Metabolomics (2008) 4:81–89
DOI 10.1007/s11306-007-0099-6

ORIGINAL ARTICLE

Assessment of PLSDA cross validation

Johan A. Westerhuis · Huub C. J. Hoefsloot · Suzanne Smit · Daniel J. Vis · Age K. Smilde · Ewoud J. J. van Velzen · John P. M. van Duijnhoven · Ferdi A. van Dorsten

Received: 9 July 2007 / Accepted: 23 October 2007 / Published online: 24 January 2008

Abstract Classifying groups of individuals based on their metabolic profile is one of the main topics in metabolomics research. Due to the low number of individuals compared to the large number of variables, this is not an easy task. PLSDA is one of the data analysis methods used for the classification. Unfortunately this method eagerly overfits the data and rigorous validation is necessary. The validation however is far from straightforward. Is this paper we will discuss a strategy based on cross model validation and permutation testing to validate the classification models. It is also shown that too optimistic results are obtained when the validation is not done properly. Furthermore, we advocate against the use of PLSDA score plots for inference of class differences.

Keywords Cross model validation · Permutation testing · Classification · PLSDA

1 Introduction

The research area of metabolomics is growing fast due to an enormous improvement of analytical technology as LCMS, GCMS and NMR (Bollard et al. 2005; Van Der Greef and Smilde 2005). The application field is rather wide, ranging from plants (Bino et al. 2004; Fiehn 2002) to microbial (van der Werf et al. 2005), medical (Clayton et al. 2006) and even nutritional applications (Van Dorsten et al. 2006; van Ommen 2004). The typical metabolomics study involves two groups of individuals, often called case and control (Broadhurst and Kell 2006). Such a study can be used in an exploratory way or in a predictive way. An *explorative* study is used to see whether the specific data contains sufficient information to distinguish between the two groups. E.g. recently the use of MALDI to detect metabolites has been explored, but it was unknown whether these data contained sufficient information to make a distinction between diseased and control groups (Ragazzi et al. 2006; Vaidyanathan and Goodacre 2007). Then there is need for a *predictive* model that can predict whether an unseen individual belongs to the case or control group.

An often used data analysis tool for classification in the metabolomics area is PLSDA (Barker and Rayens 2003) or OPLSDA (Bylesjo et al. 2006; Trygg 2002; Trygg and Wold 2002). These classification tools are based on the PLS model in which the dependent variable is chosen to represent the class membership. The large number of peaks in these spectra that are all potential biomarkers create modelling and validation challenges. The number of samples needed to accurately describe such a classification problem increases exponentially with the number of variables measured. However, the number of samples used in these applications is usually much smaller than the number of variables. This can easily lead to chance classifications, i.e. models that just by chance give a good classification of the two groups.

A good start for the analysis of any data analysis method is to use a set of random data and see how the method deals with it. A convincing example that validation of PLSDA models is of major importance is the classification of a set

J. A. Westerhuis (✉) · H. C. J. Hoefsloot · S. Smit · D. J. Vis · A. K. Smilde · E. J. J. van Velzen
Biosystems Data Analysis, Swammerdam Institute for Life Sciences, Universiteit van Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherland
e-mail: j.a.westerhuis@uva.nl

E. J. J. van Velzen · J. P. M. van Duijnhoven · F. A. van Dorsten
Unilever Food and Health Research Institute, Olivier van Noortlaan 120, 3133 AT Vlaardingen, The Netherlands

of random data. Using PLSDA to discriminate a random data set of size e.g. 40 × 100 (comparable to the size of metabolomics data) into two groups does almost always give a PLS score plot with perfect separation between the two arbitrary classes. Please try this using your own software.

A similar example is given in Fig. 1. Here NMR spectra (382 channels) of a group of 23 healthy volunteers are arbitrarily divided into two classes of 11 and 12 class members. Cross validation revealed a Q^2 value of −0.18 which is usually considered not to be a good classification model. However, the PLSDA score plot in Fig. 1 shows a clear separation between the two classes. PLSDA is eager to please and thus its results should be handled with great care. The problem is that in the 382 dimensional megavariate space there is almost always a perfect separation possible between the 23 samples and PLSDA has no problems finding it. The PLSDA score plot therefore does not give a good representation of class difference between the groups. This is probably the reason why others already started to show cross validated score plots (Cloarec et al. 2005). However, the problem with that plot is that each score value is based on a different loading and therefore they should not be plotted into one figure as they cannot directly be compared.

