algorithm, as is the case for the wrapper and embedded approaches.

This work evaluates the combination of three traditional filter methods with the Label Powerset approach to achieve multi-label feature selection. Each method considers a specific importance measure to evaluate features.

3.1 Traditional algorithms

The traditional algorithms considered in this work are Correlation-based Feature Selection (CFS), ReliefF (RF) and Information Gain (IG). In what follows, each algorithm is briefly described.

CFS evaluates the quality of a feature subset X' based on the predictive ability of each feature, *i.e.*, the correlation between the feature and the class, as well as the degree of correlation between features within the subset [8]. Thus, CFS rewards subsets of features highly correlated to the class and with low redundancy.

RF, in turn, rewards a feature X_j, $j = 1 \ldots M$, for having different values on a pair of similar instances (neighbors) from different classes, as well as penalizes it for having different values on similar instances from the same class [9]. Although RF evaluates one feature at a time, it differentiates from strictly univariate measures such as IG due to the consideration of the effect of interacting features, as all features are used to search for neighbors.

Finally, IG evaluates each feature X_j, $j = 1 \ldots M$, according to the dependence between X_j and the class, as defined by Equation 1. To do so, IG calculates the difference between the entropy of the dataset D and the weighted sum of the entropy of each subset $D_v \subseteq D$, where D_v consists in the set of examples where X_j has the value v [10]. Different from CFS and RF, IG requires a previous discretization of numerical features.

$$IG(D, X_j) = entropy(D) - (\sum_v |D_v| \, entropy(D_v) / |D|). \quad (1)$$

It should be emphasized that the CFS algorithm deals with features according to a multivariate perspective, *i.e.*, it searches for subsets of features, evaluating each possible subset at a time. On the other hand, RF and IG follow a univariate perspective, evaluating each feature individually and ranking all of them accordingly. Therefore, in their cases, a given threshold may be established on the number of features to be chosen to specify feature subsets to be submitted for FS assessment. Section 4 reports the setting considered by us to define these subsets.

3.2 Multi-label feature selection

An alternative to apply CFS, RF and IG for a multi-label dataset starts by applying a problem transformation approach, such as Label Powerset, to obtain a single-label dataset. Afterwards, the FS algorithm can be employed in the resulting data.

Related work contains examples of combinations between single-label FS algorithms and problem transformation approaches [5, 6]. In [11], CFS is associated with an approach similar to LP after the use of another data transformation strategy for feature selection. In addition, the feature ranking strategies RF and IG were applied after LP in [12], yielding the methods RF-LP and IG-LP. In this work, we extended the latter idea by experimentally comparing these methods with the combination between Correlation-based Feature Selection and Label Powerset (CFS-LP). It should be emphasized that, different from [11], we focus on only one transformation approach to reduce external influences on the feature evaluation.

4 Experimental setting

The experiments were carried out using 5 benchmark multi-label datasets obtained from the Mulan repository[1]. Table 2 shows, for each dataset, the number of examples (N); the number of nominal (d) and numeric (n) features that sum up to M; the number of labels ($|L|$); the Label Cardinality (LC), which is the average number of single-labels associated with each example; the Label Density (LD), which is the normalized cardinality; and the number of Distinct Combinations of labels (DC).

Table 2 Description of the datasets used in the experiment.

dataset	*M*	*d*	*n*	*|L|*	*LC*	*LD*	*DC*
1-corel5k	5000	499	0	374	3.52	0.01	3175
2-corel16k001	13766	500	0	153	2.86	0.02	4803

[1] http://mulan.sourceforge.net/datasets-mlc.html

3-emotions	593	0	72	6	1.87	0.31	27
4-flags	194	9	10	7	3.39	0.49	54
5-scene	2407	0	294	6	1.07	0.18	15

Each dataset mentioned in Table 2 is submitted to the 10-fold cross-validation strategy. For each dataset, the training fold *i, i = 1, ..., 10*, is then used as the input for a feature selection method. The reduced version of this fold, described only by the chosen features, is employed to build a *BRkNN-b* classifier (*k=10* nearest neighbors). The model in turn is evaluated in the corresponding reduced test fold. After applying the training and testing procedures for all folds in a dataset, different multi-label evaluation measures are averaged across the 10 test folds to calculate the classification performance. The better the *BRkNN-b* performance, the better the FS method ability to support multi-label learning.

In this work, we used code publicly available in the Weka [13] and Mulan [14] frameworks. The algorithms CFS, RF and IG were applied with the parameter settings recommended by Weka. In particular, RF employs the Euclidean and the Overlap dissimilarity measures to find the distance between instances. IG in turn considers the Minimum Description Length discretization technique [15].

From the final feature ranking found by RF-LP and IG-LP in a dataset fold, nine subsets of the best features $X' \subset X$, $|X| = M$, $|X'| = 10\%M, 20\%M, ..., 90\%M$ are specified. On the other hand, CFS-LP, which evaluates subsets of features, already yields a unique subset of the best features $X' \subset X$, $|X'| = h\%M$. It should be emphasized that the h value is found by CFS-LP and is specific for each fold.

