BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 10 2011, pages 1442–1443
doi:10.1093/bioinformatics/btr157

Gene expression

Advance Access publication March 30, 2011

pathClass: an R-package for integration of pathway knowledge into support vector machines for biomarker discovery

Marc Johannes[1,*], Holger Fröhlich[2], Holger Sültmann[1] and Tim Beißbarth[3,*]

[1]German Cancer Research Center, Cancer Genome Research, Im Neuenheimer Feld 460, 69120 Heidelberg, [2]Bonn-Aachen International Center for IT, Algorithmic Bioinformatics, Dahlmannstrasse 2, 53113 Bonn and [3]University Medical Center Göttingen, Medical Statistics, 37099 Göttingen, Germany

Associate Editor: David Rocke

ABSTRACT

Summary: Prognostic and diagnostic biomarker discovery is one of the key issues for a successful stratification of patients according to clinical risk factors. For this purpose, statistical classification methods, such as support vector machines (SVM), are frequently used tools. Different groups have recently shown that the usage of *prior* biological knowledge significantly improves the classification results in terms of accuracy as well as reproducibility and interpretability of gene lists. Here, we introduce `pathClass`, a collection of different SVM-based classification methods for improved gene selection and classfication performance. The methods contained in `pathClass` do not merely rely on gene expression data but also exploit the information that is carried in gene network data.

Availability: `pathClass` is open source and freely available as an R-Package on the CRAN repository at http://cran.r-project.org

Contact: m.johannes@dkfz-heidelberg.de; tim.beissbarth@ams.med.uni-goettingen.de

Received and revised on February 5, 2011; accepted on March 23, 2011

1 INTRODUCTION

Microarray studies are commonly used in clinical cancer research to investigate molecular profiles associated with tumor development and progression. Several thousand genes are measured to identify predictive or prognostic gene signatures comprising only a couple of genes. The goal is to use these gene signatures for stratification of patients according to clinical relevant endpoints.

Classification algorithms are frequently used to identify gene signatures. The support vector machine (SVM, Boser *et al.* 1992) is a well-known example of such a classification algorithm. Several groups have proposed the so-called feature / gene selection methods for the SVM since the algorithm does not offer an embedded feature selection mechanism. Despite the identification of biomarkers, the aim of feature selection is to reduce the dimensionality of the feature space in order to avoid the so-called *curse of dimensionalty* (Bellman, 1961), that is overfitting. Overfitting occurs when the number of features (genes) is large and the amount of samples (i.e. patients) is comparatively small, which is often the case when performing microarray studies. Thus, the algorithm can easily find a hyperplane that separates the training examples. However, the generalization performance of such a classifier will be poor and the stability of identified gene signatures will be low (Ein-Dor *et al.*, 2005).

In order to overcome these problems, several groups have recently proposed to adapt classification methods in such a way that the algorithms can benefit from using *prior* biological knowledge (Chuang *et al.*, 2007; Johannes *et al.*, 2010; Rapaport *et al.*, 2007). Common sources of such knowledge are databases that contain pathway information or protein–protein interactions for a review see Porzelius *et al.* (2011).

`pathClass` aims at providing the user with comprehensive implementations of these methods in a unified framework in order to allow easy and transparent benchmarking. To our knowledge, it is the first package implementing several SVM-based algorithms that are capable of incorporating network knowledge into the classification process. It is, however, worth mentioning that all methods available in `pathClass` have previously been published and shown their predominance over standard algorithms. For benchmarking of the more 'classical' methods not using *prior* knowledge, the user is referred to the packages `CMA` (Slawski *et al.*, 2008) and `MCRestimate` (Ruschhaupt *et al.*, 2004). A boosting approach that is capable of using *prior* knowledge (Binder and Schumacher, 2009) can be found in the package `GAMBoost`, which is also available on CRAN.

2 PACKAGE FEATURES

To this end, the *prior* knowledge used by the algorithms implemented in `pathClass` is represented as a network structure or graph, i.e. it carries information on the connectivity of features. In the following sections, the different algorithms and the way they use the network data are briefly explained.

For illustration purposes, we show a benchmark of all three algorithms in Figure 1. However, it is important to note, that these results are specific for this particular dataset. All steps needed to reproduce the result of Figure 1 are given in the package vignette.

2.1 Reweighted Recursive Feature Elimination

We recently proposed an extension of SVM-RFE (Guyon *et al.*, 2002), called Reweighted Recursive Feature Elimination (RRFE), which exploits both gene expression data as well as pathway knowledge (Johannes *et al.*, 2010). The assumption of RRFE is that a gene which is not differentially expressed but connected to other genes that show a differential expression should have a higher

*To whom correspondence should be addressed.

