ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

Using a Feature Selection Ensemble on DNA Microarray Datasets

Borja Seijo-Pardo, Verónica Bolón-Canedo
and Amparo Alonso-Betanzos *

Department of Computer Science - University of A Coruña
Campus de Elviña s/n 15071 - A Coruña, Spain

Abstract. DNA microarray has brought a difficult challenge for researchers due to the high number of gene expression contained and the small samples size. Therefore, feature selection has become an indispensable preprocessing step. In this paper we propose an ensemble for feature selection based on combining rankings of features. The individual rankings are combined with different aggregation methods, and a practical subset of features is selected according to a data complexity measure –the inverse of Fisher discriminant ratio–. The proposed ensemble, tested on seven different DNA microarray datasets using a Support Vector Machine as classifier, was able to obtain the best results in different scenarios.

1 Introduction

In recent years, the real-world scenarios have incremented considerably their dimension and size. Specifically, DNA microarray experiments generate a lot of gene expressions for a low number of patients. In DNA microarray data, every sample corresponds with a patient and each feature represents a gene expression coefficient corresponding to the abundance of *Messenger Ribonucleic Acid (mRNA)* in a concrete sample. An important application of DNA microarray data is to separate healthy patients from cancer patients based on their gene expression "profile". The DNA microarray datasets pose an enormous challenge for feature selection researchers due to the high number of gene expression contained and the small samples size.

Several studies have shown that most genes in a DNA microarray experiment are not relevant for an accurate classification among different classes of the problem [1]. To reduce the data dimensionality, feature selection preprocess plays a crucial role in DNA microarray analysis. This process obtains an ideal subset of relevant features of the data, eliminating irrelevant and redundant information. As a result, the dimension of the datasets is reduced, allowing a lower use of the storage size and an improvement of the computational time of machine learning algorithms. In feature selection, there are two different approaches when conducting the evaluation of the features of a dataset: (i) *individual evaluation* and (ii) *subset evaluation* [2]. In the first case, a ranking

*This research has been financially supported in part by the Spanish Ministerio de Economía y Competitividad (research projects TIN 2012-37954 and TIN2015-65069-C2-1-R), by European Union FEDER funds and by the Consellería de Industria of the Xunta de Galicia (research project GRC2014/035). V. Bolón-Canedo acknowledges Xunta de Galicia postdoctoral funding (ED481B 2014/164-0).

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

of features is returned by assigning a level of relevance to each of these features. The feature selection methods with this evaluation approach are also known as *ranker methods*. In the second case, successive subsets of features are generated and evaluated iteratively, according to an optimality criterion, until reaching the final subset of selected features.

Machine learning methods have traditionally used a single learning model to solve a given problem. However, it has been recently observed that by combining different learning models, better results could be obtained (called *ensemble learning*). Ensemble learning has been normally applied to classification process but can also be thought as a means of improving other machine learning disciplines such as feature selection.

The idea of this paper is to use an ensemble of ranker methods instead of a single method. Seven different ensemble configurations will be presented, depending on the combination method used to generate the final ranking. Since the ensemble is formed by ranker methods, a threshold is necessary to obtain a practical subset of features. The novelty of the approach presented in this work is that we use complexity measures to automatically establish this threshold, in particular we employ the inverse of Fisher discriminant ratio. The results obtained in a previous work on ensembles for feature selection [3] are taken as baseline to compare the results of the proposed approach. An extensive experimentation on different DNA microarray datasets, using a Support Vector Machine (SVM) as a classifier, shows the adequacy of the proposed ensemble.

2 Ensemble feature selection

Datasets can be very large, both in number of samples or in number of features, and can also be redundant, noisy, multivariate and nonlinear. In order to make a correct choice, a user not only needs to know the domain and the characteristics of each dataset well, but also is expected to understand technical details of available algorithms [4]. In this sense, a possible way to confront this situation is to use an ensemble of feature selection algorithms. Specifically, in this study we use ranker methods, i.e. they return an ordered ranking of all the features. Thus, a threshold is necessary in order to obtain a practical subset of features. In this work, instead of choosing an arbitrary threshold, we have opted for the use of a data complexity measure to establish the threshold value automatically and tailored for the dataset, since they are a recent proposal to represent characteristics of the data which are considered difficult in classification tasks beyond estimates of error rates [5]. We assume that good candidate features would contribute to decrease the complexity and must be maintained. Figure 1 shows the proposed approach, and the pseudo-code can be seen in Algorithm 1.