In this work, we evaluate the learning performance according to three frequently used evaluation measures described in [4]: Example-based F-measure, Hamming Loss and Accuracy. These measures range in the interval [0,1]. For Hamming Loss, the smaller the value, the better the multi-label classifier is, whereas higher values for the other measures indicate better classifiers.

To verify the competitiveness of CFS-LP, RF-LP and IG-LP, a *baseline* given by a *BRkNN-b* classifier built using all features, *i.e.*, without feature selection, is included in the experimental comparison presented in the next section.

5 Results and discussion

As mentioned, RF-LP and IG-LP yield nine X' subsets with a predefined number of features $|X'| = 10\%M, 20\%M, ..., 90\%M$. On the other hand, CFS-LP selects a single subset X', $|X'| = h\%M$, and specifies the h value for each dataset fold. In what follows, the average h_a values calculated across the 10 folds are giv-

en for each dataset: $h_a = 24\%M$ (corel5k), $h_a = 24\%$ (corel16k), $h_a = 20\%$ (emotions), $h_a = 25\%$ (flags) and $h_a = 40\%$ (scene). As CFS-LP yields a single feature subset, we compared it with each subset derived from RF-LP and IG-LP.

Regarding the evaluation of multi-label classifiers built from the FS outputs, Table 3 summarizes the results by showing the classifier rankings, averaged across the 5 datasets, for each feature subset size $|X'|$. In this table, the best rankings achieved for each measure are highlighted in bold. Note that the lower the F-measure and Accuracy rankings, the better the classifier performances are, whereas higher Hamming Loss rankings indicate better results. Figure 1 shows the number of times that each FS method was highlighted in Table 3. Besides including the ranking value achieved for each subset size and dataset, Tables A.2, A.3 and A.4 in Appendix detail the results in terms of the average learning performance and the corresponding standard deviation.

Table 3 Average ranking (and standard deviation) estimated according to different multi-label evaluation measures.

$\|X'\|$	CFS-LP	RF-LP	IG-LP
	F-measure		
10%M	2.10 (1.02)	**1.90 (0.65)**	2.00 (0.79)
20%M	2.10 (1.02)	2.10 (0.74)	**1.80 (0.84)**
30%M	2.30 (0.97)	**1.60 (0.42)**	2.10 (0.89)
40%M	2.30 (0.97)	**1.80 (0.76)**	1.90 (0.74)
50%M	2.20 (1.10)	**1.80 (0.57)**	2.00 (0.79)
60%M	2.20 (1.10)	**1.70 (0.57)**	2.10 (0.65)
70%M	2.00 (1.00)	2.10 (0.55)	**1.90 (0.74)**
80%M	1.90 (1.02)	2.30 (0.84)	**1.80 (0.57)**
90%M	**1.80 (1.10)**	2.10 (0.65)	2.10 (0.65)
	Hamming Loss		
10%M	**2,20 (0,84)**	1,90 (0,42)	1,90 (0,42)
20%M	**2,00 (0,71)**	**2,00 (0,35)**	**2,00 (0,35)**
30%M	1,60 (0,55)	**2,20 (0,27)**	**2,20 (0,27)**
40%M	1,40 (0,55)	**2,30 (0,27)**	**2,30 (0,27)**
50%M	1,20 (0,45)	**2,40 (0,22)**	**2,40 (0,22)**
60%M	1,40 (0,55)	**2,30 (0,27)**	**2,30 (0,27)**
70%M	1,80 (0,84)	**2,10 (0,42)**	**2,10 (0,42)**
80%M	**2,00 (1,00)**	**2,00 (0,50)**	**2,00 (0,50)**
90%M	**2,00 (1,00)**	**2,00 (0,50)**	**2,00 (0,50)**
	Accuracy		
10%M	2,00 (0,71)	2,10 (0.55)	**1,90 (0,74)**

20%M	2,00 (0,71)	**1.80 (0.45)**	2.20 (0.45)
30%M	2,00 (1,00)	**1.80 (0.57)**	2.20 (0.57)
40%M	2.10 (1.02)	**1.80 (0.76)**	2.10 (0.65)
50%M	2.20 (1.10)	**1.70 (0.45)**	2.10 (0.89)
60%M	2.20 (1.10)	**1.60 (0.65)**	2.20 (0.57)
70%M	**2.00 (1.00)**	**2.00 (0.79)**	**2.00 (0.35)**
80%M	**1.90 (1.02)**	2.20 (0.91)	**1.90 (0.42)**
90%M	**1.80 (1.10)**	2.00 (0.50)	2.20 (0.67)

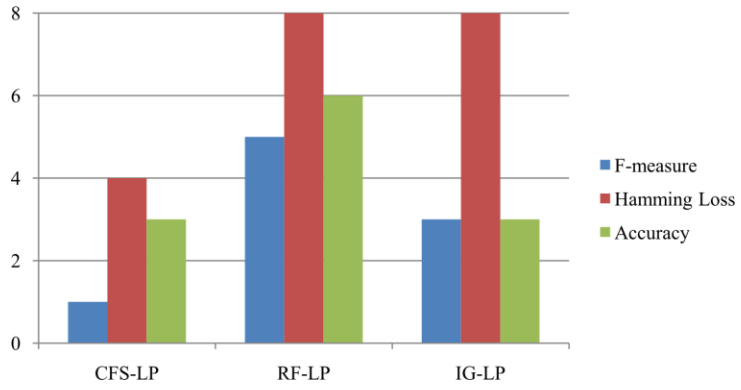

Fig. 1 Number of times that each FS method achieved the highest average ranking for each multi-label evaluation measure.

