Filter Approach Feature Selection Methods to Support Multi-label Learning Based on ReliefF and Information Gain

Newton Spolaôr[1,2], Everton Alvares Cherman[1], Maria Carolina Monard[1], and Huei Diana Lee[2]

[1] Laboratory of Computational Intelligence
Institute of Mathematics and Computer Science
University of São Paulo
São Carlos, Brazil
[2] Laboratory of Bioinformatics
Western Paraná State University
Foz do Iguaçu, Brazil
newtonspolaor@gmail.com,
{echerman,mcmonard}@icmc.usp.br,hueidianalee@gmail.br

Abstract. In multi-label learning, each example in the dataset is associated with a set of labels, and the task of the generated classifier is to predict the label set of unseen examples. Feature selection is an important task in machine learning, which aims to find a small number of features that describes the dataset as well as, or even better, than the original set of features does. This can be achieved by removing irrelevant and/or redundant features according to some importance criterion. Although effective feature selection methods to support classification for single-label data are abound, this is not the case for multi-label data. This work proposes two multi-label feature selection methods which use the filter approach. This approach evaluates statistics of the data independently of any particular classifier. To this end, ReliefF, a single-label feature selection method and an adaptation of the Information Gain measure for multi-label data are used to find the features that should be selected. Both methods were experimentally evaluated in ten benchmark datasets, taking into account the reduction in the number of features as well as the quality of the generated classifiers, showing promising results.

1 Introduction

Machine Learning (ML), which has significant overlapping with data mining, pattern recognition and parts of statistics, is an important field of Artificial Intelligence. ML deals with the fundamental problem of using a dataset to reproduce the process that generated the data.

Multi-label learning deals with the classification problem where each example (or instance) in the training dataset is associated with a set of labels, *i.e.* each example can belong to multiple different classes simultaneously. Multi-label

learning is an emerging research topic due to the increasing number of applications where examples are annotated with more than one label. Multi-label classification has been used in applications such as semantic annotation of video and image, bioinformatics, text categorization and categorization of music into emotions [15].

The task of a multi-label classifier is to predict the label set of unseen examples. Thus, multi-label learning is more general than single-label learning, in which each example in the training dataset is associated with only one class, which can assume several values. Whenever there are more than two class values in single-label learning, it is called multi-class classification. Case the class value is Yes/No, it is called binary classification. In fact, the main difference between multi-label and single-label learning is that classes in multi-label learning are often correlated while the class values in single-label learning are mutually exclusive.

Several approaches have been proposed for multi-label learning, which are well described in [15], where the existing methods for multi-label classification are divided into two main categories: problem transformation and algorithm adaptation. The first category considers methods which transform the multi-label classification problem into either one multi-class classification problem or several binary classification problems. Thus, state of the art algorithms such as SVM can then be used directly. The second category consists of methods that extend specific algorithms such that they can handle multi-label data directly.

Similarly to other data mining and machine learning tasks, multi-label learning also experiences the *curse of dimensionality*, which may cause problems when learning from high-dimensional data. Dimensionality reduction can be tackled, among others, through Feature Selection (FS), which aims to find a small number of features that describes the dataset as well as, or even better, than the original set of features does [8]. This can be achieved by removing irrelevant and/or redundant features according to some importance criterion. Although effective feature selection methods to support classification for single-label data have been extensively studied for many years, few results on multi-label dimensionality reduction have been reported.

This work proposes two multi-label feature selection methods which use the filter approach. This approach evaluates statistics of the data irrespective of any particular classifier. The first method uses the standard approach, which consists in measuring the contribution of each feature according to each label. Afterwards, the average of the score of each feature across all labels is considered, and features with an averaged score greater than a threshold are selected. To this end, ReliefF, a single-label feature selection method is used. Although this approach to multi-label feature selection is standard, to the best of our knowledge this is the first time that ReliefF is used for this purpose. However, this approach does not consider label correlations. The second method uses an adaptation of the Information Gain (IG) measure for multi-label data, and features which have an IG greater than a threshold are the ones selected.