Although the score plot should not be used to infer class separation, it might reveal structure (e.g. subgroups) within a class. Since the model is not forced to show this difference, this is not a result of overfit, and thus such information could be inferred from the score plot.

Validation of PLSDA models has received a lot of attention. Very recently a number of papers appeared discussing various aspects of the validation and claiming that in many applications, proper validation of the classification

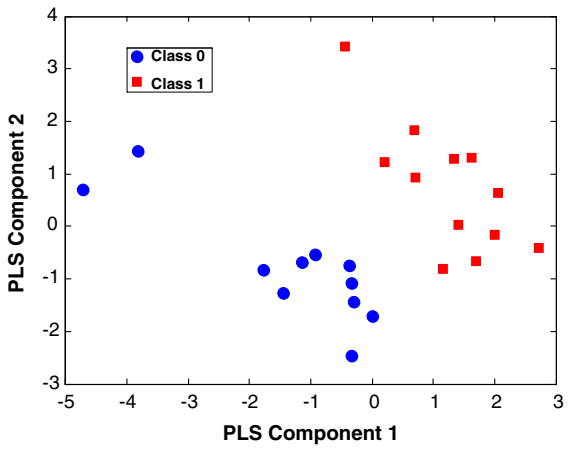

Fig. 1 PLSDA score plot of an NMR data set of healthy volunteers which were arbitrarily divided into two classes. Q^2 value of this model was −0.18. Still a clear separation between the classes is observed in this score plot

models were lacking (Brereton 2006; Broadhurst and Kell 2006; Harrington 2006; Rubingh et al. 2006). Broadhurst and Kell summarize the main problems in the analysis of megavariate data. The most important one might be the too small sample size. Due to the very expensive experiments, the number of samples is usually too small. A good criterion of sample size is hard to give as multivariate power calculations are not well understood at the moment. A good starting point here is the use of Monte Carlo simulations to determine the appropriate number of samples (Martens et al. 2000). Most important however is that the samples chosen contain sufficient information to provide the answers searched for (Trygg et al. 2007).

Cross validation is performed in most cases to validate the results found, but often not performed in a proper manner (Anderssen et al. 2006). Various parameters have been used to quantify a certain classification, e.g. Q^2 values, number of misclassifications, many combinations of sums or ratios of True Positives, True Negatives and False positives and False Negatives of a confusion matrix. Also the area under curve of a receiver operating characteristic (ROC) curve (AUROC) is used often. Problem with all these measures is that it is unknown which value corresponds to a good discrimination between the groups. A measure for statistical significance (e.g. a *P*-value) is usually not given. Furthermore, this value depends on the number of samples in the training and test set. Another problem with the cross validation procedure is that in each of the many models built during the cross validation, a different number of PLS components seem important for each of the submodels. At the moment there are no accepted criteria for the way to choose the overall model, when returning to the full dataset based on the conclusion obtained from the many models developed from subsets of the samples.

In this paper we will tackle most of the problems discussed above. The main message we will bring in this paper is that by analysing many versions of the data with randomly assigned class labels, a reference distribution for the H0 hypothesis that no difference exist between the two classes is obtained. Using these permutations we will show that improper use of cross validation leads to a too optimistic classification result. Although many papers have pointed to this problem, we clearly show that too few misclassifications are obtained when cross validation is used wrongly. Furthermore using the permutations each quality parameter used to assess the quality of the classification (e.g. Q^2, AUROC or the number of misclassifications) is accompanied with its own H0 distribution of values that can be obtained in case of no difference between the classes. From this it can be observed e.g. which Q^2 value corresponds to a statistically significant classification model. We argue against the use of a single final model, but instead promote

the use of many slightly different models to obtain a range of class membership predictions. This range can be used as a confidence measure for class membership assignment. Note that although we use PLSDA here as the test case, the same approach can be used for other classification methods.

2 Theory

2.1 PLS-DA

Partial least squares discriminant analysis (PLSDA) is a frequently used classification method and is based on the PLS approach (Barker and Rayens 2003). The standard PLS algorithm can be used and for the dependent **y** vector, class labels can be used. In the two-class case, usually the values of the dependent variable are given 1 for one class and 0 or −1 for the other class. In case of more than 2 classes, dummy variables are defined and a PLS2 algorithm is used. A much used variant of PLSDA is OPLSDA where the first components orthogonal to the dependent variable are removed from the data (Trygg and Wold 2002). This gives a model with a single classification component while the other components describe the other variation that is orthogonal to the class information. OPLS enhances the interpretation of PLS by forcing all classification information into a single component. The prediction power of both models is usually the same (Trygg and Wold 2002).

