ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

Feature selection for label ranking

Noelia Sánchez-Maroño and Beatriz Pérez-Sánchez *

University of A Coruña
Research Center on Information and Communication Technologies (CITIC)
Campus de Elviña, 15071 - Spain

Abstract. Over the last years, feature selection and label ranking have attracted considerable attention in Artificial Intelligence research. Feature selection has been applied to many machine learning problems with excellent results. However, studies about its combination with label ranking are undeveloped. This paper presents a novelty work that uses feature selection filters as a preprocessing step for label ranking. Experimental results show a significant reduction, up to 33%, in the number of features used for the label ranking problems whereas the performance results are competitive in terms of similarity measure.

1 Introduction

Feature selection (FS) has been embraced as one of the high research areas during the last few years, due to the appearance of datasets containing hundreds of thousands of features, such as microarray or text categorization datasets [1]. FS algorithms choose the relevant features while discard the redundant or irrelevant ones. There are many potential benefits of FS: facilitating data visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times, defying the curse of dimensionality to improve prediction performance [2]. Due to all these advantages, FS has been applied to different types of problems, although classification tasks have focused most of the attention.

On the other hand, the topic of *preferences* has attracted considerable attention in Artificial Intelligence (AI) research, notably in fields such as autonomous agents, non-monotonic reasoning, constraint satisfaction, planning, and qualitative decision theory [3]. Roughly, one can say that preference learning is about inducing predictive preference models from empirical data. The problem of *learning to rank*, which has been studied extensively in recent years, is an important special case; here, the goal is to predict preference models in the form of total orders of a set of alternatives [4]. In general, a preference learning task consists of some set of items for which preferences are known, and the task is to learn a function that predicts preferences for a new set o items, or for the same set of items in a different context [3]. There are three types of ranking problems, namely label ranking, instance ranking and object ranking. In this work, we will focus on label ranking which can be seen as a generalization of

*This research has been supported in part by the Spanish Government's Secretaría de Estado de Investigación (Grant TIN2015-65069-C2-1-R), the Xunta de Galicia (Grants GRC2014/035, ED431G/01 and ED341D R2016/045) and European Union FEDER funds.

conventional classification, where a complete ranking of class labels is associated with an instance instead of a single class label.

The goal of this paper is to combine both techniques: feature selection and label ranking in order to get all the advantages of the former in the latter. Besides the importance of both topics, to the knowledge of the authors, there are no papers in the literature covering this issue, maybe because most of the attention has been devoted to ranking objects or instances. This paper presents a novelty work that uses well-known FS methods as a preprocessing step for label ranking. The methodology proposed is applied over different datasets in this area demonstrating similar levels of performance whereas the number of features is significantly reduced.

2 Proposed methodology

In label ranking, the problem is to predict, for any instance $\boldsymbol{x}$ (e.g. a person) from an instance space X, a preference relation $\succ_x \subseteq L \times L$ among a finite set $L= \{\lambda_1, \lambda_2, \ldots, \lambda_c\}$ of labels or alternatives, where $\lambda_i \succ_x \lambda_j$ means that instance $\boldsymbol{x}$ prefers label λ_i to λ_j [5]. The training information consists of a set of instances where each one is represented by a feature vector of length n, i.e., $\boldsymbol{x}= \{x_1, x_2, \ldots, x_n\}$. Besides, each instance is related to a subset of all pairwise preferences of the form $\lambda_i \succ_x \lambda_j$, indicating the preferences for all set of labels L.

The goal of this work is to apply FS methods as a preprocessing step of label ranking algorithms in order to reduce the feature vector, so each instance $\boldsymbol{x}$ would be represented by a feature vector of length r, being $r < n$. The idea is to use well-known and publicly available FS methods. However, these FS methods are not prepared to deal with label ranking. Therefore, each problem has been divided into different binary problems, so each one only considers two labels L_i and L_j and the desired output is 1 when $\lambda_i \succ_x \lambda_j$ and 0 otherwise. For the experimental study, cross validation has been used to determine the overall performance. Then, as illustrated in Fig 1, for each fold, the procedure consists on dividing each problem into one versus one binary problems, then applies FS methods over each binary problem and combines all features to derive an unified and reduced subset of features. Finally, label ranking methods have been applied over the training set of instances using only the features selected. For simplicity, union of features has been employed as strategy to determine the features to retain. Next subsections briefly present the methods used for the experimental study.

