BIOINFORMATICS APPLICATIONS NOTE

Vol. 25 no. 8 2009, pages 1076–1077
doi:10.1093/bioinformatics/btp094

Gene expression

Supervised feature selection in mass spectrometry-based proteomic profiling by blockwise boosting

Jan Gertheiss* and Gerhard Tutz

Department of Statistics, Ludwig-Maximilians-Universität, Munich D-80799, Germany

Received on October 15, 2008; revised on January 30, 2009; accepted on February 16, 2009

Advance Access publication February 20, 2009

Associate Editor: John Quackenbush

ABSTRACT

Summary: When feature selection in mass spectrometry is based on single *m/z* values, problems arise from the fact that variability is not only in vertical but also in horizontal direction, i.e. also slightly differing *m/z* values may correspond to the same feature. Hence, we propose to use the full spectra as input to a classifier, but to select small groups – or blocks – of adjacent *m/z* values, instead of single *m/z* values only. For that purpose we modify the LogitBoost to obtain a version of the so-called blockwise boosting procedure for classification. It is shown that blockwise boosting has high potential in predictive proteomics.

Availability: R-code is freely available at http://www.statistik.lmu.de/~gertheiss/research.html.

Contact: jan.gertheiss@stat.uni-muenchen.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Methods for feature selection in mass spectrometry (MS)-based predictive proteomics with categorial outcome can be roughly categorized into one of two groups. The first set of approaches uses the full spectrum as input to a classifier that is able to select variables. Thus features are selected in terms of single *m/z* values. Due to horizontal variability, however, results are difficult to interpret, because also slightly differing *m/z* values may correspond to the same feature. That is why, for example, Hoefsloot *et al.* (2008) manually clustered the selected *m/z* values at the end of the analysis. Alternatively, feature (pre)selection in terms of peak detection and peak alignment (Tibshirani *et al.*, 2004) is often performed before employing a classifier. The workflow is nicely summarized in Barla *et al.* (2008). But for the latter approach initial feature selection is done in an unsupervised way.

An early application of boosting to MS data is Yasui *et al.* (2003). The procedure is based on (unsupervised) dichotomization of intensities into peak/non-peak binary data (to address unreliability in peak heights). In contrast, in the following a modification of the LogitBoost (Friedman *et al.*, 2000) using the raw spectra is presented which yields a version of so-called blockwise boosting for dichotomous outcomes. To attack the problem of horizontal variability, we propose to select small subsets of adjacent *m/z* values instead of single ones. For regularization a difference penalty is employed. The high potential of this approach is demonstrated on a publicly available dataset from proteomics. All computations are carried out using R (R Development Core Team, 2007).

2 METHODS

One building block of the presented approach is boosting, which has been introduced by Freund and Schapire (1996). Boosting has been shown to be a powerful classification tool, especially for high-dimensional problems. A classifier $\mathcal{C}(x)$ as a function of predictors x is based on an ensemble of so-called *weak* (or *base*) learners $f^{(m)}(x)$. We will use a modification of the LogitBoost (Friedman *et al.*, 2000). The LogitBoost was used, e.g. by Dettling and Bühlmann (2003) to discriminate microarrays of gene expression data. Given dichotomous responses $y_i \in \{0,1\}$, each single $f^{(m)}$ is fitted by weighted least squares using weights

$$w_i^{(m)} = p^{(m-1)}(x_i)\cdot(1-p^{(m-1)}(x_i))$$

and working responses

$$z_i^{(m)} = \frac{(y_i - p^{(m-1)}(x_i))}{w^{(m)}(x_i)},$$

with $p^{(m)}(x_i)$ denoting the estimated probability of observation i belonging to class 1. For details of the algorithm, see Friedman *et al.* (2000) or Dettling and Bühlmann (2003).