Among the broad suite of feature selection methods available in the literature, four well-known filter methods were chosen to conform the final ensemble (*Chi-Square* [6], *Information Gain* [7], *mRMR* [8] and *ReliefF* [9]). The A_r rankings generated using the different feature selection methods, all of them using the same training data, must be combined in order to produce a unique final output

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

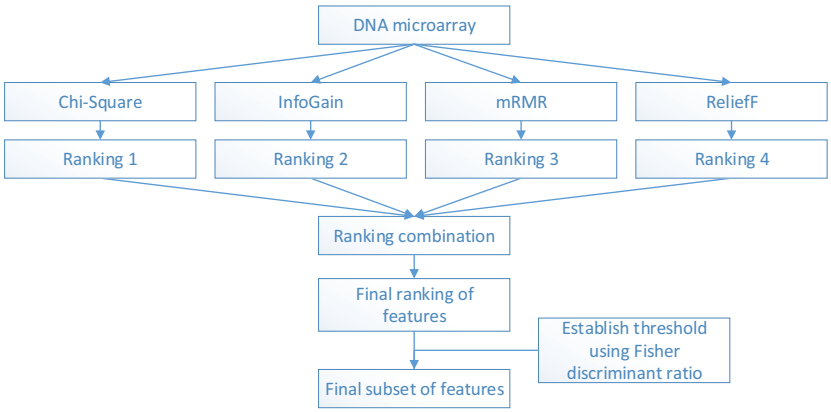

Fig. 1: Diagram of the proposed ensemble method.

Algorithm 1: Pseudo-code of the proposed ensemble method

Data: R — number of ranker methods
Data: T — number of features to be selected
Data: V — vector of complexity measure values
Result: E — classification test error

Calculate and store in V the complexity measure value of each feature.
for *each* r *from 1 to* R **do**
  Obtaining ranking A_r using feature selection method r
end
A = combining rankings A_r with a ranking combination method
Use V on A to obtain the optimal T value.
A_t = Select T top attributes from A
Build classifier SVM-RBF with the selected attributes A_t
Obtain test error E

A. The combination methods, also known as 'aggregators', are responsible for conducting the fusion of several rankings in order to obtain a final ranking. To perform the combination of the input rankings several different measures are used, ranging from simple calculation to more sophisticated measures. The different ranking combination functions selected to combine the several rankings in this study were: *SVM-Rank* [10] (a SVM-based method for learning of ranking functions), the *Min*, *Median*, *Mean* and *GeomMean* methods [11] (based on simple arithmetic operations) and the *Stuart* [12] and *RRA* [13] methods (based on statistical sorting distributions).

Finally, we have opted for the use of a data complexity measure to establish the threshold value T to obtain a practical subset of features A_t. In this paper, the measure selected was the Fisher discriminant ratio f [5], defined as:

$$f = \frac{\sum_{i=1,j=1,i\neq j}^{c} p_i p_j (\mu_i - \mu_j)^2}{\sum_{i=1}^{c} p_i \sigma_i^2}, \tag{1}$$

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

where μ_i, σ_i^2, and p_i are the mean, variance, and proportion of the ith class C, respectively. The Fisher discriminant ratio values are calculated for each feature of the dataset individually. The formula applied to establish the threshold and obtain the final subset of features is defined as:

$$e[v] = \alpha \times 1/f + (1 - \alpha) \times \rho \tag{2}$$

where α is a parameter with value in the interval $[0, 1]$ ($\alpha = 0.75$ for this work), ρ is the percentage of features retained (ranging from one to the total number of features of the dataset) and $1/f$ is the inverse of the Fisher discriminant ratio. A small complexity value $e[v]$ represents an easier problem.

