A classification model for the Leiden proteomics competition

Hoefsloot, H.C.J.; Smit, S.; Smilde, A.K.

Publication date
2008

Published in
Statistical Applications in Genetics and Molecular Biology

Statistical Applications in Genetics and Molecular Biology

Volume 7, *Issue* 2 2008 *Article* 8

COMPETITION ON CLINICAL MASS SPECTROMETRY BASED PROTEOMIC DIAGNOSIS

A Classification Model for the Leiden Proteomics Competition

Huub C. J. Hoefsloot* Suzanne Smit†

Age K. Smilde‡

*University of Amsterdam, h.c.j.hoefsloot@uva.nl

†University of Amsterdam, ssmit@science.uva.nl

‡University of Amsterdam, asmilde@science.uva.nl

A Classification Model for the Leiden Proteomics Competition

Huub C. J. Hoefsloot, Suzanne Smit, and Age K. Smilde

Abstract

A strategy is presented to build a discrimination model in proteomics studies. The model is built using cross-validation. This cross-validation step can simply be combined with a variable selection method, called rank products. The strategy is especially suitable for the low-samples-to-variables-ratio (undersampling) case, as is often encountered in proteomics and metabolomics studies. As a classification method, Principal Component Discriminant Analysis is used; however, the methodology can be used with any classifier. A data set containing serum samples from breast cancer patients and healthy controls is analysed. Double cross-validation shows that the sensitivity of the model is 82% and the specificity 86%. Potential putative biomarkers are identified using the variable selection method. In each cross-validation loop a classification model is built. The final classification uses a majority voting scheme from the ensemble classifier.

KEYWORDS: classification, curse of dimensionality, statistical validation, double cross-validation, principal component discriminant analysis, biomarker discovery, rank products

Introduction

The experiment involves two groups, a breast cancer group and a healthy control group. The control group consists of 77 persons and the breast cancer group of 76 women making a total of 153 persons. The measurements are performed on a MALDI-MS machine. The mass spectrum of each person consists of 11205 mass/charge (or m/z) values.

The question here at hand is whether it is possible to build a model that is capable of discriminating between the breast cancer group and the control group. To test this model a separate validation set was made available to us after we built the model. The status of persons in the test set was initially unknown to us and only after the model predictions were presented made available to us. The test set results are discussed in a separate paper.

The data analysis may start by building a discrimination model that separates the groups. The large number of variables in this setup generates modelling and validation challenges commonly referred to as the curse of dimensionality (Hastie et al., 2001) or undersampling. In short, the curse of dimensionality means that the number of samples needed to accurately describe a (discrimination) problem increases exponentially with the number of dimensions (variables) measured. In the situation at hand this is clearly the case. A good discrimination result for the original control-diseased problem can be simply a chance effect. Therefore the use of cross-validation is strongly advocated. Using cross-validation diminishes the effect that the model is specific for the data it is build from and will not perform well on new data, the so-called over-fitting phenomena.

Principal component discriminant analysis (PCDA) is used to discriminate between the groups of protein profiles. This method is a combination of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). In the PCA step the dimensionality of the data is reduced. In this lower dimensional space a LDA is performed. The cross-validation procedure generates several models in the form of discriminant vectors. A sample is classified preliminary by each of the models. The final classification is performed by majority voting.

From the models discriminating m/z values are selected using the rank products (Breitling et al., 2004) procedure. Both PCA and LDA are well known, simple and straight forward methods. PCDA is a combination of these two methods and it retains these favourable properties.

Methodology

A simple method for discrimination between two groups is linear discriminant analysis (LDA). Good discriminating directions are directions in *m/z* space in which the differences between the groups are large compared to the differences within the groups. In the two-group case it is the vector, $\boldsymbol{d}$, that maximizes

$$\frac{\boldsymbol{d'Bd}}{\boldsymbol{d'Wd}},$$

where $\boldsymbol{W}$ is the pooled within class sample covariance matrix, and $\boldsymbol{B}$ is the between class sample covariance matrix. The discriminating direction is the first eigenvector of the matrix $\boldsymbol{W}^{-1}\boldsymbol{B}$(Vandeginste et al., 1998*)*. Because in this case there are more *m/z* values than samples, the matrix $\boldsymbol{W}$ is singular. This means that $\boldsymbol{W}^{-1}$ does not exist and LDA cannot be applied directly. This problem can be overcome by using principal component analysis (PCA), which finds principal components to describe the data. These components are linear combinations of the original *m/z* values. The first principal component (PC) describes as much of the variation in the data as possible, the second describes as much of the remaining variation as possible, etcetera. By keeping only a few of the principal components we can reduce the dimensionality of the data to a point where LDA is applicable, while preserving most of the information in the data. In this application PCA is performed on the mean-centered data matrix; no scaling is applied. The number of components in the model can be decided upon using cross-validation. The combination of LDA with PCA yields principal component discriminant analysis (PCDA) (Hoogerbrugge et al., 1983; Howland and Park, 2004; Lilien et al., 2003; Smit et al., 2007; Ye et al., 2004).

