Statistical Applications in Genetics and Molecular Biology

Volume 7, *Issue* 2 2008 *Article* 5

COMPETITION ON CLINICAL MASS SPECTROMETRY BASED PROTEOMIC DIAGNOSIS

Developing a Discrimination Rule between Breast Cancer Patients and Controls Using Proteomics Mass Spectrometric Data: A Three-Step Approach

A. Geert Heidema, *Maastricht University; National Institute of Public Health and the Environment; Wageningen University and Research Centre*

Nico Nagelkerke, *United Arab Emirates University*

Recommended Citation:
Heidema, A. Geert and Nagelkerke, Nico (2008) "Developing a Discrimination Rule between Breast Cancer Patients and Controls Using Proteomics Mass Spectrometric Data: A Three-Step Approach," *Statistical Applications in Genetics and Molecular Biology*: Vol. 7: Iss. 2, Article 5.
DOI: 10.2202/1544-6115.1341
Available at: http://www.bepress.com/sagmb/vol7/iss2/art5

Developing a Discrimination Rule between Breast Cancer Patients and Controls Using Proteomics Mass Spectrometric Data: A Three-Step Approach

A. Geert Heidema and Nico Nagelkerke

Abstract

To discriminate between breast cancer patients and controls, we used a three-step approach to obtain our decision rule. First, we ranked the mass/charge values using random forests, because it generates importance indices that take possible interactions into account. We observed that the top ranked variables consisted of highly correlated contiguous mass/charge values, which were grouped in the second step into new variables. Finally, these newly created variables were used as predictors to find a suitable discrimination rule. In this last step, we compared three different methods, namely Classification and Regression Tree (CART), logistic regression and penalized logistic regression. Logistic regression and penalized logistic regression performed equally well and both had a higher classification accuracy than CART. The model obtained with penalized logistic regression was chosen as we hypothesized that this model would provide a better classification accuracy in the validation set. The solution had a good performance on the training set with a classification accuracy of 86.3%, and a sensitivity and specificity of 86.8% and 85.7%, respectively.

Author Notes: We thank the organizers of the classification contest for their effort and for providing us the opportunity to participate. We also thank T. Travis for allowing us to make use of the computer cluster at the Rowett Research Institute for the random forests analyses.

Introduction

To develop a decision rule that discriminates well between individuals with breast cancer (cases) and individuals without (controls), we applied a three-step approach, viz. i) variable selection/reduction; ii) synthesis of new variables by grouping selected variables to make use of the correlation structure; and iii) actual discrimination and classification on the basis of the variables developed in step ii. In step i, in order to make a reduction in the large numbers of variables present in the dataset we applied random forests. Random forests appears to be the most appropriate method for prioritizing variables and selection of a small set of most important variables, i.e. variables that appear to hold promise of having discriminatory power *in conjunction with other variables*. The latter clause is important, as selection of variables on the basis of individual discriminatory power is unsatisfactory in view of the correlation between many variables. Also, as it is based on cross-validation, one can expect the selected variables to work (i.e. to be discriminatory) not only in the training data set, but also from validation data sets collected from different patients and controls from the same population. Random forests was developed by Breiman (2001). This machine learning approach has proven to have excellent performance in many classification tasks, and is now available as an off-the-shelf method. Random forests has shown to outperform other classification methods in applications to microarray data (Diaz et al., 2006) and mass spectrometry data (Wu et al., 2003). One of the features of random forests is that it provides a measure of importance for each of the variables, referred to as the importance index. The importance index was used to prioritize and select the variables that best discriminate between cases and controls.

Contiguous variables with approximately identical mass are highly correlated due to physical properties and the smoothing applied in the pre-processing steps. Therefore, it can be expected that among the highly prioritized variables highly correlated, contiguous variables will be present. Therefore, in the second step we searched over the most important variables whether groups of highly correlated variables would be present. These highly correlated variables can then be grouped into a new variable. In the third step we used these newly created variables as predictors and applied different methods to find a suitable discrimination rule. The methods compared at this step are Classification and Regression Tree (CART) (Breiman et al.,1984), logistic regression and penalized logistic regression (Le Cessie et al., 1992, Firth, 1993). The decision rules obtained by the different methods and their classification performance were compared and the decision rule with the best performance was finally chosen to be applied to the validation set.

Method

Step 1: Prioritization of variables by random forests

To reduce the number of variables to be used to make a decision rule, we used random forests to prioritize and select the apparently best discriminating candidate variables in the first step.