As Figure 1 shows, regardless of the evaluation measure considered, RF-LP achieved the best average ranking more often. This finding suggests that the feature selection algorithm ReliefF fitted better with the Label Powerset approach for multi-label learning. Moreover, one can note in Table 3 that this method achieved the best F-measure results on average when the number of features was smaller ($|X'|=10\%M$). IG-LP comes next, with similar achievements in terms of the Hamming Loss measure.

On the other hand, CFS-LP obtained few highlighted results. Thus, the evaluation of feature subsets performed by CFS benefited less from label relations, inherent to the data transformed by LP, than the feature ranking algorithms. It should be emphasized, however, that CFS-LP used less features than the remaining algorithms in most of the comparisons conducted, i.e., the size of its feature subsets was usually smaller than the ones defined for RF-LP and IG-LP.

By taking into account Tables A1, A2 and A3 in Appendix to analyze the results achieved for each dataset, one can observe that RF-LP often achieved the best Hamming Loss and Accuracy results in the relatively small dataset emotions.

CFS-LP, in turn, is associated with a similar scenario in the dataset with the largest number of distinct labels, corel16k. The same method also was highlighted in the F-measure and Accuracy values achieved in another relatively large dataset, corel5k. Finally, the IG-LP method obtained good results in flags, the unique dataset with numeric and nominal features.

An issue that arises for future work is to further study the relation between FS algorithms properties and dataset properties. For example, one could investigate if algorithms that consider some relation among features, such as RF-LP and the multivariate CFS-LP, stand out in multi-label datasets with correlated features.

The Appendix tables also contain detailed information regarding the comparison among CFS-LP, RF-LP, IG-LP and the *baseline* – a *BRkNN-b* classifier built using all features. Figures 2, 3 and 4, in turn, focus on the baseline results and the best results achieved by each FS method in each dataset, in terms of predictive performance and size of the feature subset (as a percentage of the number of features *M*). These figures also include a Reference Point (RP) given by the best values represented in each axis. This point is taken into account by the multi-objective optimization community to find a good compromise between potentially conflicting criteria [16, 17].

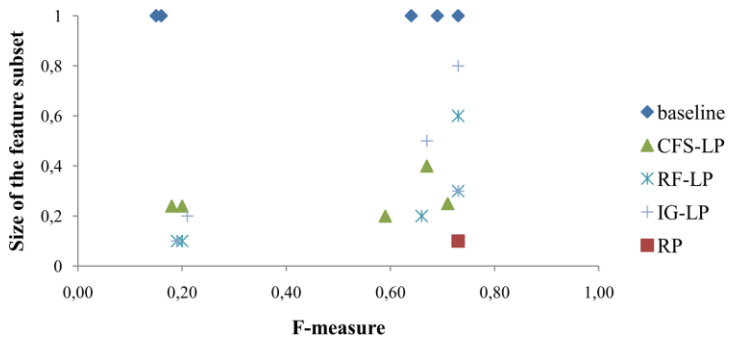

Fig. 2 Baseline results and the best FS results achieved in terms of *BRkNN-b* F-measure performance and size of the feature subset used to build the classifier.

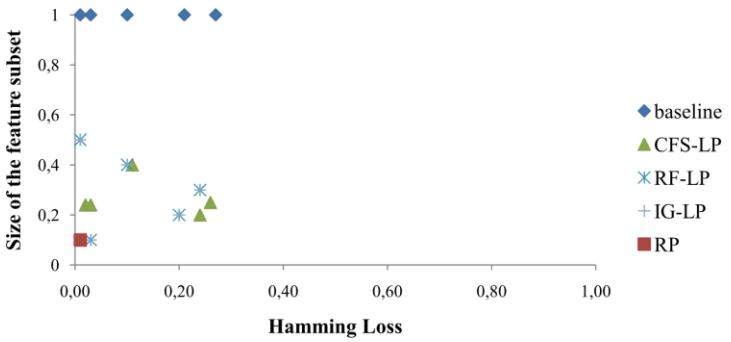

Fig 3. Baseline results and the best FS results achieved in terms of *BRkNN-b* Hamming Loss performance and size of the feature subset used to build the classifier.