The PCDA is combined with a double cross-validation (Mertens et al., 2006; Smit et al., 2007; Stone, 1974) approach. This is a nested validation scheme; the inner validation is used to determine the optimal number of principal components, see figure 1. The outer validation is used to find the cross validation error of the method.

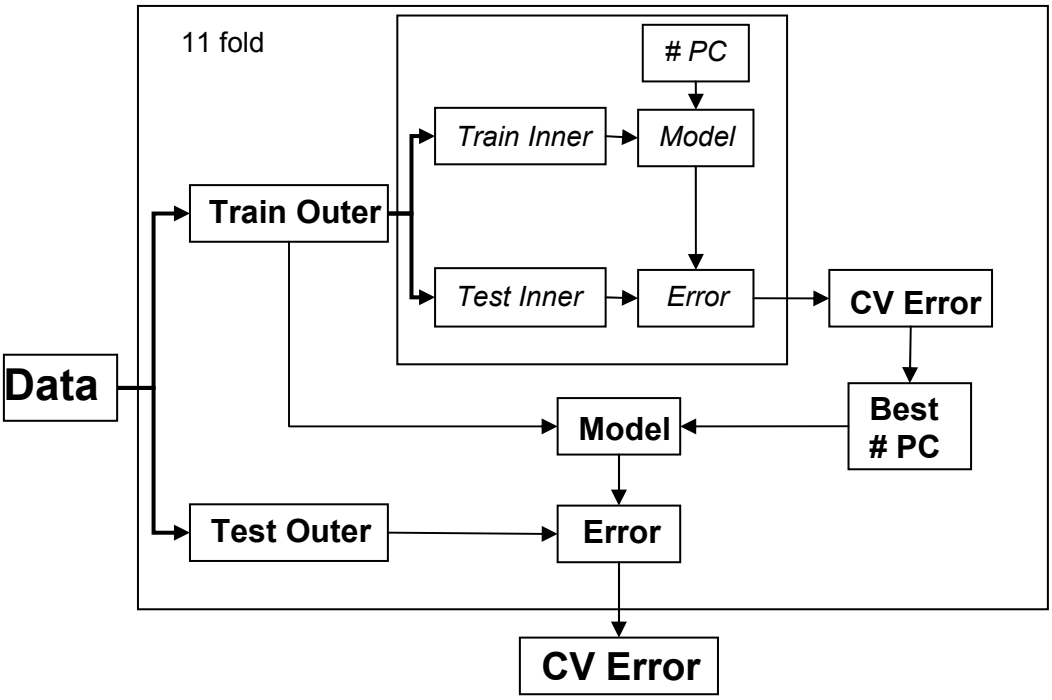

Figure 1: The double cross validation scheme. In the inner loop a cross validation is used to determine the number of principal components to be used. In the outer loop an 11 fold cross validation is used to determine the error rates.

The discriminant vector is calculated using PCDA with 11 fold outer cross-validation to determine the cross-validation errors and an inner cross-validation to determine the optimal number of principal components. The result of this procedure is 11 models each with a discriminant vector.

The entire double-cross procedure can be repeated with randomly chosen partitions of the outer training and test set. We performed the whole procedure (see figure 1) 10 times. Because a single run yields 11 discriminant vectors; 10 runs gives 110 discriminant vectors in total. The obtained models thus consist of 110 classifications rules. A sample is classified by every one of these 110 classification rules. The final classification is performed by majority voting. It is known from literature (Breiman, 1996) that aggregated predictors can have favorable properties. This is not the only reason for using this type of predictor. Our second purpose is to use this scheme to examine the 110 obtained discriminant vectors and study the variability of the model. In the ideal case the variability should be small; the model should not depend on which 90% of the people are used to build the model. Neither should the randomly chosen partition in the cross-validation procedure matter.

It is also possible to calculate the probability of a person belonging to one group. This is simply the number of times that this group is predicted for this person divided by the total number of predictions.

