Exploring signature multiplicity in microarray data using ensembles of randomized trees

Pierre Geurts[1] and Yvan Saeys[2]

[1] Department of EE and CS & GIGA-R, University of Liège, Belgium
[2] VIB-Ghent University, Ghent, Belgium
p.geurts@ulg.ac.be,yvan.saeys@ugent.be

Abstract. A challenging and novel direction for feature selection research in computational biology is the analysis of signature multiplicity. In this work, we propose to investigate the effect of signature multiplicity on feature importance scores derived from tree-based ensemble methods. We show that looking at individual tree rankings in an ensemble could highlight the existence of multiple signatures and we propose a simple post-processing method based on clustering that can return smaller signatures with better predictive performance than signatures derived from the global tree ranking at almost no additional cost.

1 Introduction

Feature selection is an important aspect of many machine learning applications in computational biology [4]. Traditionally, many standard feature selection algorithms assume the existence of a single set of "optimal" features. However, in reality, this need not necessarily be the case and there could be several, distinct or overlapping, (minimal) subsets of features that might all explain the output of interest equally well given a particular loss function. We will refer to these equivalent minimal subsets as *signatures*, and the occurence of multiple signatures as *signature multiplicity* [6]. This phenomenon arises naturally in the presence of correlated or redundant features on a pairwise basis, but multiplicity can also occur at the level of signatures of larger sizes. For some loss function, signature multiplicity can be related to the existence of multiple markov boundaries for the target variable [6]. The study of signature multiplicity, and its effect on feature selection is at the moment only in its childhood, and so far studies have mainly focused on the microarray domain [1,6].

As standard feature ranking methods are not designed to cope with multiple signatures, they often interleave the features from the different signatures. Thus, thresholding this ranking does not even ensure to give a single valid and/or minimal signature. Furthermore, signature multiplicity might have a detrimental effect on the stability of feature selection methods, as small perturbations on the training set can result in large deviations regarding the ranking of features.

In this work, we investigate the impact of signature multiplicity on tree-based ensemble methods and we propose a simple post-processing method based on clustering to retrieve multiple signatures from the individual rankings provided by individual trees in a randomized tree ensemble.

2 Exploring individual tree rankings

Classification and regression trees are non-parametric supervised learning methods that learn an input-output model in the form of a tree, combining elementary tests defined on the input features. Because of their high variance, they are typically exploited in the context of ensemble methods such as bagging or random forests. A feature importance measure can be derived in different ways from a tree. In this work, we restrict ourselves to the importance obtained by summing the impurity reduction score at each tree node where this feature is used to split[3]. These importance scores are then averaged over several trees to yield a more stable score.

While one is often interested only in the global ranking obtained by averaging the individual rankings, in the presence of multiple signatures, one can reasonably expect that each tree in an ensemble will highlight a distinct signature. Indeed, since each tree is built greedily in a top-down fashion, the selection of a feature, or group of features, in a tree branch will decrease the probability to select redundant features at deeper nodes, which will favor the appearance of features from only one signature in each tree. In addition, because of randomization, one can also expect the selected signature to be different from one tree to another.

To check this hypothesis, we carried out experiments on the TIED dataset, an artificial dataset, specifically designed to contain multiple signatures [5]. The TIED dataset was generated from a bayesian network containing 1000 discrete variables, including the four-valued target. By construction, each of the 72 signatures contains 5 variables and belongs to $\{9\} \times \{4, 8\} \times \{11, 12, 13\} \times \{18, 19, 20\} \times \{1, 2, 3, 10\}$. The upper left graph of Figure 1 shows a heatmap representing 1000 tree rankings obtained with bagging (x-axis) for the top 20 features (y-axis) in the global ranking. Features are ranked top-down according to their global importances and rankings have been ordered by hierarchical clustering (dendrogram not shown). This heatmap clearly highlights the existence of groups of rankings each corresponding to one of the signatures. While the global ranking introduces the redundant features by block (e.g., features 1,2,3, and 10 are the top 4 features which are redundant by construction), each individual ranking usually contains only one feature per group. We obtained similar results on other artificial datasets.

3 Towards an automatic identification of signatures

Assuming that we are looking for K signatures, the analysis in the previous section suggests a simple approach for retrieving the multiple signatures from T feature importance vectors; Use any clustering algorithm (k-means in our experiments) to determine K clusters of weight vectors. Then, average the weight vectors in each of the clusters, and rank the features according to their average weight. To evaluate the quality of a given signature, a model is rebuilt with any

[3] A feature not appearing in a tree receives a zero importance.

supervised learning method using the top m features in each cluster for increasing values of m. When there are multiple signatures, we expect that the model obtained from each cluster will be at least equally good as a model learned in the same manner from the global ranking, i.e. the ranking obtained by averaging over all trees, and not over the clustered ones. To determine the optimal value of the number of clusters, we propose to proceed as follows: several values of K are compared, and the one that maximizes the difference over all values of m between the error obtained from the global ranking and the average error over the clusters is considered as optimal.