Both methods were experimentally evaluated in ten benchmark datasets, taking into account the reduction in the number of features as well as the quality of the generated multi-label classifiers, showing promising results.

The rest of this paper 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, which is followed by the conclusions and future work in Section 6.

2 Multi-label Classification

This section presents basic concepts and terminology of multi-label learning, as well as the Binary Relevance multi-label transformation approach 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 $L = \{y_1, y_2, ...y_q\}$ is the set of q possible labels. Table 1 shows this representation. In this scenario, the multi-label classification task consists in generating a classifier H which, given an unknown instance $E = (\mathbf{x}, ?)$, is capable of accurately predicting its subset of labels Y, *i.e.*, $H(E) \rightarrow Y$.

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

As already mentioned, methods for multi-label classification can be divided into two main categories: problem transformation and problem adaptation. The first category considers methods which transform the multi-label classification problem into either one multi-class classification, such as the Label Powerset (LP) approach, in which each unique set of labels in the training set is considered as a class value, or several binary classification problems, such as the Binary Relevance approach described next does. In both cases, multi-class or binary, respectively, state of the art algorithms can then be used directly. The second category consists of methods that extend specific algorithms such that they can handle multi-label data directly [15].

2.2 Binary Relevance

This problem transformation approach decomposes a multi-label classification problem into several distinct binary classification problems, one for each label in the set of labels L with $|L| = q$. The Binary Relevance (BR) approach initially transforms the original training dataset into q binary datasets D_{y_j}, $j = 1..q$, where each D_{y_j} contains all examples of the original dataset, but with a single positive or negative label related to the single label y_j according to the true label subset associated with the example, *i.e.*, positive if the label set contains label y_j and negative otherwise. The other labels $(y_k, k \neq j)$ are not included in D_{y_j}. After the data is transformed, a set of q binary classifiers $H_j(E), j = 1..q$ is constructed using the correspondent training dataset D_{y_j}. In other words, the BR approach initially constructs a set of q classifiers — Equation 1:

$$H_{BR} = \{C_{y_j}((\mathbf{x}, y_j)) \rightarrow \lambda_j \in \{0, 1\} | y_j \in L : j = 1..q\} \tag{1}$$

To classify a new multi-label instance, the algorithm outputs the aggregation of the labels positively predicted by all the q independent binary classifiers.

An advantage of the BR approach is its low computational complexity compared with other multi-label methods. For a constant number of examples, BR scales linearly with size q of the label set L, which makes it appropriate for not very large q. For large numbers of labels some divide-and-conquer methods have been proposed to organize labels into a tree-shaped hierarchy where it is possible to deal with a much smaller set of labels compared to q. A disadvantage of the standard BR approach is that it completely ignores any label relationships. However, two successful methods that enable the binary classifiers to discover existing label dependency by themselves have already been proposed [2, 10].

3 Feature Selection for Multi-label Classification

Approaches to feature selection are addressed next, as well as related work in feature selection for multi-label classification.

3.1 Feature Selection Approaches

FS methods can be classified into three main categories (wrapper, embedded or filter) according to the interaction with the learning algorithm [8].

The wrapper approach uses the learning algorithm itself as a black box to evaluate candidate subsets of features, repeating the process on each feature subset until a stopping criterion is met. Thus, wrapper methods take into consideration all the important characteristics of the learning algorithm in the final decision of the feature selection process. However, its computational cost could be very high. Similarly to wrappers, FS performed by embedded methods is linked with the learning algorithm itself. However, in this case this link is stronger than in wrappers, since the FS process is included in the classifier construction. A typical example of embedded methods for feature subset selection is decision

trees [9]. Unlike these two approaches, filter methods perform a separate process that does not interact with and is independent from the learning algorithm itself. The basic idea of filters is to use general characteristics of data to select the relevant features according to these characteristics, before the construction of the classifier takes place. An advantage of filters is the fact that they are fast and simple to use.