Typical applications of two class classifications using PLSDA show a score plot of the classified samples. This plot is often accompanied with a Q^2 value indicating the validity of the discrimination. Figure 2 shows such a typical score plot of a PLSDA classification between men and women based on NMR spectra of their urine. The data presented here is described in more detail in the experimental section.

As demonstrated in the introduction of this paper, a score plot indicating separation between two groups has no meaning as similar plots can be obtained when random data is classified. It is also unknown what the corresponding Q^2 value of 0.51 indicates. Is this a good separation? As long as there are no values to compare the Q^2 value with, the value of this number is meaningless.

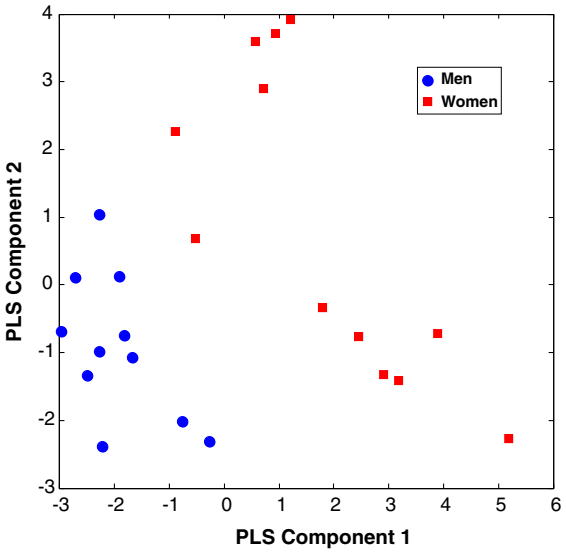

Fig. 2 Typical PLSDA application showing a score plot of the two groups of samples nicely separated. In this case Q^2 was found to be 0.51

2.2 Permutation tests

A permutation test can evaluate whether the specific classification of the individuals in the two designed groups is significantly better than any other random classification in two arbitrary groups (Golland et al. 2005; Mielke Jr and Berry 2001). In a permutation test, the class labels of case and control are permuted, they are randomly assigned to different individuals. With the 'wrong' class labels, again a classification model is calculated. The rationale behind the permutation test is that with the wrong class labels, the newly calculated classification model should not be able to predict the classes very well. As the groups are formed in a random way, the assumption is that no difference exists between them. By repeating the permutation test many times, a H0 distribution of classifications that are expected not to be significant is formed. From these classifications, H0 distributions for Q^2, AUROC and for regression coefficients etc can be obtained. The results obtained from the non permuted set of samples should be outside the 95 or 99% confidence bounds of the H0 distribution from the permuted classifications to be significant.

A major advantage of the permutation test is that the H0 reference distribution is always based on models based on the same number of samples that also show the same amount of variation, outliers, missing data etc. The reference distribution therefore perfectly matches the analysis results of the original model. The permutation test is integrated in the currently described validation procedure. It will be used to validate the metabolic differences. It will also be used to evaluate validation procedures such as cross validation.

2.3 Cross validation

Cross validation is often used for validation of a classification model due to the low number of samples available.

As separation into training, validation and test set is often not possible, cross validation makes better use of the data. However, cross validation only gives a reliable error rate when the complete modelling procedure is cross validated. The object that is predicted should in no way be used in the development of the model (Anderssen et al. 2006; Brereton 2006; Broadhurst and Kell 2006). This is repeatedly mentioned in many papers, however, it is still not common practise in the many applications found in the recent literature.

For a proper cross validation, the total data should be divided into a training set, a validation (sometimes called optimization) set and a test set. Using the validation and training set a model is developed and optimized. The test set is only used to test the model performance. By repeating the procedure in a way that each sample appears once and only once in the test set, the prediction error is representative for new samples. For a complete independent test set, it should also not be used in data pre-treatment and pre-processing, scaling etc.

In this paper we will use three versions of cross validation that will be compared using the permutation strategy.

FIT: Using the single cross validation strategy (1CV, see below), the optimal number of PLS components is found. A PLSDA model is then built on all samples using this optimal number of components. Then the class labels of all samples are "predicted" using this model. Note that here predicted is placed inside quotation marks to indicate that this is not a real prediction, but merely a resubstitution. Based on the predictions of the class labels in this way, Q2_FIT, AUROC and number of misclassifications can be obtained.