We carried out experiments with this approach on the TIED dataset. T was fixed to 1000, and the values explored for K were $\{2, 3, 5, 10, 15\}$. Features were ranked using a bagged ensemble of trees and the evaluation was done using ensembles of (100) totally randomized trees [2]. The latter method is not robust to the introduction of irrelevant features and is thus appropriate to determine minimal signatures. For the evaluation of signatures, we used 20 repetitions of a 90%-10% split of data in training and test, with the feature ranking computed only on the training sample, so as to avoid any selection bias.

The bottom left graph of Figure 1 shows in red the evolution of the error with the number of features m taken in their order in the global ensemble ranking, and in green the average error over all cluster rankings, for the value of $K = 15$ selected as just described. Blue curves show for each value of m respectively the minimal and maximal error obtained over all clusters. This graph shows that the cluster signatures are all very good and much better than the global signature for small values of m.

4 Experiments with microarray data

We have applied the same approach on several microarray datasets related to two families of problems: biomarker discovery for disease classification and regulatory network inference [3]. We only report below the results obtained on one representative problem. The graphs on the right in Figure 1 were obtained from microarray data when trying to discover the regulators of gene `tyrP` of *E. coli* using the same procedure and dataset as in [3]. The protocol was exactly the same as for the experiments on the TIED dataset.

The heatmap clearly highlights the diversity and complexity of the signatures, with for example the top feature from the global ranking not being used in many single rankings. The optimal number of clusters as determined automatically is here 5 and it leads to five signatures that are all (slightly) better than the global one.

5 Conclusion and future works

The discovery of multiple signatures is a challenging topic in the context of feature selection. In this work, we investigate the effect of signature multiplicity on tree-based feature rankings. We show that looking at individual tree rankings in an ensemble could highlight the existence of multiple signatures and we propose

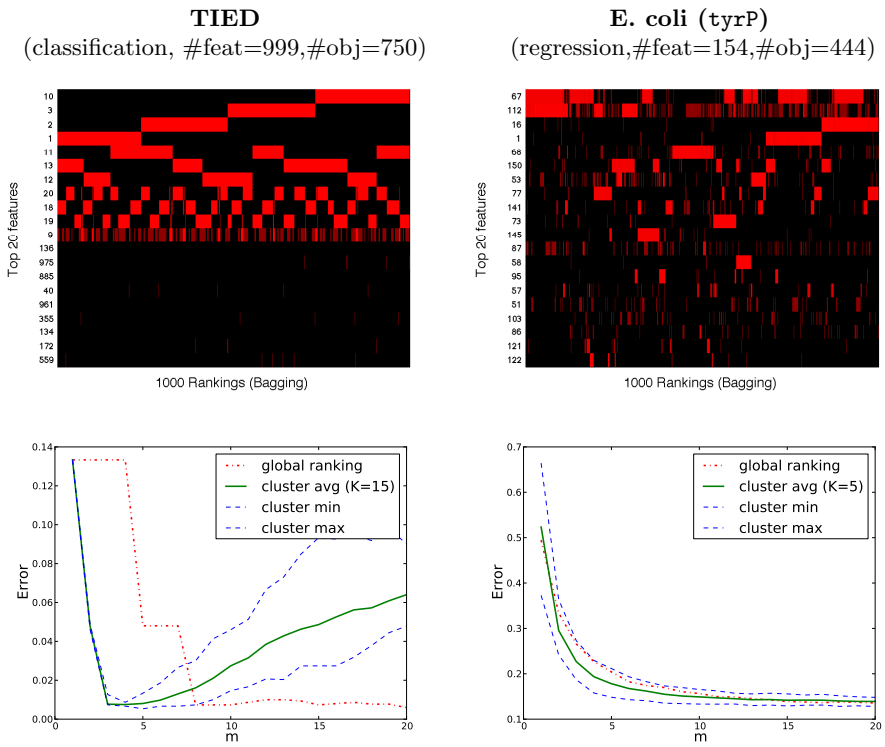

Fig. 1. Results on artificial and real datasets

a simple post-processing method based on clustering that can return smaller signatures with better predictive performance than signatures derived from the global tree ranking at almost no additional cost. In future work, we would like to explore alternative ways to extract multiple signatures from an ensemble of randomized feature rankers (not restricted to trees) and determine a measure of the multiplicity in a given dataset.

Acknowledgments

Pierre Geurts is a research associate with FNRS and Yvan Saeys is a postdoctoral fellow with FWO. This work is partially supported by the Interuniversity Attraction Poles Programme (IAP P6/25 BIOMAGNET), initiated by the Belgian State, Science Policy Office and by the European Network of Excellence, PASCAL2.

References

1. L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171, 2005.

2. P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 36(1):3–42, 2006.
3. V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts. Inferring regulatory networks from expression data using tree-based methods. *Plos ONE*, 5(9):e12776, sept 2010.
4. Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.
5. A. Statnikov and C. Aliferis. Tied: An artificially simulated dataset with multiple markov boundaries. *Journal of Machine Learning Research Workshop Conference & Proceedings*, 2009.
6. A. Statnikov and C. Aliferis. Analysis and computational dissection of molecular signature multiplicity. *Plos Comput Biol*, 6(5):e1000790, May 2010.