Consequences of sample size, variable selection, and model validation and optimisation, for predicting classification ability from analytical data

Richard G. Brereton

This article discusses problems of validating classification models especially in datasets where sample sizes are small and the number of variables is large. It describes the use of percentage correctly classified (%CC) as an indicator for success of a classification model. For small datasets, %CC should not be used uncritically and its interpretation depends on sample size. It illustrates the use of a common classification method, discriminant partial least squares (D-PLS) on a randomly generated dataset of 200 samples and 200 variables.

An aim of the classifier is to determine whether the null hypothesis (there is no distinction between two classes) can be rejected. Autoprediction gives an 84.5% CC. It is shown that, if there is variable selection, it must be performed independently on the training set to obtain a CC close to 50% on the test set; otherwise, over-optimistic and false conclusions can be reached about the ability to classify samples into groups.

Finally, two aims of determining the quality of a model are frequently confused, namely optimisation (often used to determine the most appropriate number of components in a model) and independent validation; to overcome this, the data should be split into three groups.

There are often difficulties with model building if validation and optimisation have been done on different groups of samples, especially using iterative methods, each group being modelled using properties, such as a different number of components or different variables.

Keywords: Classification model; Discriminant partial least squares; D-PLS; Optimisation; Percentage correctly classified; Validation; %CC

Richard G. Brereton
Centre for Chemometrics, School of Chemistry, University of Bristol, Cantocks Close, Bristol BS2 8DF, UK

E-mail: r.g.brereton@bris.ac.uk

1. Introduction

There is increasing use of pattern recognition in chemistry, especially in applications to biology and medicine where analytical chemical data are employed to make predictions about the origins of samples [1–3]. Many applications involve obtaining a large number of variables (e.g., chromatographic peak areas, mass spectral intensities or nuclear magnetic resonance (NMR) peak heights) and trying to use these for classification of samples (e.g., using measurements on extracts from plasma to predict disease state, or plant extracts to predict future productivity).

An example involved the use of gas chromatography combined with mass spectrometry (GC-MS) to determine the sexes of subjects from their sweat. From a population in Carinthia (Southern Austria), 910 samples were obtained, extracted using stir bars, and their GC-MS results recorded. The aim was to use the relative intensities and the presence or the absence of 337 peaks detected in the chromatograms from human emanations for predictive modelling [4]. In reality, this would allow the sex of an unknown individual to be determined from human emanations and the principles could be extended to predict other aspects of individual identity and personal habits.

With the advent of proteomics, there is an even more severe problem with variable-rich datasets often recorded on very few samples; typically, tens of thousands of variables can be obtained from 30 or 40 samples.

A common theme is that a model is formed between two blocks of data, as illustrated in Fig. 1. The model has two main purposes. The first is *exploratory.*

0165-9936/$ - see front matter doi:10.1016/j.trac.2006.10.005

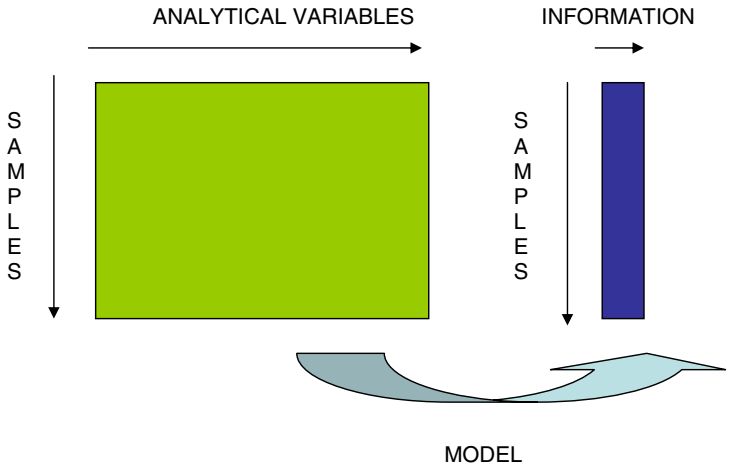

Figure 1. Modelling analytical information, relating information to a matrix of observations.

Such a model may be used to answer questions as to whether there is a genuine connection between the two sets of data (e.g., whether we can use liquid chromatography combined with mass spectrometry (LC-MS) to predict whether a patient has a disease), and also what variables are best for discriminating between classes (e.g., marker compounds). The second is *predictive*. The aim of such a model is to determine whether the origin of a sample of unknown nature can be predicted using analytical data, and, if so, how well.