The cross-validation statistics are calculated from the cross-validation errors. Each sample has been in an outer loop cross-validation 10 times. Thus every sample is classified 10 times. The final classification is by majority voting and the probabilities can be calculated by dividing the number of breast cancer classification for the specific sample by 10.

The important m/z values are calculated using a rank product approach; a possible alternative is taking the average of the ranks. For a discriminant vector the 11205 m/z values are ranked according to their absolute values. The largest absolute value gets rank 1 and the lowest absolute value gets rank 11205. This procedure is repeated for the 110 discriminant vectors. Then for all m/z values the products of the 110 ranks are calculated. These are sorted again; the m/z with the lowest rank product is seen as the most important m/z for the discrimination between breast cancer and control.

Results

The results presented here are in the form of confusion tables. These results are called the re-substitution results because for the model development the same samples are used as the samples to be predicted. The confusion table in the case of re-substitution using the majority voting scheme as described above can be seen in Table 1.

		True class	
		Control	Breast cancer
Predicted class	Control	71	8
	Breast cancer	6	68

Table 1. Confusion table of the re-substitution results

Thus the total number of misclassifications is 14, the sensitivity is 89% and the specificity is 92%. The cross-validated confusion table is calculated using the outer loop cross-validation errors and is given in Table 2.

		True class	
		Control	Breast cancer
Predicted class	Control	66	14
	Breast cancer	11	62

Table 2. Confusion table of the double cross-validation results

The total number of misclassifications is 25 with a sensitivity of 82% and a specificity of 86%.

Obviously the sensitivity and specificity of the cross-validated results are smaller than the re-substitution results. Moreover al the samples that are predicted wrongly in the re-substitution are also predicted wrongly in the cross-validation procedure. The sensitivity and specificity of the cross-validated results should be a better estimate for the sensitivity and specificity of the test set results.

Important m/z values.

The 100 m/z values with the smallest rank products are considered. The m/z value with rank product 1 is 4053.9, the second most important m/z is 4054.8 which is just next to the first one. The third one is 4055.6 which is the m/z value neighboring the second most important one. The important m/z values appear in groups. The reason for this is that adjoining m/z values form one peak in the mass spectrum. These peaks represent a protein or a protein fragment.

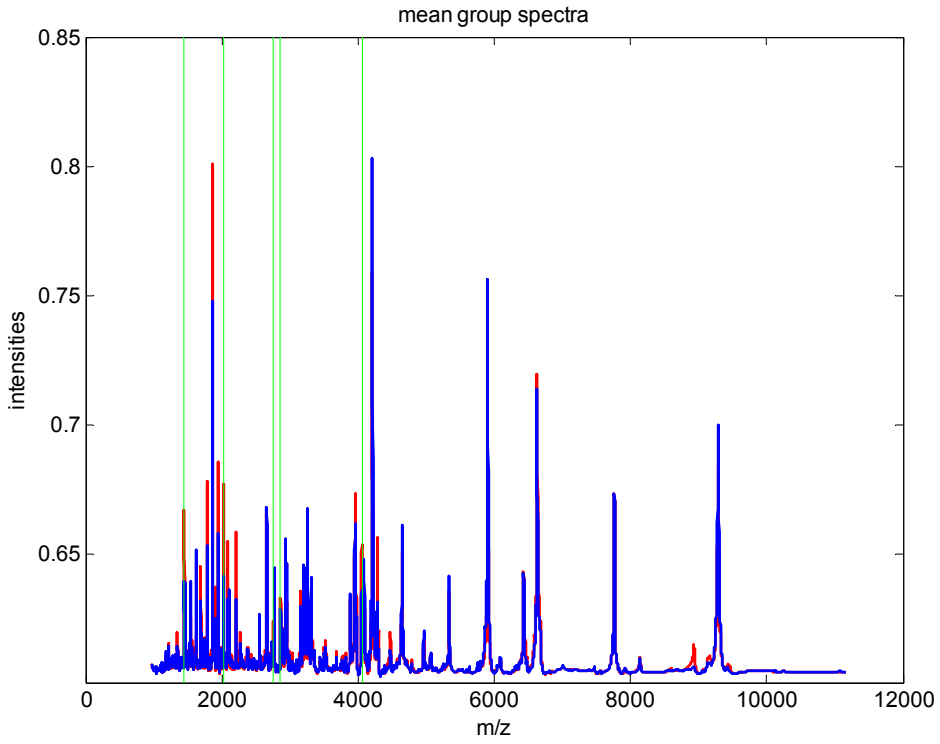

Figure 2. The 5 most important peaks(green vertical lines) in the mean spectra of the controls (blue) and the mean of the patients (red) .