3.2 Related Work

Although effective FS methods to support classification for single-label data have been extensively studied for many years, few results on multi-label dimensionality reduction have been reported. A systematic review process, a method to support bibliographic reviews, related to multi-label FS was carried out in [12]. Results show the findings of less than 50 related papers, as well as a growing interest in the subject in recent years.

Some papers use the wrapper approach addressing directly the multi-label data [17]. However, most papers consider the previous transformation of multi-label data to multi-class data (using LP) or binary data (using BR). Afterwards, the filter approach is used in the transformed data. To this end, measures related to Information Gain [1, 16], mutual information [5], chi-square [14] and others are used. Whenever the BR approach is used, each label is considered separately and the results are combined using, for example, an averaging approach. Embedded feature selection is used in [3, 6]. In addition, in [7] it is proposed to learn the label correlation and do FS simultaneously.

4 Multi-label Feature Selection Methods Proposed

The first method, named RF, was initially proposed in [13], where it was evaluated but on few multi-label datasets. RF uses ReliefF, an algorithm which 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 attributes. The idea of ReliefF and its derivatives is to reward an attribute for having different values on a pair of nearest examples from different classes, and penalize it for having different values on examples from the same class [4, 11]. For each feature, ReliefF outputs a value w, ranging from -1 to 1 with large positive w assigned to important features.

Initially, RF uses the BR approach to transform the multi-label training dataset into q binary datasets and ReliefF is used in the conventional way to evaluate the set of features $\{X_1, X_2, ..., X_M\}$ on each of the q binary datasets. The q ReliefF measure values of each feature $X_i, i = 1..M$, are then averaged and the ones with values greater than or equal to a threshold are selected. However, apart from the use of ReliefF, of which we are not aware it has been used before for multi-label feature selection, RF uses the standard multi-label filter approach, which considers each label separately. Thus, it has the disadvantage of do not considering label correlations.

The second method, named IG-ML, aims at taking into consideration label correlations. To this end, the Information Gain (IG) measure for multi-label

data proposed in [3] is used directly in the multi-label data. IG-ML evaluates the multi-label IG of the set of features $\{X_1, X_2, ..., X_M\}$ and the ones with IG values greater than or equal to a threshold are selected.

5 Experimental Evaluation

Both methods were implemented using Mulan[3], a Java package for multi-label classification based on Weka[4]. The experiments were carried out using two different base single-label classifiers from Weka: J48, an implementation of the decision tree C4.5 algorithm [9] and the support vector machine SMO learning algorithm. The methods were evaluated using 10 datasets. All the reported results were obtained by Mulan using 10-fold cross validation with paired folds.

5.1 Datasets

Table 2 describes the datasets used in the experiments, obtained from the Mulan's repository[5]. It shows the datasets domain (Domain); number of examples (N); number of features (M); number of labels ($|L|$); Label Cardinality (LC), which is the average number of labels associated with each example defined by Equation 2; Label Density (LD), which is the normalized cardinality defined by Equation 3, and the number of Distinct Combinations (DC) of labels.

$$LC(D) = \frac{1}{|D|}\sum_{i=1}^{|D|}|Y_i| \qquad (2) \qquad\qquad LD(D) = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i|}{|L|} \qquad (3)$$

Table 2. Datasets used for experiments

Dataset	Domain	N	M	|L|	LC	LD	DC
1-*bibtex*	text	7395	1836	159	2.40	0.02	2856
2-*cal500*	music	502	68	174	26.04	0.15	502
3-*corel16k001*	images	13766	500	153	2.86	0.019	4803
4-*corel5k*	images	5000	499	374	3.52	0.01	3175
5-*emotions*	music	593	72	6	1.87	0.31	27
6-*enron*	text	1702	1001	53	3.38	0.06	753
7-*genbase*	biology	662	1186	27	1.25	0.05	32
8-*medical*	text	978	1449	45	1.25	0.03	94
9-*scene*	image	2407	294	6	1.07	0.18	15
10-*yeast*	biology	2417	103	14	4.24	0.30	198