1CV: Single cross validation is generally used in many applications. With single cross validation, some of the individuals are removed from the data and used as a validation set. The remainder of the individuals that form the training set are used to develop a series of classification models with 1 to many PLS components. For each of these models, a prediction of the validation set is given and the prediction errors of all of these models are stored. This procedure is repeated for a new set of individuals until all individuals have been in the validation set once (and only once). The total prediction error for all the models over all individuals is calculated, and the model with the lowest total prediction error is selected as the best. The prediction errors obtained using the 1CV approach can be used to calculate the number of misclassifications and the Q2_1CV. It has to be realized that the same individuals (in the validation set) were also used to find the best overall model parameter and thus they are not completely independent as is requested for a proper cross validation.

2CV: In order to overcome the dependency between the prediction error for new individuals and the optimization of the model parameter, cross model validation (2CV) was suggested. In 2CV one set of individuals is set aside completely as the test set. Then the remainder of the individuals are subjected to a single cross validation regime. In this regime the remaining individuals are again split into a validation and a training set. The single cross validation will result in an optimal number of PLS components. Then all training and validation individuals are used to build a final classification model with the optimal number of PLS components. This final model is then used to predict the individuals in the test set. The whole procedure is repeated until all individuals have been in de test set once (and only once). It is important that the selection of the validation samples is done randomly to enforce different combinations of validation sets and training set for each new test set. In this way the model has been built in absolute absence of the test set, the prediction is independent of the model optimization (Stone 1974).

Summarizing, the 1CV method assesses the variability of the estimated parameters and its effect on the prediction, while the 2CV method also assesses the variability of the meta parameters and their effect on the prediction. Table 1 gives a summary of the dependence of the sample that is predicted and some of the model parameters. Only when 2CV is used for validation, the sample that is predicted is truly independent, leading to a correct prediction error.

2.4 Final calibration model

Note that now the prediction of each separate test set is performed on slightly different models that have different optimal model parameters. This complicates the development of a final calibration model, because it is not known what model parameters to choose for the final model (Brereton 2006). The precision of a final calibration model is not allowed to be smaller than the models that were developed during the cross validation procedure. However those models were always based on a subset of the samples.

Table 1 Dependence of predicted sample to model parameters for the FIT, 1CV and 2CV approaches

Method	Predicted sample used to calculate model coefficients	Predicted sample used to calculate number of components
FIT	Yes	Yes
1CV	No	Yes
2CV	No	No

Therefore the full potential of the final calibration model is never used.

However, such a final model is also not necessary. Instead of having one final model, a group of models (one for every test set) is available to classify future individuals. Instead of one class prediction, a group of class predictions is given for each individual. This group of predictions can be combined into an average prediction and it can also be used to obtain a confidence interval for the class prediction. This is related to the bagging approach introduced by Breiman (Breiman 1996, 1998). Furthermore, the various models give insight in the stability of the optimized model parameters.

2.5 Quality assessment

Many measures exist to quantify the quality of a specific classification such as measures, derived from the confusion table, which consist of the number of False Positives, False Negatives, True Positives and True Negatives exist (Broadhurst and Kell 2006). For the Case individuals, the fraction of True positives is referred to as the *sensitivity* while the fraction of False Positives is referred to as (1-*specificity*). Combining the two leads to the receiver operator characteristic (ROC) that we will use in this paper. The ROC unifies two characteristics that are often used to evaluate the performance of a (clinical) test or method. A ROC curve plots the sensitivity of a test versus 1-specificity of a test. The sensitivity is defined as the number of true positives found as a percentage of all positives (diseased people), while 1-specificity is the number of false positives as a percentage of all negatives (controls, healthy people). Sensitivities are between 0 and 1 and should be close to 1. The specificity should preferably be close to 1, and 1-specificity should be close to 0. Both specificity and sensitivity depend on the setting of the classification boundary of the classifier used by a method. By shifting the classification boundary more true positives may be detected, but the number of false positives also increases, and the other way around. The ROC curve therefore is a characteristic of a method, describing the sensitivity and specificity of a method for different classification boundaries. Each point on a line gives the sensitivity of a classifying model versus 1-specificity of the model. Each point on the ROC curve refers to a value chosen for the classifier boundary. Instead of choosing 0 as the classified boundary, one can select a slightly lower value to increase the sensitivity of class +1. Of course this goes together with a loss of 1-specificity. As a quality measure we use the area under the ROC curve (AUROC). This value goes to 1 for a perfect separation between the classes. If there is no separation then a value close to 0.5 is obtained.