In figure 2 the 5 most important peaks are indicated. It can be seen that the best discriminating peaks in the spectra are not the most abundant peaks.

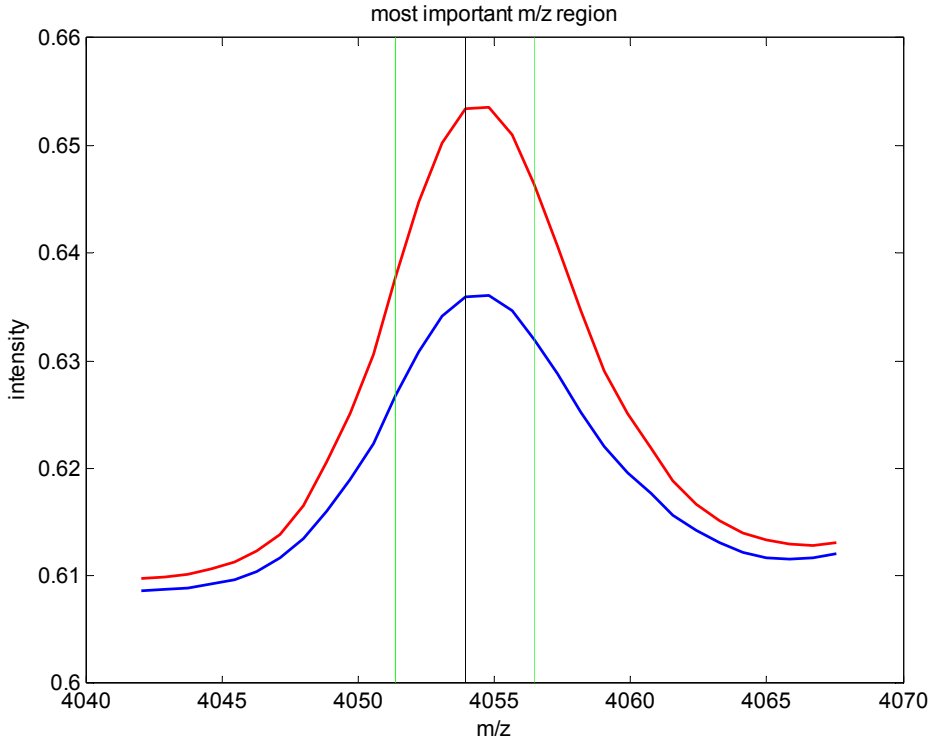

Figure 3. The most important peak for discrimination in the mean spectra of the controls (blue) and the mean of the patients (red).

In figure 3 the most important peak for discrimination can be seen. All 7 m/z values between the green lines are in the top 10 of the rank product. From this peak it can be concluded that on average the protein that is responsible for this peak is elevated in cancer patients.

Stability of the model.

In order to examine the stability of the model the standard deviation and the mean of the 110 discriminant vectors are studied.