In this article, we will discuss primarily class models, and, for simplicity, we will focus on a two-class problem, where samples originate from one of two possible sources, class A or class B. There are numerous approaches ranging from Mahalanobis distance [5], soft independent modelling of class analogy (SIMCA) [6] and discriminant partial least squares (D-PLS) [7] to support vector machines (SVM) [8]. There are several common challenges. One of the most serious relates to the ability of modern analytical methods to produce a large number of variables (e.g., in chromatography, typically several hundred peaks can be detected per chromatogram, especially using modern extraction techniques and sensitive chromatography). However, the number of samples is often limited, and we will look at the problems associated with these ''short, fat, datasets'', where variables can sometimes exceed samples by one to three orders of magnitude. Because there are so many variables, it is possible to find fortuitous correlations between variables and the origins of samples, especially if the ratio of variables to samples is high, which, if incorrectly handled, can lead to an unduly optimistic assessment of predictive ability.

2. Sample size

Typically, the success of a classification model is determined using an indicator such as the percentage of samples correctly classified (%CC). This statistic is commonly cited in the chemometrics literature, and it is usually agreed that the higher the %CC the better the model.

Often, an elaborate strategy is employed to determine whether a model is suitable or not. Usually, a number of samples is left out from the main dataset and are not used for forming the model that is then developed using the remaining (or training set) samples. There are various strategies; the simplest is to have a test set that may, for example, involve a third to a quarter of the original samples being removed. Cross-validation [9] involves removing one or more samples in turn, until each sample has been removed once. The boot-strap [10] involves removing a portion of samples and then repeating the calculation removing the same portion each time, but selecting the samples to be removed randomly; often, 100 or so iterations are performed. Common to all of these strategies, the quality of the model can be determined according to %CC of the samples that have been removed.

A fallacy in the chemometrics literature is to accept this value uncritically; however, it should be interpreted according to sample size. In order to determine whether the underlying data is indeed sufficient in quality for us to be able to assign samples to one or more groups, we are interested in whether the classification ability is significantly better than a random classifier. For example, if we record 100 GC-MS results for urine extracts from male subjects, then, if a classification predicts that 50% of the samples are from male and 50% from female subjects, it is of no predictive value and suggests that the particular analytical technique has no discriminant ability. However, if the model predicts the sex of 90% of the samples correctly as male, is this now significant evidence that a classification model can be applied to the GC-MS data to predict an unknown person's sex (even though there may be errors associated with prediction)?

It probably is, but then consider the case when the model was tested on only 10 males and 9 are predicted correctly, is there still sufficient evidence that the model contains sufficient predictive power to be useful?

In order to interpret the %CC statistic, it is useful to consider the distribution of a toss of a coin. If a coin is tossed 10 times, and it comes up Heads 8 times, and Tails 2 times, is this sufficient evidence that a coin is biassed? This is analogous to asking if a predictive model is tested on 10 samples and 8 are correctly classified, is this sufficient evidence to be convinced that this model really has potential predictive power or could this arise simply as a chance event?

The null hypothesis is that the coin is unbiassed, and, in statistical terms, we are interested in whether we can reject this null hypothesis. The binomial theorem comes to our rescue here. The chance of an outcome of M Heads when an unbiased coin is tossed N times is given by $N!/(M!(N-M)!)0.5^N$. These outcomes are illustrated in Fig. 2 for 10 and 100 tosses of a coin. It can be shown that, when an unbiassed coin is tossed 10 times, we expect to obtain 8 or more Heads 5.4% of the time: analogously, 80% CC could be obtained even using a random dataset more than one time out of 20. Hence validating models on small sample sizes can lead to false optimism about the quality of a model even when the %CC is high. If for example, 40 samples are analysed from a specific group which a quarter of these (10) are removed to test the model we end up we expect 80% CC at least once in 20 times, even if the underlying distribution of variables arises from entirely random processes.

Fig. 3 illustrates the %CC required to obtain a given level of confidence that there is genuinely a distinction between two classes for different sample sizes and so to reject the null hypothesis that the analytical data does not show a distinction between the classes (e.g., if there are 20 samples in the test set, 70% CC could be obtained from an unbiassed (random) distribution around one time in 10 (90% confidence limit)). If one quarter of the samples is removed as a test set, this implies that it is necessary to analyse 80 samples to achieve this level of confidence in prediction.