The prediction error measure Q^2, which is the default parameter used in PLSDA discriminations focuses on how well the class label can be predicted from new data. Q^2 is defined as follows:

$$Q^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

where $\hat{y}_i$ refers to the predicted value of class membership for sample i while $\bar{y}$ refers to the mean value of y for all samples. The optimal Q^2 value of 1 is difficult to reach as this requires that the class prediction of each individual should be exactly equal to its class label. This is hard to get due to the inherent variation between the individuals in the same class. The Q^2 depends on the between class separation but also on the within class variability. This makes it difficult to give a general Q^2 value that corresponds to a good classification. Therefore we use permutations to provide a whole distribution of Q^2 values for models of no effect to relate the original Q^2 value to.

The ROC and number of misclassifications are both classification error measures (they only make a distinction between good and wrongly classified). The Q^2 value is a prediction error that makes a distinction between slightly wrong and very wrong. A prediction of -0.2 is penalized more than a prediction of 0.4 for a class label of 1 while the classification error measures treat these predictions both equally as wrong. In a further study we will examine the power of these measures.

3 Experimental

3.1 Human urine samples

Twenty-four-hour urine samples were collected from 23 healthy male and female volunteers in the age range between 19 and 78 years, i.e. 12 women (mean age: 42.0 years, range: 20–78 years) and 11 men (mean age: 44.4 years, range: 19–74 years). All samples were stored frozen at -20°C prior to analysis and thawed on the day of analysis.

3.2 ^{1}H NMR Spectroscopy

^{1}H NMR spectra were acquired at 600.13 MHz on a Bruker Avance 600 spectrometer. Urine samples were prepared for ^{1}H NMR spectroscopy by diluting 300 µl of urine 1:1 with ^2H_2O, followed by spinning down non-soluble particles for 10 min at 14,000 rpm in an Eppendorf centrifuge. Urine NMR spectra were measured at 303 K using a standard

water-suppressed 1D-NOESY pulse sequence, i.e. RD-90°-t_1-90°-t_m-90°-acquire. Here, RD is a relaxation delay of 1.5 s during which the water resonance was selectively irradiated, and t_1 corresponds to a fixed delay of 0.15 s. A total of 128 transients were collected into 32 k data points, with a spectral width of 7,000 Hz. Prior to Fourier transform (FT) the free induction decays (FIDs) were multiplied by a 0.3-Hz exponential line-broadening function and zero-filled by a factor of 2.All spectra were manually phase- and baseline-corrected using XWINNMR (*Bruker GmbH, Germany*), and referenced internally to the creatinine methyl peak at δ 3.10.

3.3 NMR spectral data reduction

The NMR spectral dataset between 0.4 and 10.0 ppm was automatically reduced into regions of equal width (0.02 ppm), to minimize the effects of (pH-dependent) peak shifts, and the integral of each region was determined using AMIX software (Bruker GmbH, Germany). The spectral region from 4.0–6.0 ppm was excluded from analysis to remove the effect of variations in the suppression of the water resonance and variations in the urea signal. The peak integral within each 0.02-ppm spectral region was normalized to a constant sum of 1,000 for each spectrum, to account for differences in urinary volume.

3.4 Data analysis

In the single and double cross validation procedures, the samples were split into six groups; a 6-fold cross validation strategy is used. It was enforced that in each validation set (1CV) or test set (2CV) individuals of both classes were present. Thus six different models are created. The single cross validation inner loop of the double cross validation, which is used to obtain the optimal number of PLS components, used a leave one out CV1. The selection of the test sets, validation sets and training sets were randomized to enforce a different combination of samples. The class labels of 0 and 1 were used for men and women respectively.

4 Results and discussion

4.1 Evaluation of cross validation

In the first part of the results and discussion section we evaluate the different versions of cross validation discussed in the theory section. The number of misclassification is used here as a measure because of its known expected value in the fully randomized case. Here we mean that if the individuals are randomly permuted over the classes that the average number of misclassifications should be half of the samples in a two-class problem. Predictions of class labels using the actual data were performed 20 times in which the combination of samples in training set, validation set and test set was varied. The numbers of misclassifications were averaged over the 20 cross validation runs. The data was permuted 2,000 times. Of each permuted set, the cross validation was repeated 20 times and the average number of misclassifications was calculated.

