Feature Selection for Multi-Label Learning*

Newton Spolaôr, Maria Carolina Monard
Laboratory of Computational Intelligence
Institute of Mathematics and Computer Science
University of São Paulo
São Carlos, Brazil
newtonspolaor@gmail.com, mcmonard@icmc.usp.br

Huei Diana Lee
Laboratory of Bioinformatics
State University of West Paraná
Foz do Iguaçu, Brazil
hueidianalee@gmail.com

Abstract

Feature Selection plays an important role in machine learning and data mining, and it is often applied as a data pre-processing step. This task can speed up learning algorithms and sometimes improve their performance. In multi-label learning, label dependence is considered another aspect that can contribute to improve learning performance. A replicable and wide systematic review performed by us corroborates this idea. Based on this information, it is believed that considering label dependence during feature selection can lead to better learning performance. The hypothesis of this work is that multi-label feature selection algorithms that consider label dependence will perform better than the ones that disregard it. To this end, we propose multi-label feature selection algorithms that take into account label relations. These algorithms were experimentally compared to the standard approach for feature selection, showing good performance in terms of feature reduction and predictability of the classifiers built using the selected features.

1 Introduction

Feature Selection (FS) plays an important role in machine learning and data mining, and it is often applied as a data preprocessing step. FS aims to find a small number of features that describes the dataset as well as the original set of features does [Liu and Motoda, 2007]. Thus, it provides support to tackle the "*curse of dimensionality*" problem when learning from high-dimensional data. FS can effectively reduce data dimensionality by removing irrelevant and/or redundant features, speeding up learning algorithms and sometimes improving their performance. The filter approach is one of the most usual approaches applied to select features, as it has a potentially lower computational cost than other alternatives.

Many importance measures have been proposed for single-label data. However, it is intuitive for humans to associate their observations (examples) with two or more concepts, leading to multi-label data. The interest in knowledge extraction from this data has increased and different multi-label learning applications have emerged [Tsoumakas *et al.*, 2010].

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

The standard approach for multi-label FS, which transforms multi-label data into single-label data before using traditional FS algorithms, is implementable within the Binary Relevance (BR) approach [Tsoumakas *et al.*, 2010]. A BR drawback is that label dependence is often ignored.

Label dependence has been highlighted as an important issue by the multi-label learning community [Dembczyński *et al.*, 2012], as it is considered that this aspect may contribute to improve learning performance. A replicable and wide systematic review performed by us corroborates this idea [Spolaôr *et al.*, 2014], as taking into account label dependence for FS led to good results in related work. Based on this information, it is believed that considering label dependence during FS can contribute to obtain better learning performance.

The hypothesis of this work is that FS algorithms that consider label dependence will perform better than the ones that disregard label dependence. To this end, we propose filter multi-label feature selection algorithms that take into account label relations. These algorithms were experimentally compared to the standard approach for FS, showing good performance in terms of feature reduction and predictability of the classifiers built using the selected features [Spolaôr, 2014].

2 Proposed methods

BR transforms a multi-label dataset into q single-label binary datasets, one per label. Afterwards, it learns (or selects features) from each binary problem and combines the results. As mentioned, BR often ignores label dependence. An alternative to overcoming this drawback would be to build labels based on relations among the original labels from data and include the new labels during FS. The main idea of variable (label or feature) construction is to gather information about

*This research was supported by the São Paulo Research Foundation (FAPESP), grant 2011/02393-4.

the relations among the original variables and infer additional variables. However, to the best of our knowledge, there is little research on label construction for multi-label data.

In this work, we proposed the Label Construction for Feature Selection (*LCFS*) method to build q' binary variables (new labels) based on label relations. In particular, *LCFS* constructs each variable y_{ij} by combining the original labels within a pair (y_i, y_j), $i \neq j$, $y_i \in L$ and $y_j \in L$. To do so, two steps are required: (1) Selection and (2) Generation. The former chooses pairs of labels, whereas the latter combines the labels within each pair to generate a new label. The variables constructed are then included as new labels in the original dataset and the standard multi-label FS approach is used in the augmented dataset to select features. Afterwards, the dataset described by the selected features and the original labels can be submitted to any multi-label learning algorithm.