In the random forests approach, an ensemble of tree models is used to predict case-control status (bagging). Each tree recursively splits the total dataset into smaller and more homogenous subgroups of cases and controls, whereby the total sample for each tree is obtained by bootstrap sampling. With bootstrap sampling, sampling is performed with replacement and some individuals are sampled more than once while others are left out, while keeping the bootstrap sample size the same as that of the original sample. This method involves cross-validation as the (bootstrap) sampled observations are used to construct the classification tree whereas a prediction is obtained for each left-out individual. Aggregating the predictions over the different trees in which the individual was left-out, a prediction for this individual is obtained for the ensemble of trees, which is called the forest. The proportion of misclassified cases and controls provides the prediction error of the forest. Another important feature is that the predictor that gives the best partitioning in cases and controls at a certain split is not selected from the total number of predictors but from a smaller random sample of predictors. This parameter is referred to as m_{try}. We used the default value for m_{try}, which is the square root of the number of variables to be analyzed in the dataset (in this dataset equal to 105). Multiple thousands of trees in the forest are needed to obtain stable estimates of the importance indices (Lunetta et al., 2004). Also, each tree captures only the possible interactions for the variables selected by that tree only, and large numbers of trees are required to capture as many interactions as possible. Therefore, the number of trees in the forest was set to 30,000 for each of the different analyses. We performed several analyses with random forests to verify whether the ranking of the variables by their importance index did not change over the different analyses. This is done by using different seed values for the different analyses (the seed value controls the random number generator).

Random forests provides an importance index for each variable by comparing the predictive performance of the forest for all variables with the predictive performance of the forest for all variables but with the values for one variable randomly permuted for the left-out individuals. Larger differences in the predictive performance indicate more important variables. Permuting the predictor values for the left-out individuals does not only remove the association between the permuted predictor and the outcome variable, but also the interaction

DOI: 10.2202/1544-6115.1341

effects of the permuted predictor with other predictors, if present. Thereby, the interactions of the predictor with other predictors are taken into account in the importance index. We used the importance index as a first step to prioritize and select the best discriminating variables.

For the random forests analyses we used the R-package randomForest written by Liaw and Wiener (Liaw et al., 2002, R Development Core Team, 2004), freely available from the CRAN website (http://cran.r-project.org/). Because the predictors are all of the same type, the random forests variable importance indices obtained with the randomForest R-package can be used (Strobl et al., 2007). This R-package is based on the original FORTRAN code from Breiman et al. (2003, freely available at www.stat.berkeley.edu/users/breiman/randomforests/).

Step 2: Grouping highly correlated variables

Among the highest prioritized variables we tried to exploit the correlation between variables. Adjacent variables were very highly correlated (generally $r>0.9$) which made adjacent variables almost duplicate measurement of the same underlying "parameter". If groups of adjacent variables were identified, we combined these variables into a new variable by taking the sum of the variables. Small "gaps" were ignored, i.e. variables of which both its neighbours were selected, were included even if that variable itself was not selected.

Step 3: Obtaining the decision rule

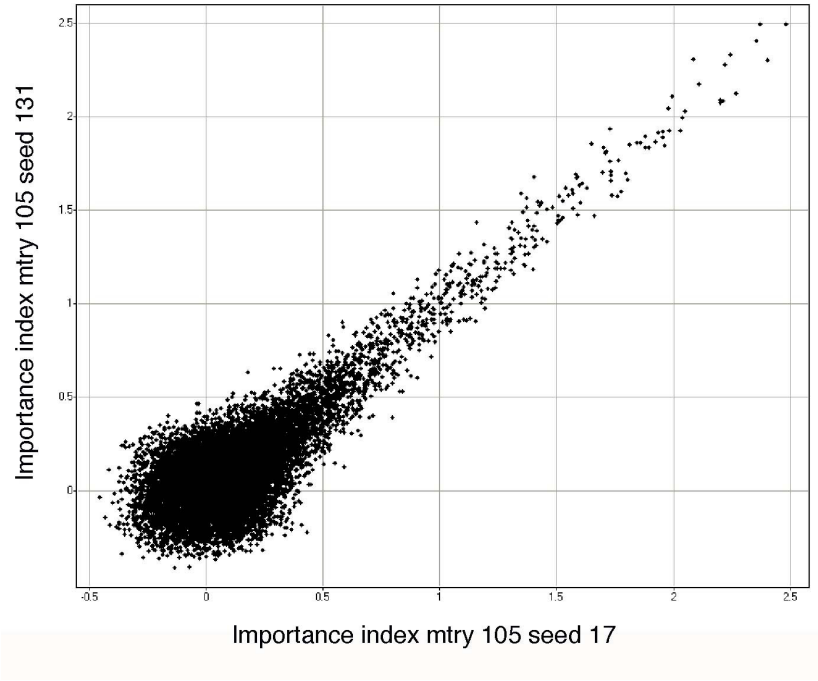
The newly created variables, each consisting of a group (sum) of highly correlated variables, were used as candidate predictors for case-control status. We tried the following methods to predict the individuals in the calibration dataset: CART, logistic regression and penalized logistic regression. For CART we used the program QUEST (Loh et al., 1997), which is freely available at http://www.stat.wisc.edu/~loh/quest.html. To perform logistic regression analysis, SPSS version 13.0 was used (SPSS, Inc., Chicago, Illinois). For penalized logistic regression, we applied the R-package brlr written by Firth (1993). This method penalizes the likelihood by the Jeffrey's prior, and has the effect of "mildly" shrinking parameter estimates to 0. The decision rule obtained by the method with the best classification performance was chosen to obtain the prediction for the individuals in the validation dataset.