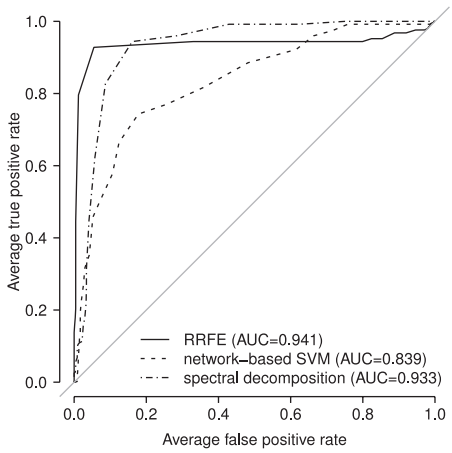

Fig. 1. Benchmarking of the implemented algorithms. The results were obtained by a five-times repeated 10-fold cross-validation on the dataset from Golub *et al*. (1999).

impact on the classifier compared to those genes that do not have any connections to other deregulated genes.

We used a modified version of Google's PageRank algorithm, called GeneRank (Morrison *et al.*, 2005), to calculate a ranking of the features based on fold change and network centrality. In a next step, this ranking is combined with the RFE ranking criterion. Based on this combination, features are ordered and 90% of the most important genes are kept for the next iteration. This procedure is then repeated until only one feature is left. Afterwards, the model with the best performance is chosen.

2.2 Network-based SVM

The approach introduced by Zhu *et al.* (2009) treat neighboring genes is called network-based SVM and uses a network-based penalty which leads to a grouped variable selection. This variable selection is archieved by penalizing the SVM objective function with an F_∞-norm, instead of the commonly used L_1 or L_2 penalization. This norm forces the simultaneous selection or elimination of a group of features from the same pathway. Zhu *et al.* (2009) treat neighboring genes in a graph as a group and construct their network-based penalty as the sum of F_∞-norms of groups of neighboring genes pairs.

2.3 Incorporating network knowledge by spectral decomposition

The method by Rapaport *et al.* (2007) defines a new metric for gene expression measurements by using diffusion kernel maps. Their assumption is that most biologically relevant information is captured in the low-frequency component of expression profiles. Hence, the projection of the low-frequency component of an expression vector on the gene metabolic network should reveal areas of positive and negative expression on the graph that are likely to correspond to the activation or inhibition of specific branches of the graph.

In `pathClass`, we combined this method with a recursive feature elimination, hence we allow to perform a feature selection when using this algorithm. However, the user also has the possibility to use the original version.

3 SUMMARY

We introduced `pathClass` a comprehensive package for classification with *prior* knowledge of feature connectivity. Up to now, it contains three SVM-based classification methods capable of using *prior* knowledge in the form of network data represented as a graph. The methods can be used directly or in a repeated cross-validation setting, which helps to estimate the average classification accuracy. The package is able to run the cross-validation in parallel and thus exploit the architecture of modern multicore computers or computing clusters. A detailed step-by-step tutorial as well as a comparison of the different methods on a cancer dataset is given in the package vignette, which is also available on CRAN.

Funding: This project was supported by the German Federal Ministry of Education and Science in the framework for medical genome research (NGFN, IG-Prostate Cancer, 01GS0890), further by the Clinical Research Group 179 through the DFG. The authors are responsible for the contents of this publication.

Conflict of Interest: none declared.

REFERENCES

Bellman,R. (1961) *Adaptive Control Processes*. Princeton University Press, Princeton, New Jersey.

Binder,H. and Schumacher,M. (2009) Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. *BMC Bioinformatics*, **10**, 18.

Boser,B.E. *et al*. (1992) A training algorithm for optimal margin classifiers. In *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, ACM, New York, NY, USA, pp. 144–152.

Chuang,H.-Y. *et al*. (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, **3**, 10.

Ein-Dor,L. *et al*. (2005) Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**, 171–178.

Golub,T.R. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Guyon,I. *et al*. (2002) Gene selection for cancer classification using support vector machines. *Mach. Learn.*, **46**, 389–422.

Johannes,M. *et al*. (2010) Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. *Bioinformatics*, **26**, 2136–2144.

Morrison,J.L. *et al*. (2005) Generank: using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics*, **6**, 233.

Porzelius,C. *et al*. (2011) Leveraging external knowledge on molecular interactions in classification methods for risk prediction of patients. *Biomet. J.*, **53**, 190–201.

Rapaport,F. *et al*. (2007) Classification of microarray data using gene networks. *BMC Bioinformatics*, **8**, 35.

Ruschhaupt,M. *et al*. (2004) A compendium to ensure computational reproducibility in high-dimensional classification tasks. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article 37.

Slawski,M. *et al*. (2008) Cma - a comprehensive bioconductor package for supervised classification with high dimensional data. *BMC Bioinformatics*, **9**, 439.

Zhu,Y. *et al*. (2009) Network-based support vector machine for classification of microarray samples. *BMC Bioinformatics*, **10** (Suppl. 1), S21.