Using the permutations, a distribution is developed for the H0 hypothesis being that no difference exists between the two classes. When no difference exists between the classes, an average of 11.5 misclassifications is expected. Thus the validation procedure should on average give 11.5 misclassifications for the permuted datasets. In the evaluation of cross validation procedures we calculate the number of misclassifications using the FIT, 1CV and 2CV approach.

Figure 3 shows histograms of the number of misclassifications of the permuted data. When the FIT approach is used, many good classifications are made on the permuted data. The average number of misclassifications using the FIT approach is 1.5, which is clearly much too optimistic as the expected number of misclassifications of the permuted data is 11.5. Note that even in the FIT approach a single cross validation step was used to select the number of PLS components. But this is in no way sufficient for a proper validation.

The single cross validation (1CV) procedure leads to an average number of misclassifications of 10.0 for the permuted data. This is still a too optimistic result. In single cross validation the predicted individuals are still not

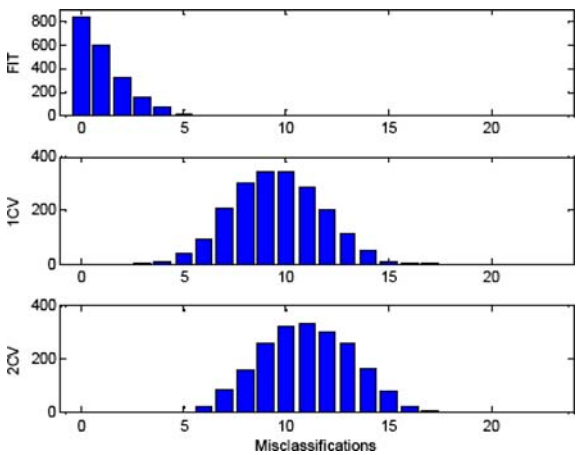

Fig. 3 Misclassifications of gender PLSDA as classification method with FIT, 1CV and 2CV as validation strategies with permutated data

representative of a completely new individual as they are also used to define the optimal number of PLSDA components. Only when a double cross validation (2CV) strategy is used, the number of misclassifications equals the expected value of 11.5.

4.2 Quality assessment

Is there a difference between the groups? How to statistically quantify whether a difference exists? If the number of misclassifications is lower than half of the number of samples, then there is a difference. If the area under the ROC curve is above the 0.5 then there is also a difference. The default parameter in PLSDA, the Q^2 value, does not provide an answer to this question. It is unknown which value Q^2 value to expect for a good classification of two groups. Therefore we will again use the permutation approach to obtain a distribution of Q^2 values that correspond to the H0 of no difference between the two classes. In this case the Q^2 values and the AUROC and also the number of misclassifications are obtained using predictions of class labels in a cross model validation setup. Thus the predictions of class labels are performed for samples that have not been used to develop the model. Figure 4 shows the number of misclassifications, the Q^2 values and the AUROC values for the original classifications in red and for the permutations in blue based on predictions in a cross model validation framework. For all assessment parameters there is a clear distinction between the permutation distribution and the original classification. This shows that the specific classification is significant. If we compare the average value of the original classification with all the permutations then a *P*-value can be obtained.

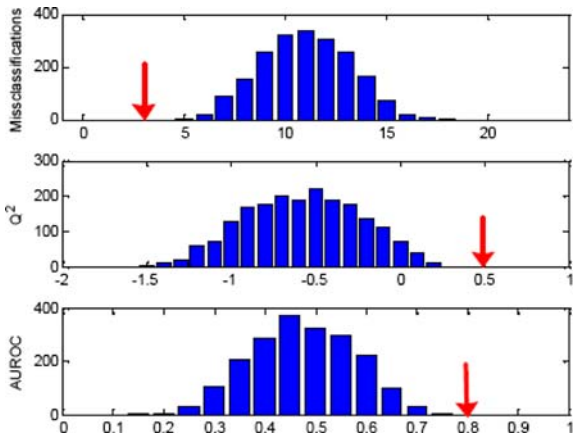

Fig. 4 Class prediction results based on cross model validation predictions of the original labeling compared to the permuted data assessed using the number of misclassifications, Q^2 and area under the ROC curve

Using double cross validation to predict the class label, the original classification has on average 3.35 misclassifications. Of the 2,000 permutations none had a number of misclassifications lower than 3.35 leading to a $P < 0.0005$. The Q^2 values of the permutations show a distribution around -0.5. Six percent of the permutations show a Q^2 value above 0. The original labelling had a Q^2 value of 0.5. This also leads to a $P < 0.0005$ as none of the permutations had an average Q^2 value above 0.5. The average AUROC value is indeed 0.5 as expected while the average AUROC of the original labelling was 0.80 ($P < 0.0005$). For the set of 2,000 permutations no difference between the three assessment criteria were found.