In common componentwise boosting the base learner f is a function of only one component of the (p-dimensional) vector $x_i = (x_{i1},\dots,x_{ip})^T$. In addition to the fitting of f a selection step is included which selects the component that produces minimum (weighted)-squared loss. Predictors that are never selected are not used for classification. In contrast, we do not select single variables but groups $x^{(s)}$ of k adjacent predictors, i.e. *m/z* values. Let $x_i^{(s)}$ denote the vector $(x_{is},\dots,x_{i,s+k-1})^T$, so that one has $s = 1,\dots,p-k+1$ groups of adjacent *m/z* values. For the regression function f we fit linear combinations of these (k adjacent) predictors, i.e. $f(x) = \beta^T x^{(s)}$. That means $f^{(m)}$ from above is represented by $\beta^{(m)}$ and a distinct block $x_i^{(s)}$. The selection step now refers to blocks instead of single components. Since adjacent measurements are highly correlated, simple (weighted) least squares estimation of β is not recommended. So we penalize the sum of squared differences between adjacent coefficients β_j. This makes good sense, since features are assumed to cover more than a single *m/z* value, and measurements at adjacent *m/z* values should be linked to the class label in a similar way. Hence, $\beta^{(m)}$ is a generalized weighted ridge estimator. In matrix notation one has

$$\beta^{(m)} = (X^{(s_m)T}W^{(m)}X^{(s_m)} + \lambda\Omega)^{-1}X^{(s_m)T}W^{(m)}z^{(m)},$$

with $X^{(s_m)T} = (x_1^{(s_m)},\dots,x_n^{(s_m)})$, $z^{(m)} = (z_1^{(m)},\dots,z_n^{(m)})^T$, $W^{(m)} = \mathrm{diag}(w_1^{(m)},\dots,w_n^{(m)})$. In iteration m, block s_m producing minimum error is selected. The penalty matrix is $\Omega = D^T D$, with $D_{11} = D_{rr} = D_{k+1,k} = 1$,

*To whom correspondence should be addressed.

$D_{r,r-1} = -1$ and 0 otherwise, $r = 2,\ldots,k$. By upper left and lower right 1s in D differences to zero coefficients of neighboring but not selected *m*/*z* values are penalized. For details of the blockwise boosting procedure, see Tutz and Gertheiss (2009) where a blockwise procedure for continuous response is used to select variable blocks in signal regression.

The result of the described algorithm can be characterized as a logistic regression model with coefficient vector β of length p, and many coefficients being zero. Blocks of non-zero coefficients have at least length k and represent selected *m*/*z* values. Since k is fixed and distances between adjacent *m*/*z* values are larger for higher *m*/*z* values, features selected in regions of higher *m*/*z* values tend to be wider, too. This, however, makes sense since features are truly larger for higher *m*/*z* ranges (cf. Barla *et al.*, 2008). Alternatively, block sizes may also be held constant on another scale, e.g. the logarithm of *m*/*z* values.

Another (difference) penalty approach applied to MS data is in Goeman (2008), where a prior distribution of (logistic) regression coefficients is derived from a stationary autoregressive process. The used Bayes point estimate (the mode of the posterior density), however, leads (almost surely) to non-zero coefficients. Hence, no feature selection is performed.

3 RESULTS

To evaluate our method we use data from Petricoin *et al.* (2002), which is publicly available from the National Cancer Institute via http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp. The data are already baseline corrected. The goal is to discriminate prostate cancer from benign prostate conditions by proteomic pattern diagnosis. All in all 69 SELDI-TOF blood serum spectra from cancer patients and 253 from patients with benign conditions are given. See Petricoin *et al.* (2002) for further information.

From the data at hand we randomly selected test data consisting of 150 observations. On the remaining data our BlockBoost was trained and test set labels were predicted. This procedure was independently repeated 20 times. For each training sample block size k and penalty λ were determined via (3-fold) cross-validation. For comparison we also considered the performance of usual componentwise LogitBoost (i.e. the special case with $k = 1$), for (3-fold) cross-validated number of boosting iterations (Boost, CV) and for the optimal number producing minimum test error (Boost,opt). Additionally, we give the error resulting from regularized linear discriminant analysis (R-package `rda`; Guo *et al.*, 2005) as proposed by Guo *et al.* (2007), if oracle tuning parameters (minimizing test errors) are chosen (RDA,opt). Note, oracle tuning parameters give a (somewhat unrealistic) lower bound for test set errors. Since the full spectrum is used as input, componentwise boosting and RDA select features in terms of single *m*/*z* values.