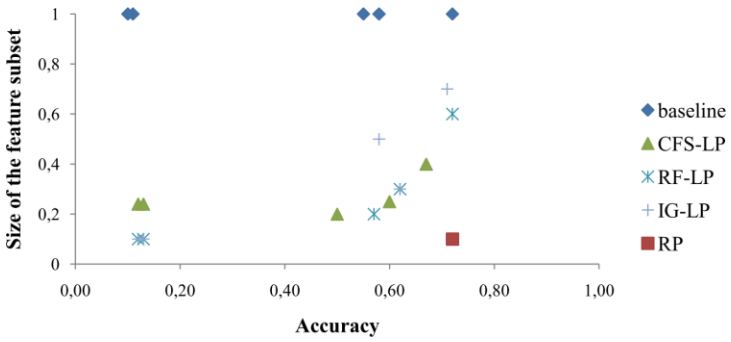

Fig. 4 Baseline results and the best FS results achieved in terms of *BRkNN-b* Accuracy performance and size of the feature subset used to build the classifier.

All FS methods outperformed one or more baseline results in terms of F-measure, Hamming Loss and Accuracy. Moreover, RF-LP and IG-LP always achieved results better than, or equal to, the best baseline result in all evaluation measures. The figures also suggest equilibrium among the FS methods, as they yielded a few compromises similar among themselves and close to RP.

The lazy algorithm *BRkNN-b* is a good candidate to support FS evaluation, as it is sensitive to irrelevant features. However, some findings can be biased towards this algorithm. Future work will deal with this issue by extending the experimental comparison with more multi-label learning algorithms.

6 Conclusion

This work combined the Label Powerset problem transformation approach with 3 traditional algorithms: CFS, ReliefF and Information Gain. The corresponding multi-label feature selection methods (CFS-LP, RF-LP and IG-LP) were compared in terms of their ability to support the building of *BRkNN-b* classifiers in 5 benchmark datasets.

RF-LP was the best method in most of the cases, including scenarios with small feature subsets. This finding suggests that ReliefF fits better with Label Powerset to support multi-label learning with *BRkNN-b*. IG-LP and CFS-LP came next and achieved a few highlighted results. When compared with a classifier built using all features (*baseline*), the classifiers derived from all FS methods were superior in many cases. In particular, feature selection based on ReliefF and Information Gain contributed to the building of competitive classifiers in all evaluation measures.

An idea to improve RF-LP results in datasets with numeric and nominal features, such as flags, is to replace the Euclidean and Overlap dissimilarity measures with the Heterogeneous Value Difference Metric. The latter was already found as a superior setting in BRkNN for multi-label learning [18].

As future work, we plan to investigate other problem transformation approaches that take into account label dependence, such as the pruned problem transformation approach [19]. The adaptation of traditional algorithms for multi-label data, illustrated in references indicated in [5, 6], consists in another alternative, as they can explore label relations.

Acknowledgments **This research was partially supported by the Brazilian National Counsel of Technological and Scientific Development.**

Appendix

Table A1 ***BRkNN-b*** **experimental results according to F-measure.**

dataset	CFS-LP *|X'|=h%M*	RF-LP *|X'|=10%M*	IG-LP *|X'|=10%M*	CFS-LP *|X'|=h%M*	RF-LP *|X'|=20%M*	IG-LP *|X'|=20%M*
corel5k	0.20 (0.01) [1.0]	0.19 (0.01) [2.5]	0.19 (0.01) [2.5]	0.20 (0.01) [1.0]	0.18 (0.01) [3.0]	0.19 (0.01) [2.0]
corel16k	0.18 (0.01) [3.0]	0.20 (0.01) [1.5]	0.20 (0.01) [1.5]	0.18 (0.01) [3.0]	0.20 (0.01) [2.0]	0.21 (0.01) [1.0]
emotions	0.59 (0.05) [3.0]	0.61 (0.04) [1.0]	0.60 (0.04) [2.0]	0.59 (0.05) [3.0]	0.66 (0.06) [1.0]	0.62 (0.03) [2.0]