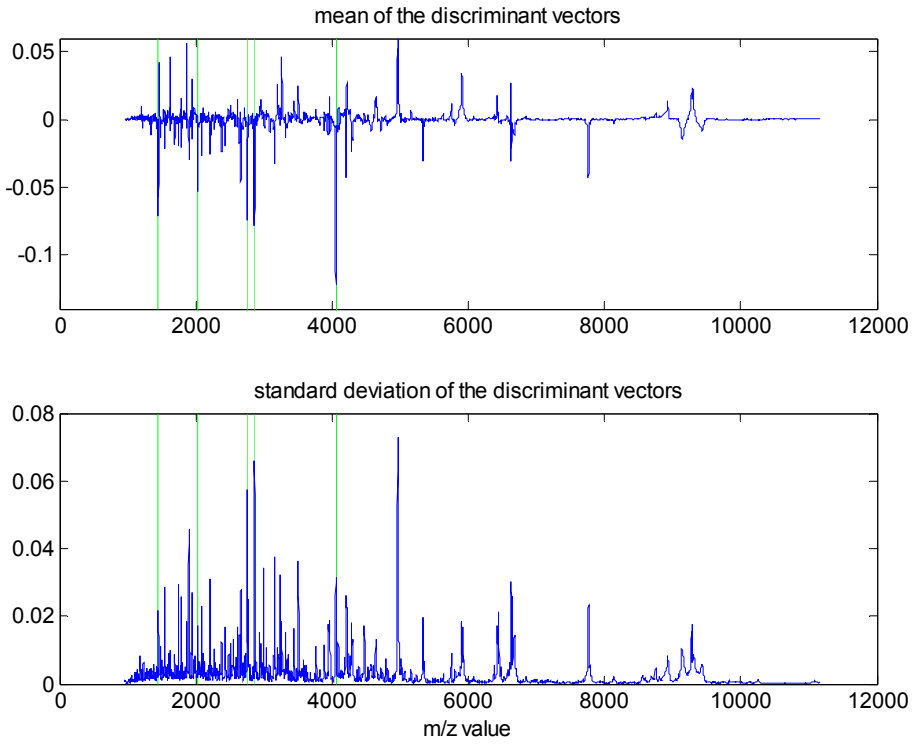

Figure 4. The mean of the discriminant vectors, upper panel and the standard deviation of the 110 discriminant vectors lower panel. The 5 most important peaks are indicated by the vertical green line.

From figure 4 it can be seen that the important m/z values as reported in the previous sub-section coincide with the large values of the mean of the discriminant vectors. Comparing the upper and the lower panel of figure 4 it can be concluded that the standard deviation is relatively large for many m/z values. For a proteomics study the sample size is relatively large but the discriminant vectors are still not very stable for many m/z values.

The situation that a model is not very stable is favorable for the so-called bagging approach (Breiman, 1996). In this approach models are built from bootstraps. It is shown in the Breiman paper that having an aggregated model has beneficial properties if the variation in the obtained models is relatively large. In our study this is the case as is shown above. Although we do not perform a bootstrap but derive our different models from a double cross-validation, the arguments put forward in favor of bagging do in our opinion also hold for our approach.

Conclusions

In this paper a double cross validation approach is put forward in a classification problem. This procedure is followed to obtain an estimate of the specificity and the sensitivity for an unknown sample. In the cross-validation loops the importance of m/z value is evaluated using rank products. Several important regions in the spectrum are found to be important in the classification.

If the test set is measured together with the training set we expect the performance of our classifier to be comparable to the results of the double cross-validation. If the test set is not measured together with the trainings-set the results are probably a little worse than the cross-validated results reported in table 2.

We combine the concept of aggregated models and double cross-validation with a robust and simple method. From the cross validation results it can be concluded that PCDA is a good classifier for the problem at hand.

References

Breiman, L. 1996. Bagging Predictors. Machine learning 24:123-140.

Breitling, R., P. Armengaud, A. Amtmann, and P. Herzyk. 2004. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. Febs Letters 573:83.

Hastie, T., J. Friedman, and R. Tibshiranie. 2001. The Elements of Statistical Learning. Data mining, Inference and Prediction. Springer, New York.

Hoogerbrugge, R., S. J. Willig, and P. G. Kistemaker. 1983. Discriminant Analysis by Double Stage Principal Component Analysis. Analytical chemistry 55:1710.

Howland, P., and H. Park. 2004. Generalizing discriminant analysis using the generalized singular value decomposition. Ieee Transactions on Pattern Analysis and Machine Intelligence 26:995.

Lilien, R. H., H. Farid, and B. R. Donald. 2003. Probabilistic Disease Classification of Expression-Dependent Proteomic Data from Mass Spectrometry of Human Serum. Journal of Computational Biology 10:925.

Mertens, B.J.A., De Noo M.E., Tollenaar R.A.E.M., and D. A.M. 2006. Mass Spectrometry Proteomic Diagnosis: Enacting the Double Cross-Validatory Paradigm. Journal of Computational Biology 13:1591 -1605.

Smit, S., M. van Breemen, H. C. J. Hoefsloot, A. K. Smilde, J. M. F. G. Aerts, and C. G. de Koster. 2007. Assessing the statistical validity of proteomics based biomarkers. Analytica Chimica Acta 592:210-217.

Stone, M. 1974. Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society.Series B (Methodological) 36:111.

Vandeginste, B. G. M., D. L. Massart, L. M. C. Buydens, S. De Jong, P. J. Lewi, and J. Smeyers-Verbeke. 1998. Handbook of Chemometrics and Qualimetrics: Part B. Elsevier, Amsterdam.

Ye, J., T. Li, T. Xiong, and R. Janardan. 2004. Using Uncorrelated Discriminant Analysis for Tissue Classification with Gene Expression Data. IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS 1:181.