5.2 Performance Measures

The performance of multi-label classifiers can be evaluated using different measures. Some of these measures are adaptations from the single-label classification

[3] `http://mulan.sourceforge.net`
[4] `http://www.cs.waikato.ac.nz/ml/weka/`
[5] `http://mulan.sourceforge.net/datasets.html`

problem, while others were specifically defined for multi-label tasks. In what follows, we briefly describe the measures used in this work to compare both methods. These measures are *Hamming Loss*, *Accuracy*, *F-Measure* and *Subset Accuracy*, defined by Equations 4 to 7 respectively, where Δ represents the symmetric difference between two sets, Y_i is the set of true labels, Z_i is the set of predicted labels and $I(\text{true}) = 1$ and $I(\text{false}) = 0$.

$$Hamming\,Loss(H, D) = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \Delta Z_i|}{|L|} \tag{4}$$

$$Accuracy(H, D) = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \tag{5}$$

$$F\text{-}Measure(H, D) = \frac{1}{N}\sum_{i=1}^{N}\frac{2|Y_i \cap Z_i|}{|Z_i| + |Y_i|} \tag{6}$$

$$Subset\,Accuracy(H, D) = \frac{1}{N}\sum_{i=1}^{N} I(Z_i = Y_i) \tag{7}$$

All these performance measures have values in the interval [0..1]. For *Hamming Loss*, the smaller the value, the better the algorithm performance is, while for the other measures greater values indicate better performance. Note that *Subset Accuracy* is a very strict evaluation measure as it requires an exact match of the predicted and the true set of labels. Furthermore, as one of the advantages of FS is to reduce the data dimensionality, the average percentage of feature reduction in RF and IG-ML was also considered in the experimental evaluation.

5.3 Results and Discussion

Table 3 presents the average feature reduction and the standard deviation (in brackets) carried out by RF and IG-ML, using as a threshold 0.01 and 0.1 respectively, which can be considered conservative [11].

Table 3. Average percent of feature reduction (and standard deviation).

Dataset	RF	IG-ML
1-*bibtex*	78.31(0.31)	84.79(0.30)
2-*cal500*	8.82(0.98)	0.00(0.00)
3-*corel16k001*	70.10(0.67)	–
4-*corel5k*	43.99(1.62)	95.93(0.31)
5-*emotions*	23.89(1.94)	1.53(0.44)
6-*enron*	1.27(0.30)	6.22 (0.54)
7-*genbase*	95.51(0.21)	97.05 (0.07)
8-*medical*	86.62(1.06)	95.89 (0.05)
9-*scene*	19.15(0.58)	11.63 (1.64)
10-*yeast*	40.58(2.74)	15.73 (2.85)

Note that the IG values of all features for dataset *corel16k001* were lower than the threshold (8 cases). As can be observed, for some datasets both methods reduced the number of features by more than 75% (*bibtex*, *genbase* and *medical* datasets). For *cal500* RF reduced less than 10% while IG-ML no reduced dimensionality. However, for *corel5k*, IG-ML was able to reduce twice as much features than RF. Thus, we can conclude that, in general, RF and IG-ML select different features. Nevertheless, the features selected must be useful for the multi-label learning algorithm. To this end, using the BR approach and the two different base-learning algorithms J48 and SMO, the classifiers constructed using all features, and the features selected by RF and IG-ML were analyzed.

For all datasets and base-learning algorithms, the average (and the standard deviation) of the four multi-label performance measures were tabulated. Observe that the value of these measures for the classifiers constructed using all features represent a good *Baseline* for the ones obtained with the features selected by RF and IG-ML. However, due to lack of space, these tabulated results are not shown in this paper, but they can be found at `http://www.labic.icmc.usp.br/pub/mcmonard/ExperimentalResults/SBIA2012.pdf`.