2.1 Feature selection methods

FS has gained so much attention that the number of available methods is continuously growing. We have selected well-known filter methods because of their speed and good performance, covering both types of filters, i.e, subset (CFS) and ranker (IG and RF). Next, we briefly present these methods. For a broader explanation of them, the interested reader can consult the book by Bolón-Canedo

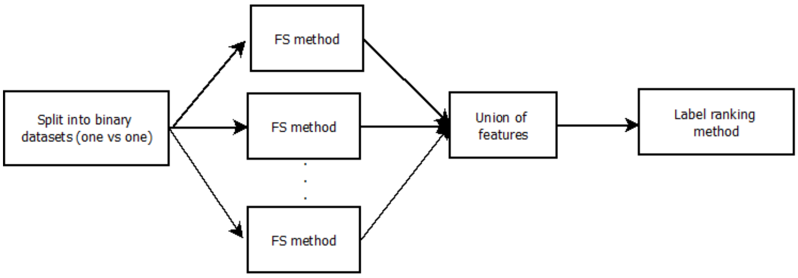

Fig. 1: Schema for each fold in the proposed methodology

et al. [1]. All these methods have been used with default parameters provided by Weka platform [6].

- Correlation-based Feature Selection (CFS) is a simple multivariate filter algorithm that ranks feature subsets according to a correlation based heuristic evaluation function. The bias of the evaluation function is toward subsets that contain features that are highly correlated with the class and uncorrelated with each other.
- Information Gain (IG) evaluates the worth of a feature by measuring the information gain with respect to the class.
- ReliefF (RF) is an extension of the original Relief algorithm. The original Relief works by randomly sampling an instance from the data and then locating its nearest neighbor from the same and opposite class. The values of the attributes of the nearest neighbors are compared to the sampled instance and used to update the relevance scores for each attribute. This method may be applied in all situations, has low bias, includes interaction among features and may capture local dependencies which other methods miss.

2.2 Label ranking

Contrary to FS, there are not many algorithms available for label ranking, we have selected Ranking by pairwise comparison and label ranking trees. These algorithms are available at Weka-LR, a label ranking extension for WEKA[1].

- Ranking by Pairwise Comparison (RPC) consists on reducing the problem of label ranking to several binary classification problems. The predictions of this ensemble of binary classifiers can then be combined into a ranking using a separate ranking algorithm [5].

[1] Online at `https://www-old.cs.uni-paderborn.de/fachgebiete/intelligente-systeme/software/weka-lr-a-label-ranking-extension-for-weka.html`

- Label Ranking Trees (LRT), as indicated by their name, learn decision trees the leaf nodes of which are associated with (possibly incomplete) label ranking [7].

3 Experimental study

In this section the results achieved using different datasets are introduced. Table 1 summarizes the main characteristics (first four columns) of the datasets used in this study. Next columns in table 1 indicate the number of features retained for each FS method used (see subsection 2.1) fixing the threshold to 70% of features for rankers. It is important to note that the number of features is calculated as the length of the set formed with the features chose at each fold (without repetition), and in turn, the set of features at each fold is formed by adding the features used for each binary problem (see Figure 1). This strategy is very conservative and retains many features, but even so the reduction is significant. Other strategies (for instance, maintaining those features that appear at least in b binary problems) could lead to much smaller number of features.

Dataset	Instances	Classes	Features	CFS	IG	RF
cold [5]	2465	4	24	17	18	22
diau [5]	2465	7	24	10	11	16
heat [5]	2465	6	24	16	15	18
german [8]	411	5	29	21	17	15
analcat [9]	841	4	70	68	53	53

Table 1: Characteristics of the datasets used in the experimental study and number of features selected by each filter method

For analyzing the performance of the proposed methodology, the well-known Spearman rank correlation has been used as similarity measure [10], it takes values between -1 (no similarity) and 1 (full similarity). Table 2 shows the mean and standard deviation of this measure. Note that RPC algorithm transforms the problem into binary problems, so different classifiers can be used for each one; for that reason, three different classifiers have been selected: logistic (established as default in Weka-LR), LibSVM and C4.5. For the sake of completeness, label ranking algorithms have been also applied without using feature selection techniques (column NoFS in table 2). Looking at this table, we can see that the better performance for 3 of 5 datasets (cold, diau and heat) is achieved without FS and using the RPC algorithm based on LibSVM classifiers which behave well with large number of features and do not suffer overfitting. However, these datasets also achieved very good results using Relief algorithm and RPC with LibSVM classifiers and the reduction of features is important (8%, 33% and 25% for cold, diau and heat, respectively as shown in table 1). For german and analcat datasets, the performance results are better using a FS technique, IG combined with RPC based on C4.5 for the former, and Relief with RPC based on LibSVM for the latter. Besides, it is important to highlight the reduction in the number of features, german with IG filter only needs 59% of features whereas analcat with ReliefF uses just 75%.