The single-label algorithms ReliefF and RReliefF use the importance measure Relief (RF) to evaluate features [Robnik-Sikonja and Kononenko, 2003]. Unlike strictly univariate measures, RF considers the effect of interacting features, as all features are used to search for nearest examples (neighbors). Furthermore, these traditional algorithms deal with numerical data directly, forgoing data discretization.

We proposed different ReliefF and RReliefF extensions for multi-label feature selection, such as *RF-ML* [Spolaôr, 2014]. Unlike *LCFS*, *RF-ML* deals with multi-label data directly. In particular, *RF-ML* extends RReliefF by using a dissimilarity function $mld(Y_i, Y_j)$ between multi-labels Y_i and Y_j. This function models the probability that the predictions of two examples are different. By using mld, only one search for nearest neighbors is performed, yielding an algorithm with smaller complexity than the combination between the single-label ReliefF and BR (*RF-BR*). It should be emphasized that any dissimilarity function between sets can implement mld.

3 Main results

Experimental evaluations conducted in 10 benchmark datasets showed that the use of random selection and the operators XOR and XNOR was the best *LCFS* setting for FS based on the traditional Information Gain importance measure [Spolaôr *et al.*, 2014]. Furthermore, this setting gave rise to significantly better classifiers than the ones obtained by FS based on BR when the number of selected features is small. The setting was also competitive with the ones achieved by FS based on Label Powerset (LP), a problem transformation approach that considers label relations [Spolaôr, 2014].

An evaluation conducted in 45 synthetic datasets with different noise levels showed that *RF-ML* with Hamming distance was the best proposal setting, significantly outperforming *RF-BR* and the combination between RF and LP (*RF-LP*). In addition, experimental evaluations carried out in 10 benchmark datasets suggest that the proposed extension is competitive with the standard approach for FS, based on BR and LP, in terms of feature reduction and predictability of the classifiers built using the selected features [Spolaôr and Monard, 2014]. It should also be noted that our proposal avoids disadvantages specific to the problem transformation approaches. LP, for example, considers only the multi-labels contained in the training set.

4 Conclusion

We proposed filter multi-label feature selection methods that take into account label relations. *LCFS* and *RF-ML* are publicly available at `http://goo.gl/sPMHmm` and at `http://goo.gl/pSwzgp`. The following three contributions are directly related to the objective of this work.

The systematic review conducted by us depicted an up-to-date panorama on multi-label FS by synthesizing the 74 papers cited at `https://db.tt/p8WzHlHZ`. Regarding the objective, it showed that considering label dependence during FS led to good results [Spolaôr *et al.*, 2014].

LCFS is an alternative to include label relations into the traditional BR approach. The *LCFS* experimental evaluation showed that considering label relations is important for FS based on problem transformation. As future work, we plan to evaluate other *LCFS* strategies, as well as expanding the binary combination of labels to combinations of higher order.

RF-ML is a pioneer adaptation of ReliefF and RReliefF to perform multi-label FS directly. By using a dissimilarity function between multi-labels, *RF-ML* preserves label relations for FS. As future work, we plan to investigate a potential relation between learning evaluation measures and mld, as some of the best results, according to the Hamming Loss evaluation measure, were obtained with Hamming distance.

References

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

[Liu and Motoda, 2007] H. Liu and H. Motoda. *Computational Methods of Feature Selection*. Chapman & Hall/CRC, 2007.

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

[Spolaôr and Monard, 2014] N. Spolaôr and M. C. Monard. Evaluating ReliefF-based multi-label feature selection algorithm. In A. L. C. Bazzan and K. Pichara, editors, *Advances in Artificial Intelligence – IBERAMIA 2014*, volume 8864 of *Lecture Notes in Computer Science*, pages 194–205. Springer International Publishing, 2014.

[Spolaôr *et al.*, 2014] N. Spolaôr, M. C. Monard, G. Tsoumakas, and H. D. Lee. Label construction for multi-label feature selection. In *Brazilian Conference on Intelligent Systems*, pages 247–252. IEEE, 2014.

[Spolaôr, 2014] N. Spolaôr. *Feature Selection for Multi-label Learning (in Portuguese)*. Phd thesis, ICMC-University of São Paulo - Brazil, 2014.

[Tsoumakas *et al.*, 2010] G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining multi-label data. In O. Maimon and L. Rokach, editors, *Data Mining and Knowledge Discovery*, pages 1–19. Springer United States, 2010.