These results show that from the total of 152 performance measure values tabulated (2 base-learning algorithms $\times$ 2 FS methods $\times$ 4 performance measures $\times$ 10 datasets -8), and considering the standard deviation (which has 0.07 as its maximum value), only 21 of them show a degradation compared to the correspondent *Baseline* performance measure. This represents less than 14%, which can be considered a good result. Furthermore, 9 of these 21 cases were obtained when using J48, and the remaining 12 by SMO. Regarding the datasets, most of the performance measures were significantly worse than those for the correspondent *Baseline* for *corel5k* (10 cases, 5 by RF and 5 by IG-ML). From the remaining 11 cases, 10 were obtained using the features selected by RF in the following datasets using the specified base-learning algorithm (this behavior happens to both algorithms case the learning algorithm is not specified), and performance measures: 1-*bibtex* (*Subset Accuracy* & SMO, *F-Measure* and *Accuracy*); 3-*corel16k001* (*F-Measure* and *Accuracy*, *Subset Accuracy* & J48). The last case was obtained by IG-ML in the *bibtex* dataset (*Hamming Loss & SMO*).

Nevertheless, in order to evaluate FS two aspects should be considered simultaneously: the reduction in the number of features *versus* the performance measure values of the classifier generated using the features selected. To this end, a graphical analysis is more appropriate. It should be observed that the evaluation of multi-label learning is more difficult than for single-label learning. In fact, the classification of a new instance by single-label classifiers has only two possible outcomes: correct or incorrect. On the other hand, multi-label classifiers should also take into account partially correct classification. Thus, several performance measures have to be analyzed. Due to lack of space, only one graph with this kind of analysis is shown. Graphs for all datasets and the four performance measures used in this work can be found at `http://www.labic.icmc.usp.br/pub/mcmonard/ExperimentalResults/SBIA2012.pdf`.

To illustrate, Figure 1 shows this information for *Hamming Loss* and *F-Measure* using *bibtex* dataset. *Baseline* refers to the performance measure values obtained using all features. One can note that for *Hamming Loss*, results nearer to the left-hand bottom corner of the figure are the best, while for *F-Measure* the best results are the ones nearer to the right-hand bottom corner.

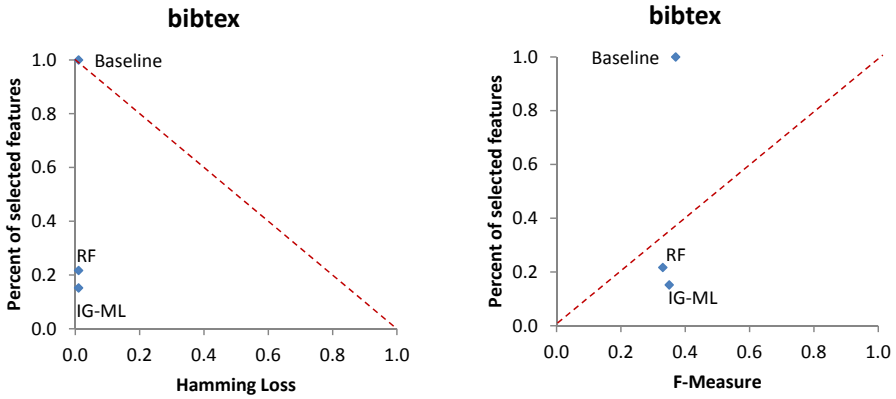

Fig. 1. *Bibtex* dataset FS evaluation using J48 as base-learning algorithm.

As can be observed in Figure 1, both FS methods obtained good results, since the performance measure values are similar to the *Baseline*, but were obtained using less features. Moreover, for this dataset, IG-ML is slightly better than RF since it was able to reduce the number of features more. Similar results for this dataset were obtained using SMO as base-learning algorithm.