flags	0.71 (0.03) [2.5]	0.71 (0.03) [2.5]	0.72 (0.05) [1.0]	0.71 (0.03) [2.5]	0.71 (0.04) [2.5]	0.72 (0.03) [1.0]
scene	0.67 (0.04) [1.0]	0.59 (0.04) [2.0]	0.54 (0.03) [3.0]	0.67 (0.04) [1.0]	0.65 (0.04) [2.0]	0.58 (0.03) [3.0]
dataset	CFS-LP *|X'|=h%M*	RF-LP *|X'|=30%M*	IG-LP *|X'|=30%M*	CFS-LP *|X'|=h%M*	RF-LP *|X'|=40%M*	IG-LP *|X'|=40%M*
corel5k	0.20 (0.01) [1.0]	0.17 (0.01) [2.0]	0.16 (0.01) [3.0]	0.20 (0.01) [1.0]	0.16 (0.01) [3.0]	0.18 (0.01) [2.0]
corel16k	0.18 (0.01) [3.0]	0.19 (0.01) [2.0]	0.20 (0.01) [1.0]	0.18 (0.01) [1.5]	0.18 (0.01) [1.5]	0.17 (0.01) [3.0]
emotions	0.59 (0.05) [3.0]	0.64 (0.04) [1.0]	0.63 (0.04) [2.0]	0.59 (0.05) [3.0]	0.65 (0.05) [1.5]	0.65 (0.03) [1.5]
flags	0.71 (0.03) [3.0]	0.73 (0.03) [1.5]	0.73 (0.03) [1.5]	0.71 (0.03) [3.0]	0.72 (0.03) [2.0]	0.73 (0.04) [1.0]
scene	0.67 (0.04) [1.5]	0.67 (0.03) [1.5]	0.66 (0.03) [3.0]	0.67 (0.04) [3.0]	0.70 (0.03) [1.0]	0.68 (0.02) [2.0]
dataset	CFS-LP *|X'|=h%M*	RF-LP *|X'|=50%M*	IG-LP *|X'|=50%M*	CFS-LP *|X'|=h%M*	RF-LP *|X'|=60%M*	IG-LP *|X'|=60%M*
corel5k	0.20 (0.01) [1.0]	0.16 (0.01) [2.5]	0.16 (0.01) [2.5]	0.20 (0.01) [1.0]	0.12 (0.01) [2.5]	0.12 (0.02) [2.5]
corel16k	0.18 (0.01) [1.0]	0.16 (0.01) [2.0]	0.13 (0.02) [3.0]	0.18 (0.01) [1.0]	0.09 (0.01) [2.0]	0.05 (0.01) [3.0]
emotions	0.59 (0.05) [3.0]	0.64 (0.04) [2.0]	0.67 (0.04) [1.0]	0.59 (0.05) [3.0]	0.65 (0.03) [1.5]	0.65 (0.05) [1.5]
flags	0.71 (0.03) [3.0]	0.72 (0.04) [1.5]	0.72 (0.04) [1.5]	0.71 (0.03) [3.0]	0.72 (0.03) [1.5]	0.72 (0.04) [1.5]
scene	0.67 (0.04) [3.0]	0.72 (0.04) [1.0]	0.70 (0.03) [2.0]	0.67 (0.04) [3.0]	0.73 (0.03) [1.0]	0.71 (0.03) [2.0]
dataset	CFS-LP *|X'|=h%M*	RF-LP *|X'|=70%M*	IG-LP *|X'|=70%M*	CFS-LP *|X'|=h%M*	RF-LP *|X'|=80%M*	IG-LP *|X'|=80%M*
corel5k	0.20 (0.01) [1.0]	0.08 (0.01) [3.0]	0.13 (0.02) [2.0]	0.20 (0.01) [1.0]	0.12 (0.02) [2.5]	0.12 (0.02) [2.5]
corel16k	0.18 (0.01) [1.0]	0.05 (0.01) [2.0]	0.04 (0.01) [3.0]	0.18 (0.01) [1.0]	0.04 (0.01) [3.0]	0.06 (0.02) [2.0]
emotions	0.59 (0.05) [3.0]	0.65 (0.04) [2.0]	0.66 (0.05) [1.0]	0.59 (0.05) [3.0]	0.66 (0.03) [1.0]	0.64 (0.03) [2.0]
flags	0.71 (0.03) [2.0]	0.71 (0.04) [2.0]	0.71 (0.03) [2.0]	0.71 (0.03) [1.5]	0.70 (0.03) [3.0]	0.71 (0.04) [1.5]
scene	0.67 (0.04) [3.0]	0.72 (0.03) [1.5]	0.72 (0.02) [1.5]	0.67 (0.04) [3.0]	0.72 (0.03) [2.0]	0.73 (0.02) [1.0]

dataset	CFS-LP *|X'|=h%M*	RF-LP *|X'|=90%M*	IG-LP *|X'|=90%M*	*baseline*
corel5k	0.20 (0.01) [1.0]	0.14 (0.01) [2.5]	0.14 (0.01) [2.5]	0.16 (0.01)

corel16k	0.18 (0.01) [1.0]	0.07 (0.00) [3.0]	0.08 (0.00) [2.0]	0.15 (0.01)
emotions	0.59 (0.05) [3.0]	0.65 (0.04) [1.5]	0.65 (0.04) [1.5]	0.64 (0.05)
flags	0.71 (0.03) [1.0]	0.70 (0.04) [2.0]	0.69 (0.03) [3.0]	0.69 (0.04)
scene	0.67 (0.04) [3.0]	0.72 (0.03) [1.5]	0.72 (0.02) [1.5]	0.73 (0.02)