4.3 Variation in class prediction using PLSDA

There is a major problem to translate the optimization results obtained during the cross validation to a final model. If a successful predictive model is obtained, it is then necessary to choose a model that will be applicable to new samples using all samples available. However, the assessment of the model has been performed on multiple subsets of samples, each with its own number of components, variables selected, scaling etc. At the moment no consensus exist how to choose the overall model based on submodel results (Brereton 2006). Therefore instead of having one final model with a single class prediction for new samples, multiple class predictions can be obtained from many different models that were developed during the cross validation procedure. Instead of having a single prediction of gender, it is interesting to know the variability in class prediction when many related models are used. Furthermore, often the average of multiple models is a better estimate of class, then a single prediction. This idea goes back to bagging predictors (Breiman 1996, 1998). The average of a set of predictors can have a smaller variance than a single predictor. In cross model validation each of the multiple models is developed using a smaller sample size and therefore will probably have a somewhat higher prediction error. However, because of the multiple related models, a good idea is obtained of the spread in class prediction. This spread can be used as a confidence measure for the class prediction, while the average of all prediction will probably have a lower variance. Therefore 1,000 models were developed within the cross model validation, each time with a different combination of samples. The prediction of the class labels is performed using models that are unrelated to the sample of which the class is predicted. The class predictions are presented in Fig. 5.

Figure 5 shows the spread in **y**-value prediction and thus the classification. Samples 1–11 are males and were coded 0. Samples 12–23 are females and were coded as 1. It can

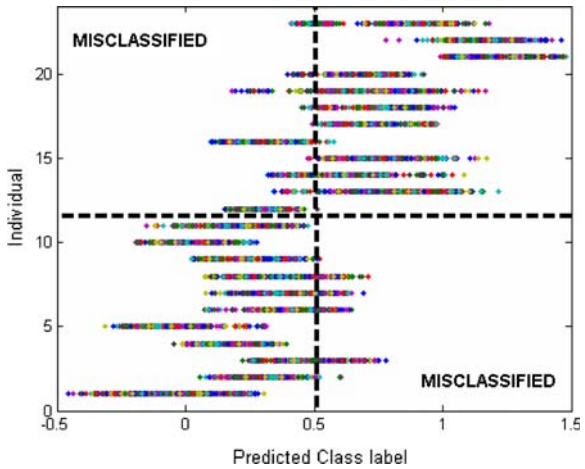

Fig. 5 Predicted **y**-values for sex using double cross validation. Individuals 1–11 were coded 0 and individuals 12–23 were coded 1. Individuals in the upper left and lower right corner are misclassified

be seen that some males are always classified correctly (1, 4, 5, 10 and 11) and also some females are always classified correctly (21 and 22). Individual 12 is almost always classified wrongly, while others are in almost half of the cases classified wrongly. The predicted range is indicative of the confidence in class membership.

5 Conclusion

Classification problems in metabolomics data analysis are complex due to the many variables few samples issue. This makes that many solutions can be found to separate the classes. Most models therefore suffer from overfit, meaning that the model classifies the training data well, but future samples are classified poorly.

In this paper we describe the use of permutation testing and cross model validation to assess the validation of classification models. Permutation tests show that when cross validation is not applied appropriately, it leads to overoptimistic results. Moreover, the permutation test gives a sensible measure for the Q^2 value. Cross model validation leads to good results. Although not new and shown in other fields, it is unfortunately almost completely ignored in the metabolomics research field.

The PLSDA score plots as presented in most classification applications in the metabolomics research field present an overoptimistic view of the separation between the classes. Although within class variation can be revealed by this plot, it should not be used to examine between class separation. This was clearly shown by the classification of randomly assigned class membership. Class separation results should be based on predictions instead of on fitted values (such as the PLSDA scores). It is therefore much more informative to present class predictions as was shown here in this paper.

Acknowledgement We acknowledge the financial support of the Community under the Marie Curie Host Fellowships for the Transfer of Knowledge—Industry-Academia Strategic Partnership Scheme, specifically MTKI-CT-2006-042786, GUTSYSTEM.

References

Anderssen, E., Dyrstad, K., Westad, F., & Martens, H. (2006). Reducing over-optimism in variable selection by cross-model validation. *Chemometrics and Intelligent Laboratory Systems, 84*(1–2), 69–74.

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics, 17*(3), 166–173.