6 Conclusion

Selecting features is an important task in machine learning in order to take care of the *curse of dimensionality* problem. This work analyses the behavior of two feature selection methods for multi-label learning which use the filter approach. The first method, RF, uses the multi-label feature selection standard approach. This approach considers a feature evaluation measure for each label separately, which are further composed in only one evaluation measure used to select the features. In this work, we proposed the use of ReliefF to evaluate each feature separately. However, the standard approach fails to consider any correlation among labels. The second method, IG-ML, aims to take into account the correlation among labels. To this end, the new information gain measure proposed in the literature for multi-label learning is used directly in the multi-label data as a feature evaluation measure. Both FS methods were thoroughly evaluated experimentally in ten benchmark datasets, showing promising results.

As future work, we plan to investigate the possibility of extending the ideas behind ReliefF for multi-label data.

Acknowledgment: The authors would like to thank the anonymous referees for their insightful comments on this paper. This research was supported by the Brazilian Research Council FAPESP.

References

1. Chen, W., Yan, J., Zhang, B., Chen, Z., Yang, Q.: Document transformation for multi-label feature selection in text categorization. In: IEEE International Conference on Data Mining. pp. 451–456 (2007)
2. Cherman, E.A., Metz, J., Monard, M.C.: Incorporating label dependency into the binary relevance framework for multi-label classification. Expert Systems with Applications 39(2), 1647–1655 (2012)
3. Clare, A., King, R.D.: Knowledge discovery in multi-label phenotype data. In: 5th European Conference on Principles of Data Mining and Knowledge Discovery. pp. 42–53. Springer-Verlag (2001)
4. Demšar, J.: Algorithms for subsetting attribute values with Relief. Machine Learning 78, 421–428 (2010)
5. Doquire, G., Verleysen, M.: Feature selection for multi-label classification problems. In: Cabestany, J., Rojas, I., Joya, G. (eds.) Advances in Computational Intelligence, chap. 2, pp. 9–16. Springer-Verlag/Heidelberg (2011)
6. Esuli, A., Fagni, T., Sebastiani, F.: Boosting multi-label hierarchical text categorization. Information Retrieval 11(4), 287–313 (2008)
7. Gu, Q., Li, Z., Han, J.: Correlated multi-label feature selection. In: ACM International Conference on Information and Knowledge Management. pp. 1087–1096 (2011)
8. Liu, H., Motoda, H.: Computational Methods of Feature Selection. Chapman & Hall/CRC (2008)
9. Quinlan, J.: C4.5: Programs for Machine Learning. Morgan Kaufmann (1993)
10. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. In: Buntine, W., Grobelnik, M., Mladenic, D., Shawe-Taylor, J. (eds.) Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science, vol. 5782, pp. 254–269. Springer Berlin / Heidelberg (2009)
11. Robnik-Sikonja, M., Kononenko, I.: Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning 53(1–2), 23–69 (2003)
12. Spolaôr, N., Monard, M.C., Lee, H.D.: A systematic review to identify feature selection publications in multi-labeled data. ICMC Technical Report N° 374. 31 pg. (2012), University of São Paulo
13. Spolaôr, N., Cherman, E.A., Monard, M.C.: Using ReliefF for multi-label feature selection (in portuguese). In: Conferencia Latinoamericana de Informática. pp. 960–975 (2011)
14. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.: Multi-label classification of music into emotions. In: International Conference on Music Information Retrieval. pp. 1–6 (2008)
15. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. Data Mining and Knowledge Discovery Handbook pp. 1–19 (2009)
16. Wei, Q., Yang, Z., Junping, Z., Wang, Y.: Semi-supervised multi-label learning algorithm using dependency among labels. In: International Conference on Machine Learning and Computing. pp. 112–116 (2009)
17. Zhang, M.L., Peña, J.M., Robles, V.: Feature selection for multi-label naive bayes classification. Information Sciences 179, 3218–3229 (2009)