3 Experimental study

The experiments performed on seven DNA binary microarray datasets, which are listed in Table 1, consisted of a comparison between the different ensemble configurations described in the Section 2, taking as baseline the methods used in a previous study [3]. The datasets originally divided into training and test sets were maintained, whilst, those with only training set were randomly divided using the common rule 2/3 for training and 1/3 for testing for the sake of comparison. This division introduces a more challenging scenario since, in some datasets, the distribution of the classes in the training set differs from the one in the test set. Table 1 shows the number of attributes and samples and also the distribution of the binary classes, indicating if the data is unbalanced.

Dataset	Features	Samples		Train distribution (%)	Test distribution (%)
		Train	Test		
Colon	2 000	42	20	67 - 33	60 - 40
DLBCL	4 026	32	15	50 - 50	53 - 47
CNS	7 129	40	20	65 - 35	65 - 35
Leukemia	7 129	38	34	71 - 29	59 - 41
Lung	12 533	32	149	50 - 50	90 - 10
Prostate	12 600	102	34	49 - 51	26 - 74
Ovarian	15 154	169	84	35 - 65	38 - 62

Table 1: Binary microarray datasets employed in the experimental study.

A Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel has been chosen for comparing the proposed ensemble configurations with the baseline methods in terms of classification error. The results obtained by the different approaches are displayed in Table 2. This table shows the test classification error of different methods and is divided in two parts: (i) the first seven rows represent the different ensemble configurations (*E-AggregatorMehod*) described in this paper (Section 2), and (ii) the last six rows correspond with the baseline methods (obtained from [3]) consisting of an ensemble method (*E2*) and five individual filter methods.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

Ranker	Colon	DLBCL	CNS	Leukemia	Lung	Prostate	Ovarian
E-SVMRank	20.00 (5)	**6.67** (5)	35.00 (5)	20.59 (5)	8.05 (5)	26.47 (10)	14.29 (5)
E-Min	**15.00** (10)	13.33 (5)	40.00 (5)	**8.82** (5)	**0.67** (5)	26.47 (5)	**0.00** (5)
E-Median	**15.00** (5)	13.33 (5)	35.00 (5)	11.77 (5)	3.36 (5)	26.47 (5)	**0.00** (5)
E-Mean	20.00 (5)	13.33 (5)	35.00 (5)	**8.82** (5)	3.37 (5)	26.47 (5)	**0.00** (5)
E-GeomMean	20.00 (5)	13.33 (5)	**25.00** (5)	**8.82** (5)	3.36 (5)	26.47 (5)	**0.00** (5)
E-Stuart	20.00 (5)	13.33 (5)	30.00 (5)	11.77 (5)	5.37 (5)	26.47 (5)	**0.00** (5)
E-RRA	20.00 (5)	13.33 (5)	35.00 (5)	14.71 (5)	5.37 (5)	26.47 (5)	1.19 (5)
E2	20.00 (34)	**6.67** (47)	35.00 (60)	11.76 (36)	1.34 (40)	**2.94** (89)	**0.00** (37)
CFS	**15.00** (19)	**6.67** (47)	35.00 (60)	11.76 (36)	1.34 (40)	**2.94** (89)	**0.00** (37)
Cons	20.00 (3)	13.33 (2)	35.00 (3)	20.59 (1)	18.12 (1)	73.53 (4)	**0.00** (3)
INT	**15.00** (16)	13.33 (36)	40.00 (47)	11.76 (36)	1.34 (40)	29.41 (73)	**0.00** (27)
InfoGain	**15.00** (25)	**6.67** (25)	35.00 (25)	11.76 (25)	**0.67** (25)	**2.94** (25)	1.19 (25)
ReliefF	**15.00** (25)	**6.67** (25)	30.00 (25)	17.65 (25)	2.01 (25)	5.88 (25)	**0.00** (25)

Table 2: Test classification error for DNA microarray datasets. The value shown in parenthesis represents the number of relevant features selected by each method. The best error for each dataset is highlighted in bold face.