	RPC-log			
	NoFS	CFS	IG	RF
cold	.256 ± .029	*.260 ± .025*	.260 ± .026	.253 ± .033
diau	*.423 ± .015*	.424 ± .016	.423 ± .017	.418 ± .018
heat	.155 ± .025	.155 ± .027	.155 ± .026	*.156 ± .026*
german	.934 ± .015	.908 ± .064	.935 ± .014	*.940 ± .005*
analcat	.934 ± .009	.941 ± .011	*.944 ± .005*	.933 ± .008
	RPC-C4.5			
	NoFS	CFS	IG	RF
cold	.246 ± .009	.247 ± .010	*.248 ± .017*	.228 ± .015
diau	.419 ± .018	.423 ± .023	*.425 ± .024*	.419 ± .021
heat	.137 ± .020	*.149 ± .019*	.148 ± .018	.134 ± .014
german	.955 ± .006	.952 ± .013	**.957 ± .004**	.953 ± .003
analcat	.910 ± .008	.910 ± .012	*.915 ± .003*	.870 ± .108
	RPC-LibSVM			
	NoFS	CFS	IG	RF
cold	**.288 ± .020**	.281 ± .015	.281 ± .019	.279 ± .026
diau	**.433 ± .019**	.430 ± .022	.428 ± .022	.431 ± .016
heat	**.181 ± .029**	.174 ± .028	.173 ± .026	.180 ± .028
german	*.813 ± .014*	*.813 ± .014*	*.813 ± .014*	*.813 ± .014*
analcat	.953 ± .011	.956 ± .008	.959 ± .010	**.965 ± .006**
	LRT			
	NoFS	CFS	IG	RF
cold	.144 ± .015	.158 ± .023	*.162 ± .030*	.150 ± .021
diau	.277 ± .019	.270 ± .026	*.282 ± .021*	.270 ± .018
heat	.082 ± .018	.101 ± .021	*.115 ± .025*	.084 ± .022
german	.940 ± .007	.929 ± .017	.938 ± .005	*.940 ± .004*
analcat	.913 ± .012	.914 ± .012	.915 ± .011	*.918 ± .016*

Table 2: Mean and standard deviation for the Spearman rank correlation. Best mean results for each dataset and label ranking method are marked in italic. Beast mean results for each dataset also marked in bold.

4 Conclusions

In this paper, feature selection methods have been used as a previous step to the application of label ranking algorithms. Thanks to this, we have achieved a significant reduction in the number of features used for the label ranking problems (over 25% in 4 of 5 datasets) whereas the performance has been maintained or even slightly overcome. Future work includes an extension of the experimental study, considering more datasets and more algorithms from both types: feature selection and label ranking. Besides, different strategies to compound the final

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

set of features could be explored. Finally, adapting an existing FS method to tackle with label ranking datasets instead of combining the results of different binary problems could be an interesting line of research.

References

[1] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos. *Feature selection for high-dimensional data*. Springer, 2015.

[2] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[3] J. Fürnkranz and E. Hüllermeier. Preference learning: An introduction. In *Preference learning*, pages 1–17. Springer, 2010.

[4] E. Hüllermeier. Preference learning. In *Preference Learning Workshop, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2010)*, 2010.

[5] E. Hüllermeier, J. Fürnkranz, W. Cheng, and K. Brinker. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172(16):1897–1916, 2008.

[6] E. Frank, M. A. Hall, and I. H. Witten. The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques".

[7] W. Cheng, J. Hühn, and E. Hüllermeier. Decision tree and instance-based learning for label ranking. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 161–168. ACM, 2009.

[8] C. Rebelo de Sá, Label Ranking datasets, Mendeley Data, v1. http://dx.doi.org/10.17632/3mv94c8jpc.1, [Online; accessed February 2018].

[9] StatLib—Datasets Archive. http://lib.stat.cmu.edu/datasets/, [Online; accessed February 2018].

[10] Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.