Table A2 ***BRkNN-b*** **experimental results according to Hamming-Loss.**

	CFS-LP	RF-LP	IG-LP	CFS-LP	RF-LP	IG-LP
dataset	*|X'|=h%M*	*|X'|=10%M*	*|X'|=10%M*	*|X'|=h%M*	*|X'|=20%M*	*|X'|=20%M*
corel5k	0.02 (0.01) [2.0]	0.02 (0.01) [2.0]	0.02 (0.01) [2.0]	0.02 (0.01) [2.0]	0.02 (0.01) [2.0]	0.02 (0.01) [2.0]
corel16k	0.03 (0.01) [2.0]	0.03 (0.01) [2.0]	0.03 (0.01) [2.0]	0.03 (0.01) [2.0]	0.03 (0.01) [2.0]	0.03 (0.01) [2.0]
emotions	0.24 (0.05) [1.0]	0.23 (0.04) [2.5]	0.23 (0.04) [2.5]	0.24 (0.05) [1.0]	0.20 (0.06) [2.5]	0.20 (0.03) [2.5]
Flags	0.26 (0.03) [3.0]	0.27 (0.03) [1.5]	0.27 (0.05) [1.5]	0.26 (0.03) [2.0]	0.26 (0.04) [2.0]	0.26 (0.03) [2.0]
scene	0.11 (0.04) [3.0]	0.14 (0.04) [1.5]	0.14 (0.03) [1.5]	0.11 (0.04) [3.0]	0.12 (0.04) [1.5]	0.12 (0.03) [1.5]
	CFS-LP	RF-LP	IG-LP	CFS-LP	RF-LP	IG-LP
dataset	*|X'|=h%M*	*|X'|=30%M*	*|X'|=30%M*	*|X'|=h%M*	*|X'|=40%M*	*|X'|=40%M*
corel5k	0.02 (0.01) [2.0]	0.02 (0.01) [2.0]	0.02 (0.01) [2.0]	0.02 (0.01) [2.0]	0.02 (0.01) [2.0]	0.02 (0.01) [2.0]
corel16k	0.03 (0.01) [2.0]	0.03 (0.01) [2.0]	0.03 (0.01) [2.0]	0.03 (0.01) [2.0]	0.03 (0.01) [2.0]	0.03 (0.01) [2.0]
emotions	0.24 (0.05) [1.0]	0.21 (0.04) [2.5]	0.21 (0.04) [2.5]	0.24 (0.05) [1.0]	0.21 (0.05) [2.5]	0.21 (0.03) [2.5]
Flags	0.26 (0.03) [1.0]	0.24 (0.03) [2.5]	0.24 (0.03) [2.5]	0.26 (0.03) [1.0]	0.25 (0.03) [2.5]	0.25 (0.04) [2.5]
Scene	0.11 (0.04) [2.0]	0.11 (0.03) [2.0]	0.11 (0.03) [2.0]	0.11 (0.04) [1.0]	0.10 (0.03) [2.5]	0.10 (0.02) [2.5]
	CFS-LP	RF-LP	IG-LP	CFS-LP	RF-LP	IG-LP
dataset	*|X'|=h%M*	*|X'|=50%M*	*|X'|=50%M*	*|X'|=h%M*	*|X'|=60%M*	*|X'|=60%M*
corel5k	0.02 (0.01) [1.0]	0.01 (0.01) [2.5]	0.01 (0.01) [2.5]	0.02 (0.01) [2.0]	0.02 (0.01) [2.0]	0.02 (0.02) [2.0]
corel16k	0.03 (0.01) [2.0]	0.03 (0.01) [2.0]	0.03 (0.02) [2.0]	0.03 (0.01) [2.0]	0.03 (0.01) [2.0]	0.03 (0.01) [2.0]
emotions	0.24 (0.05) [1.0]	0.21 (0.04) [2.5]	0.21 (0.04) [2.5]	0.24 (0.05) [1.0]	0.20 (0.03) [2.5]	0.20 (0.05) [2.5]
flags	0.26 (0.03)	0.25 (0.04)	0.25 (0.04)	0.26 (0.03)	0.25 (0.03)	0.25 (0.04)

	[1.0]	[2.5]	[2.5]	[1.0]	[2.5]	[2.5]
scene	0.11 (0.04) [1.0]	0.10 (0.04) [2.5]	0.10 (0.03) [2.5]	0.11 (0.04) [1.0]	0.10 (0.03) [2.5]	0.10 (0.03) [2.5]

dataset	CFS-LP *|X'|=h%M*	RF-LP *|X'|=70%M*	IG-LP *|X'|=70%M*	CFS-LP *|X'|=h%M*	RF-LP *|X'|=80%M*	IG-LP *|X'|=80%M*
corel5k	0.02 (0.01) [2.0]	0.02 (0.01) [2.0]	0.02 (0.02) [2.0]	0.02 (0.01) [2.0]	0.02 (0.02) [2.0]	0.02 (0.02) [2.0]
corel16k	0.03 (0.01) [3.0]	0.04 (0.01) [1.5]	0.04 (0.01) [1.5]	0.03 (0.01) [3.0]	0.04 (0.01) [1.5]	0.04 (0.02) [1.5]
emotions	0.24 (0.05) [1.0]	0.20 (0.04) [2.5]	0.20 (0.05) [2.5]	0.24 (0.05) [1.0]	0.20 (0.03) [2.5]	0.20 (0.03) [2.5]
flags	0.26 (0.03) [2.0]	0.26 (0.04) [2.0]	0.26 (0.03) [2.0]	0.26 (0.03) [3.0]	0.27 (0.03) [1.5]	0.27 (0.04) [1.5]
scene	0.11 (0.04) [1.0]	0.10 (0.03) [2.5]	0.10 (0.02) [2.5]	0.11 (0.04) [1.0]	0.10 (0.03) [2.5]	0.10 (0.02) [2.5]

dataset	CFS-LP *|X'|=h%M*	RF-LP *|X'|=90%M*	IG-LP *|X'|=90%M*	*baseline*
corel5k	0.02 (0.01) [2.0]	0.02 (0.01) [2.0]	0.02 (0.01) [2.0]	0.01 (0.01)
corel16k	0.03 (0.01) [3.0]	0.04 (0.00) [1.5]	0.04 (0.00) [1.5]	0.03 (0.01)
emotions	0.24 (0.05) [1.0]	0.21 (0.04) [2.5]	0.21 (0.04) [2.5]	0.21 (0.05)
flags	0.26 (0.03) [3.0]	0.27 (0.04) [1.5]	0.27 (0.03) [1.5]	0.27 (0.04)
scene	0.11 (0.04) [1.0]	0.10 (0.03) [2.5]	0.10 (0.02) [2.5]	0.10 (0.02)