Bino, R. J., Hall, R. D., Fiehn, O., et al. (2004). Potential of metabolomics as a functional genomics tool. *Trends in Plant Science, 9*(9), 418–425.

Bollard, M. E., Stanley, E. G., Lindon, J. C., Nicholson, J. K., & Holmes, E. (2005). NMR-based metabonomic approaches for evaluating physiological influences on biofluid composition. *NMR in Biomedicine, 18*(3), 143–162.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.

Breiman, L. (1998). Arcing classifiers. *Annals of Statistics, 26*(3), 801–824.

Brereton, R. G. (2006). Consequences of sample size, variable selection, and model validation and optimisation, for predicting classification ability from analytical data. *Trac-Trends in Analytical Chemistry, 25*(11), 1103–1111.

Broadhurst, D. I., & Kell, D. B. (2006). Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics, 2*(4), 171–196.

Bylesjo, M., Rantalainen, M., Cloarec, O., et al. (2006). OPLS discriminant analysis: Combining the strengths of PLS-DA and SIMCA classification. *Journal of Chemometrics, 20*(8–10), 341–351.

Clayton, T. A., Lindon, J. C., Cloarec, O., et al. (2006). Pharmaco-metabonomic phenotyping and personalized drug treatment. *Nature, 440*(7087), 1073–1077.

Cloarec, O., Dumas, M. E., Craig, A., et al. (2005). Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic H-1 NMR data sets. *Analytical Chemistry, 77*(5), 1282–1289.

Fiehn, O. (2002). Metabolomics - the link between genotypes and phenotypes. *Plant Molecular Biology, 48*(1–2), 155–171.

Golland, P., Liang, F., Mukherjee, S., & Panchenko, D. (2005). Permutation tests for classification. *Lecture notes in Computer Science, 3559*, 501–515.

Harrington, P. D. B. (2006). Statistical validation of classification and calibration models using bootstrapped Latin partitions. *Trac-Trends in Analytical Chemistry, 25*(11), 1112–1124.

Martens, H., Dijksterhuis, G. B., & Byrne, D. V. (2000). Power of experimental designs, estimated by Monte Carlo simulation. *Journal of Chemometrics, 14*(5–6), 441–462.

Mielke, P. W. Jr, & Berry, H. (2001). *Permutation methods: A distance function approach*. New York: Springer.

Ragazzi, E., Pucciarelli, S., Seraglia, R., et al. (2006). Multivariate analysis approach to the plasma protein profile of patients with advanced colorectal cancer. *Journal of Mass Spectrometry, 41*(12), 1546–1553.

Rubingh, C. M., Bijlsma, S., Derks, E. P. P. A., et al. (2006). Assessing the performance of statistical validation tools for megavariate metabolomics data. *Metabolomics, 2*(2), 53–61.

Stone, M. (1974). Cross validatory choice and assessement of statistical predictions. *Journal of Royal Statistical Society B, 36*, 111–147.

Trygg, J. (2002). O2-PLS for qualitative and quantitative analysis in multivariate calibration. *Journal of Chemometrics, 16*(6), 283–293.

Trygg, J., Holmes, E., & Lundstedt, T. (2007). Chemometrics in metabonomics. *Journal of Proteome Research, 6*(2), 469–479.

Trygg, J., & Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics, 16*(3), 119–128.

Vaidyanathan, S., & Goodacre, R. (2007). Quantitative detection of metabolites using matrix-assisted laser desorption/ionization mass spectrometry with 9-aminoacridine as the matrix. *Rapid Communications in Mass Spectrometry, 21*(13), 2072–2078.

Van Der Greef, J., & Smilde, A. K. (2005). Symbiosis of chemometrics and metabolomics: Past, present, and future. *Journal of Chemometrics, 19*(5–7), 376–386.

van der Werf, M. J., Jellema, R. H., & Hankemeier, T. (2005). Microbial metabolomics: Replacing trial-and-error by the unbiased selection and ranking of targets. *Journal of Industrial Microbiology & Biotechnology, 32*(6), 234–252.

Van Dorsten, F. A., Daykin, C. A., Mulder, T. P. J., & Van Duynhoven, J. P. M. (2006). Metabonomics approach to determine metabolic differences between green tea and black tea consumption. *Journal of Agricultural and Food Chemistry, 54*(18), 6929–6938.

van Ommen, B. (2004). Nutrigenomics: Exploiting systems biology in the nutrition and health arenas. *Nutrition, 20*(1), 4–8.