The experimental results demonstrate the adequacy of the proposed ensemble, since they match or improve upon the results obtained in [3]. In fact, it can be seen that some of the proposed ensemble configurations (*E-Min*, *E-Mean* and *E-GeomMean*) improve the baseline error results on the *Leukemia* dataset, and the *E-GeomMean* configuration improves the baseline error results on the *CNS* dataset. In addition to this, at least one ensemble configuration matches the best baseline error when it is applied on the datasets *Colon*, *DLBCL*, *Lung* or *Ovarian*. In both cases, the proposed ensemble has the added benefit of reducing significantly the dataset dimension (using only five features, in most configurations, for the classification process). The *Prostate* dataset is the only one that does not improve their results when the proposed ensemble is applied. Notice that the *Prostate* dataset has been a big challenge for machine learning algorithms since the training and test sets were extracted from different experiments. Thus, the common assumption that the training and test data follow the same distribution is, in this case, violated. In fact, in a previous work [14] we have pointed out that some classifiers just assign all the samples to the minority class, leading to a poor classification performance around 26% global accuracy, as happens with our proposed ensemble.

4 Conclusions and discusion

DNA microarray data has brought a difficult challenge for researchers due to the high number of gene expression contained and the small samples size. For these reason, a feature selection preprocess is essential to confront the dimensionality problem. In this paper, an ensemble for feature selection was presented. The idea is to use an ensemble of methods rather than a single method, in order to take advantage of their individual strengths and overcome their weak points at the same time. This will have the added benefits of releasing the user from the task of knowing technical details about the DNA microarray scenario.

The particularity of the proposed ensemble is that it works with ordered

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

rankings of features, and therefore, a threshold is necessary in order to obtain a practical subset of features. Four well-known algorithms were chosen to form part of the ensemble and the individual rankings were combined with different combination methods. Since we have an ordered ranking of all the features we have opted for the use of data complexity measures to establish the threshold value, releasing the user from the task of choosing it in advance. In this study, the inverse of Fisher discriminant ratio was selected as data complexity measure. The experiments on seven DNA microarray datasets showed that our proposal was able to obtain competitive results when compared with results achieved in a previous work, with the added benefit of selecting automatically the threshold to establish the final number of features to consider in the classification stage.

References

[1] Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, Mark A Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537, 1999.

[2] L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, 5:1205–1224, 2004.

[3] Verónica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. Data classification using an ensemble of filters. *Neurocomputing*, 135:13–20, 2014.

[4] H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 17(4):491–502, 2005.

[5] Mitra Basu and Tin Kam Ho. *Data complexity in pattern recognition*. Springer Science & Business Media, 2006.

[6] H. Liu and R. Setiono. Chi2: Feature selection and discretization of numeric attributes. In *2012 IEEE 24th International Conference on Tools with Artificial Intelligence*, pages 388–388. IEEE Computer Society, 1995.

[7] J.R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[8] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005.

[9] I. Kononenko. Estimating attributes: analysis and extensions of relief. In *Machine Learning: ECML-94*, pages 171–182. Springer, 1994.

[10] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.

[11] Peter Willett. Combination of similarity rankings using data fusion. *Journal of chemical information and modeling*, 53(1):1–10, 2013.

[12] Stein Aerts, Diether Lambrechts, Sunit Maity, Peter Van Loo, Bert Coessens, Frederik De Smet, Leon-Charles Tranchevent, Bart De Moor, Peter Marynen, Bassem Hassan, et al. Gene prioritization through genomic data fusion. *Nature biotechnology*, 24(5):537–544, 2006.

[13] Raivo Kolde, Sven Laur, Priit Adler, and Jaak Vilo. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, 28(4):573–580, 2012.

[14] Veronica Bolon-Canedo, Laura Moran-Fernandez, and Amparo Alonso-Betanzos. An insight on complexity measures and classification in microarray data. In *Neural Networks (IJCNN), 2015 International Joint Conference on*, pages 42–49. IEEE, 2015.