Table A3 *BRkNN-b* experimental results according to Accuracy.

dataset	CFS-LP *|X'|=h%M*	RF-LP *|X'|=10%M*	IG-LP *|X'|=10%M*	CFS-LP *|X'|=h%M*	RF-LP *|X'|=20%M*	IG-LP *|X'|=20%M*
corel5k	0.12 (0.01) [2.0]	0.12 (0.00) [2.0]	0.12 (0.00) [2.0]	0.12 (0.01) [2.0]	0.12 (0.01) [2.0]	0.12 (0.01) [2.0]
corel16k	0.13 (0.00) [2.0]	0.13 (0.00) [2.0]	0.13 (0.00) [2.0]	0.13 (0.00) [2.0]	0.13 (0.00) [2.0]	0.13 (0.00) [2.0]
emotions	0.50 (0.05) [3.0]	0.51 (0.04) [1.5]	0.51 (0.04) [1.5]	0.50 (0.05) [3.0]	0.57 (0.05) [1.0]	0.53 (0.04) [2.0]
flags	0.60 (0.03) [2.0]	0.58 (0.04) [3.0]	0.61 (0.06) [1.0]	0.60 (0.03) [2.0]	0.60 (0.06) [2.0]	0.60 (0.04) [2.0]
scene	0.67 (0.04) [1.0]	0.58 (0.04) [2.0]	0.53 (0.03) [3.0]	0.67 (0.04) [1.0]	0.64 (0.04) [2.0]	0.57 (0.03) [3.0]

dataset	CFS-LP	RF-LP	IG-LP	CFS-LP	RF-LP	IG-LP

	|X'|=h%M	*|X'|=30%M*	*|X'|=30%M*	*|X'|=h%M*	*|X'|=40%M*	*|X'|=40%M*
corel5k	0.12 (0.01) [1.0]	0.11 (0.01) [2.5]	0.11 (0.01) [2.5]	0.12 (0.01) [1.0]	0.10 (0.01) [3.0]	0.11 (0.01) [2.0]
corel16k	0.13 (0.00) [2.0]	0.13 (0.00) [2.0]	0.13 (0.01) [2.0]	0.13 (0.00) [1.0]	0.12 (0.00) [2.0]	0.11 (0.01) [3.0]
emotions	0.50 (0.05) [3.0]	0.55 (0.03) [1.0]	0.54 (0.04) [2.0]	0.50 (0.05) [3.0]	0.56 (0.05) [1.5]	0.56 (0.03) [1.5]
flags	0.60 (0.03) [3.0]	0.62 (0.04) [1.5]	0.62 (0.04) [1.5]	0.60 (0.03) [3.0]	0.62 (0.04) [1.5]	0.62 (0.05) [1.5]
scene	0.67 (0.04) [1.0]	0.66 (0.03) [2.0]	0.65 (0.03) [3.0]	0.67 (0.04) [2.5]	0.69 (0.03) [1.0]	0.67 (0.02) [2.5]

dataset	CFS-LP *|X'|=h%M*	RF-LP *|X'|=50%M*	IG-LP *|X'|=50%M*	CFS-LP *|X'|=h%M*	RF-LP *|X'|=60%M*	IG-LP *|X'|=60%M*
corel5k	0.12 (0.01) [1.0]	0.11 (0.01) [2.0]	0.10 (0.01) [3.0]	0.12 (0.01) [1.0]	0.08 (0.01) [2.5]	0.08 (0.01) [2.5]
corel16k	0.13 (0.00) [1.0]	0.10 (0.01) [2.0]	0.08 (0.01) [3.0]	0.13 (0.00) [1.0]	0.06 (0.00) [2.0]	0.03 (0.01) [3.0]
emotions	0.50 (0.05) [3.0]	0.55 (0.05) [2.0]	0.58 (0.04) [1.0]	0.50 (0.05) [3.0]	0.57 (0.03) [1.0]	0.56 (0.06) [2.0]
flags	0.60 (0.03) [3.0]	0.61 (0.05) [1.5]	0.61 (0.05) [1.5]	0.60 (0.03) [3.0]	0.61 (0.05) [1.5]	0.61 (0.05) [1.5]
scene	0.67 (0.04) [3.0]	0.71 (0.04) [1.0]	0.69 (0.03) [2.0]	0.67 (0.04) [3.0]	0.72 (0.03) [1.0]	0.70 (0.03) [2.0]