Results

Step 1:

The prioritization of variables by random forests for two different seed values are shown in figure 1. For both seed values the same variables were highly prioritized. Thus random forests provides similar results over different analyses, indicating the robustness of the method.

Figure 1: Random forests results. Prioritization of m/z values by their importance index for two different seed values.



Step 2: Grouping highly correlated variables

As expected, visual inspection of the most important variables showed that highly prioritized m/z values consisted of different groups of contiguous variables. Therefore, we combined adjacent variables into a new variable, summing the scores of the individual variables. In this way, nine new variables were formed (see Box 1).

DOI: 10.2202/1544-6115.1341

Box 1: New variables (Y1-Y9) formed by summing the scores of adjacent individual variables. The numbers of the individual variables represent mass/charge values.
Y1 = v3454 + v3455 + v3456 + v3457 + v3458 + v3459 + v3460.
Y2 = v3496 + v3497 + v3498 + v3499.
Y3 = v3830 + v3831 + v3832 + v3833 + v3834 + v3835 + v3836.
Y4 = v3844 + v3845 + v3846 + v3847 + v3848 + v3849 + v3850 + v3851 + v3852 + v3853 + v3854 + v3855 + v3856 + v3857 + v3858.
Y5 = v3924 + v3925 + v3926 + v3927 + v3928 + v3929 + v3930.
Y6 = v6607 + v6608 + v6609 + v6610.
Y7 = v9531 + v9532 + v9533 + v9534 + v9535.
Y8 = v5380 + v5381.
Y9 = v6606 + v6607.

Step 3: Obtaining the decision rule

The nine variables obtained in step 2 were used as predictors to obtain a decision rule. At this step we compared the classification performance of CART, logistic regression and penalized logistic regression. Table 1 shows the classification accuracy for the different methods.

Table 1: Classification accuracy obtained with CART, logistic regression and penalized logistic regression.

Method	Classification accuracy (%)
CART	78.1
Logistic regression	86.3
Penalized logistic regression	86.3

Logistic regression and penalized logistic regression both had a higher classification accuracy compared to CART. For both logistic regression and penalized logistic regression neither interactions nor logarithmically transformed variables did improve the classification performance and were therefore not included in the final model. Both types of logistic regression performed equally well, in fact they gave identical classification results and we chose the model

obtained with penalized logistic regression as we hypothesized that this solution, as its estimators have been designed to have less bias, would give a better classification accuracy for the validation set. Thus, using penalized logistic regression, we obtained the following solution:
Logit{Pr(individual belongs to group (Cases))} = 1572.13 – 57.36·Y1 + 4.98·Y2 – 13.32·Y3 – 13.04·Y4 + 4.36·Y5 + 69.07·Y6 – 19.88·Y7 + 124.08·Y8 – 1191.26·Y9

Performance on training data

The cross-classification table is shown in table 2. The classification accuracy equals 86.3%. For sensitivity and specificity, very similar percentages were obtained (86.8% and 85.7%, respectively).

Table 2: Cross-classification table, based on the decision rule obtained with penalized logistic regression at the third step of the three-step approach.

	True Group Control	True group Cases
Assigned group Control	66	10
Assigned group Cases	11	66

Discussion

The three-step approach we applied to obtain a decision rule to discriminate between cases of breast cancer and controls has several advantages. The use of random forests in the first step has the advantage that the interdependence between variables is taken into account in the importance index, and therefore in the prioritization and selection of variables. Also, standard stepwise procedures tend to reject or select variables on the basis of their individual discriminating power, which may be far from optimal in a context of many highly correlated variables. The high correlation between (prioritized) variables is also taken into account in the second step by combining adjacent variables into new variables, with the idea that these variables essentially measure the same "peak" or other feature and that therefore the measurement errors of these new "sum" variables is less than that of individual variables. Furthermore, a clear interpretation of the predictors on which the decision rule is based can be made in the third step.