Figure 1 gives a summary of test set errors in comparison to blockwise boosting with (3-fold) cross-validated number of iterations (BlockBoost,CV) and for the optimal number of iterations (BlockBoost,opt). It is seen that even with iterations chosen by CV BlockBoost is competitive to RDA with oracle tuning parameters. The original LogitBoost was outperformed distinctly. Apparently selecting blocks of *m*/*z* values (instead of single ones) improves prediction accuracy.

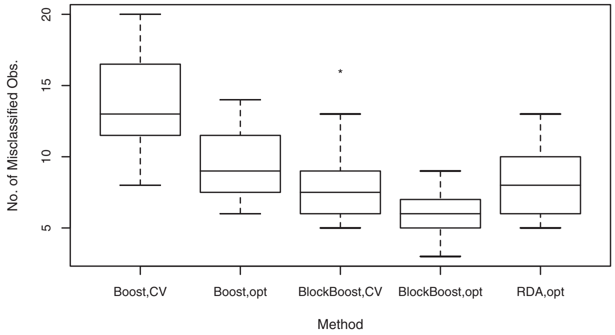

Fig. 1. Errors of componentwise and blockwise boosting, each with cross validated as well as optimal number of boosting iterations, and regularized discriminant analysis with optimal tuning parameters.

4 SUMMARY AND DISCUSSION

We presented a modification of the LogitBoost algorithm that is especially suited to the challenges of MS-based proteomic profiling (a corresponding R package is in preparation). Superiority over usual componentwise boosting was illustrated. In every iteration blocks instead of single *m*/*z* values are selected and differences between effects of adjacent *m*/*z* values are penalized. Block size k and penalty parameter λ, however, need to be chosen by the user. Block size k should depend on the quality of the spectra, like tuning parameters in unsupervised feature preselection. Penalty λ (and also k) may be determined via (k-fold) cross-validation, as done here. The same applies to the number of boosting iterations M. In general, however, the method seems to be quite resistant to overfitting caused by too many iterations (cf. Dettling and Bühlmann, 2003).

Conflict of Interest: none declared.

REFERENCES

Barla,A. *et al.* (2008) Machine learning methods for predictive proteomics. *Brief. Bioinform.*, **9**, 119–128.

Dettling,M. and Bühlmann,P. (2003) Boosting for tumor classification with gene expression data. *Bioinformatics*, **19**, 1061–1069.

Freund,Y. and Schapire,R.E. (1996) Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, CA, pp. 148–156.

Friedman,J.H. *et al.* (2000) Additive logistic regression: a statistical view of boosting. *Ann. Stat.*, **28**, 337–407.

Goeman,J.J. (2008) Autocorrelated logistic ridge regression for prediction based on proteomics spectra. *Stat. Appl. Genet. Mol. Biol.*, **7**, Article 10.

Guo,Y. *et al.* (2005) *rda: Shrunken Centroids Regularized Discriminant Analysis*. R package version 1.0.

Guo,Y. *et al.* (2007) Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, **8**, 86–100.

Hoefsloot,H.C.J. *et al*. (2008). A classification model for the Leiden proteomics competition. *Stat. Appl. Genet. Mol. Biol.*, **7**, Article 8.

Petricoin,E.F. *et al*. (2002) Serum proteomic patterns for detection of prostate cancer. *J. Natl Cancer Inst.*, **94**, 1576–1578.

R Development Core Team (2007) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Tibshirani,R. *et al.* (2004) Sample classification from protein mass spectrometry, by 'peak probability contrasts'. *Bioinformatics*, **20**, 3034–3044.

Tutz,G. and Gertheiss,J. (2009) Feature extraction in signal regression: a boosting technique for functional data regression. *J. Comput. Graphical Stat.*, (accepted for publication).

Yasui,Y. *et al*. (2003) A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics*, **4**, 449–463.