dataset	CFS-LP *|X'|=h%M*	RF-LP *|X'|=70%M*	IG-LP *|X'|=70%M*	CFS-LP *|X'|=h%M*	RF-LP *|X'|=80%M*	IG-LP *|X'|=80%M*
corel5k	0.12 (0.01) [1.0]	0.05 (0.01) [3.0]	0.08 (0.01) [2.0]	0.12 (0.01) [1.0]	0.08 (0.01) [2.5]	0.08 (0.02) [2.5]
corel16k	0.13 (0.00) [1.0]	0.03 (0.00) [2.5]	0.03 (0.00) [2.5]	0.13 (0.00) [1.0]	0.03 (0.01) [3.0]	0.04 (0.01) [2.0]
emotions	0.50 (0.05) [3.0]	0.57 (0.04) [1.0]	0.56 (0.05) [2.0]	0.50 (0.05) [3.0]	0.57 (0.03) [1.0]	0.55 (0.03) [2.0]
flags	0.60 (0.03) [2.0]	0.60 (0.04) [2.0]	0.60 (0.04) [2.0]	0.60 (0.03) [1.5]	0.58 (0.04) [3.0]	0.60 (0.04) [1.5]
scene	0.67 (0.04) [3.0]	0.71 (0.03) [1.5]	0.71 (0.02) [1.5]	0.67 (0.04) [3.0]	0.71 (0.03) [1.5]	0.71 (0.02) [1.5]

dataset	CFS-LP *|X'|=h%M*	RF-LP *|X'|=90%M*	IG-LP *|X'|=90%M*	*baseline*
corel5k	0.12 (0.01) [1.0]	0.09 (0.01) [2.5]	0.09 (0.01) [2.5]	0.11 (0.01)
corel16k	0.13 (0.00) [1.0]	0.05 (0.00) [2.5]	0.05 (0.00) [2.5]	0.10 (0.01)
emotions	0.50 (0.05) [3.0]	0.56 (0.04) [1.5]	0.56 (0.05) [1.5]	0.55 (0.05)
flags	0.60 (0.03) [1.0]	0.58 (0.05) [2.0]	0.57 (0.04) [3.0]	0.58 (0.04)

	0.67 (0.04)	0.71 (0.02)	0.71 (0.02)	
scene	[3.0]	[1.5]	[1.5]	0.72 (0.02)

References

[1] Han, J., Kamber, M.: Data mining: concepts and techniques. Morgan Kaufmann, 3 edn. (2011)

[2] Liu, H., Motoda, H.: Computational Methods of Feature Selection. Chapman & Hall/CRC (2007)

[3] Zhang, M.L., Zhou, Z.H.: A review on multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering 26(8), 1819–1837 (2014)

[4] Tsoumakas, G., Katakis, I., Vlahavas, I.P.: Mining multi-label data. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 667–685. Springer US (2010)

[5] Spolaôr, N., Monard, M.C., Tsoumakas, G., Lee, H.D.: A systematic review of multi-label feature selection and a new method based on label construction. Neurocomputing 180, 3–15 (2016)

[6] Spolaôr, N., Lee, H.D., Takaki, W.S.R., Wu, F.C.: Feature selection for multi-label learning: A systematic literature review and some experimental evaluations. International Journal of Computational Intelligence Systems 8(sup2), 3–15 (2015)

[7] Spyromitros, E., Tsoumakas, G., Vlahavas, I.: An empirical study of lazy multilabel classification algorithms. In: Hellenic conference on Artificial Intelligence. pp. 401–406. Springer-Verlag (2008)

[8] Hall, M.A.: Correlation-based feature selection for discrete and numeric class machine learning. In: International Conference on Machine Learning. pp. 359–366 (2000)

[9] Robnik-Šikonja, M., Kononenko, I.: Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning 53(1–2), 23–69 (2003)

[10] QuinLan, J.: C4.5: Programs for Machine Learning. Morgan Kaufmann (1993)

[11] Zdravevski, E., Lameski, P., Kulakov, A., Gjorgjevikj, D.: Feature selection and allocation to diverse subsets for multi-label learning problems with large datasets. In: Federated Conference on Computer Science and Information Systems. pp. 387–394 (2014)

[12] Spolaôr, N., Cherman, E.A., Monard, M.C., Lee, H.D.: A comparison of multi-label feature selection methods using the problem transformation approach. Electronic Notes in Theoretical Computer Science 292, 135–151 (2013)

[13] Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 3 edn. (2011)

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

[15] Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuous-valued attributes for classification learning. In: International Joint Conference on Artificial Intelligence. pp. 1022–1029 (1993)

[16] Deb, K.: Multi-objective evolutionary algorithms: Introducing bias among pareto-optimal solutions. Evolutionary Computation 8, 173–195 (1999)

[17] Zeleny, M.: An introduction to multiobjetive optimization. In: Cochrane, J.L., Zeleny, M. (eds.) Multiple criteria decision making, pp. 262–301. Univ. of South Carolina Press (1973)

[18] Reis, D.M.D., Cherman, E.A., Spolaôr, N., Monard, M.C.: Extensions of the multi-label machine learning algorithm BRkNN (in Portuguese). In: ENIA. pp. 1–12 (2012)

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