However, there are also limitations to this three-step approach. There is a methodological discrepancy or disconnect between using trees in variable

DOI: 10.2202/1544-6115.1341

selection and using logistic regression in the final (third step) analysis. As logistic regression was chosen because it clearly outperformed CART, one may conjecture whether in the selection phase a logistic regression approach, but one that would make use of bootstrapping and cross-validation in a similar way as Random Forests would have been better at selecting variables. Unfortunately, this method is not available off-the-shelf using readily available software. Also, the grouping of variables in the second step of our analysis was done, more or less *ad hoc*, by eye and hand, and therefore this step is not amenable to cross validation. This step should have been formalized and automated and used in cross validation. Finally, Jeffreys prior may well be too "flat" and heavier shrinkage might have yielded classifiers with great predictive efficiency.

Another limitation of our approach, or any other approach that is based on variable selection, is that it is based on an untested assumption, *viz.* that the classification problem is "sparse" in the sense that only a small minority of variables have any discriminating power and that the rest is essentially "noise". While this seems to be supported by the finding that only the importance index of variables with a high importance index tends to be reproducible across different runs of the random forests program, this is by no means certain. If this "sparseness" does not hold true then the additional discriminating power of many weakly informative variables is ignored by our approach. With such a limited sample size however, the task of making effective use of such variables would seem daunting.

A further limitation of our three step approach is lack of methodological coherence. This was largely due to our objective of developing an easy-to-apply discrimination score, and our idea that we had to take into account the correlation structure among (neighbouring) variables. However, this mixture of methods makes it harder to identify the causes of misclassifications. These are much easier to identify and perhaps correct when only a single method is used, for example random forests. For the application of only random forests the selection of variables for classification as performed by Diaz et al. (2006) could be used. Although this approach leads to a small set of variables that still has good performance, it does not lead to readily usable classification rules for clinical diagnostic purposes, and neither does it give rise to classification rules that are easy to interpret.

Finally we want to address the possible sensitivity of our approach to experimental effects. Plates on which the experiments were run were not known to us. Thus our method may be sensitive to "plate effects". If some plates yield (locally) systematically higher or lower values than other plates this may influence and bias the classification results. Perhaps, instead of new synthetic variables consisting of sums of variables, sums of contrasts, i.e. sums of signed variables, with as many positive as negative signs, would be less sensitive to such

plate effects. Preferably, perhaps, any variables with negative signs should be matched with, and chosen relatively close to variables with a positive sign. However, of course, such control variables should be chosen sufficiently distant to avoid strong correlations with their positive "matches".

References

Breiman L: Random forests. Machine learning 2001, 45:5-32.
http://www.springerlink.com/content/u0p06167n6173512/fulltext.pdf

Breiman L, Cutler A: Random forests. Version 4.0. 2003
[http://www.stat.berkeley.edu/users/breiman/RandomForests/]

Breiman L, Friedman JH, Olshen RA, Stone CJ: Classication and RegressionTrees, 1984, Wadsworth, Belmont.

Díaz-Uriarte R, Alvarez de Andrés S: Gene selection and classification of microarray data using random forest. BMC Bioinformatics 2006, 7:3.
http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=1363357&blobtype=pdf

Firth D: Bias reduction of maximum likelihood estimates. Biometrika 1993, 80:27–38. http://biomet.oxfordjournals.org/cgi/content/abstract/80/1/27

Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic-regression. Appl Stat J of the Royal Stat Soc Series C 1992, 41(1):191-201.
http://links.jstor.org/sici?sici=0035-9254(1992)41%3A1%3C191%3AREILR%3E2.0.CO%3B2-W

Liaw A, Wiener M. 2002. Classification and regression by random-Forest. Rnews 2:18-22. http://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf

Loh W-Y, Shih Y-S: Split selection methods for classification trees. Statistica Sinica 1997, 7:815-840.
http://www3.stat.sinica.edu.tw/statistica/oldpdf/A7n41.pdf

Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P: Screening large-scale association study data: exploiting interactions using random forests. BMC Genet 2004, 5:32.
http://www.biomedcentral.com/content/pdf/1471-2156-5-32.pdf

DOI: 10.2202/1544-6115.1341

R Development Core Team: R: A language and environment for statistical computing. 2004 [http://www.R-project.org]. R Foundation for Statistical Computing, Vienna,.

Strobl C, Boulesteiz A-L, Zeileis A, Hothorn T: Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 2007, 8:25. http://www.biomedcentral.com/content/pdf/1471-2105-8-25.pdf

Wu B, Abbott T, Fishman D, McMurray W, Mor G, Stone K, Ward D, Williams K, Zhao H: Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics 2003, 19:1636-1643. http://bioinformatics.oxfordjournals.org/cgi/screenpdf/19/13/1636