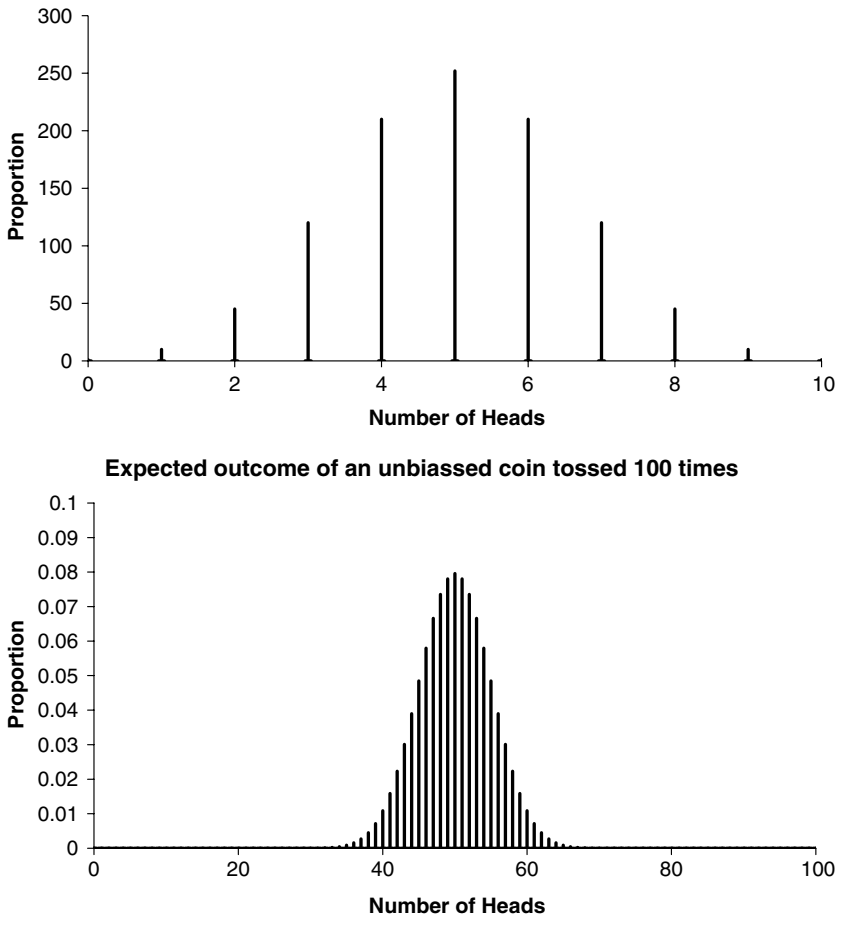

Figure 2. Expected number of Heads for an unbiassed coin.

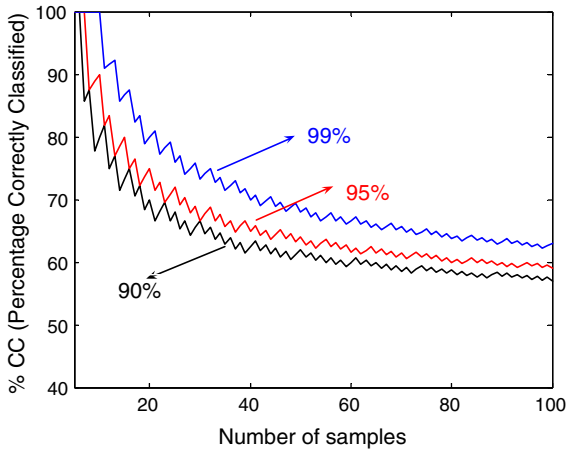

Figure 3. Percent correctly classified (%CC) against sample size required to provide differing degrees of confidence in the classification model. Note that the numbers have been rounded up to the nearest integer and divided by the number of samples.

It can therefore be dangerous to use the %CC as an indication of the quality of a model on small datasets, so the size of the experiments must be planned in advance. In addition, tables of %CC in the analytical chemistry literature should ideally be accompanied by a statement of the number of samples used to obtain this statistic. It is also possible to attach confidence limits to the %CC that the null hypothesis can be rejected.

It is quite common for analytical chemists, especially in biological and medical areas where sampling can be expensive, to report %CC on quite small test sets. However, as will be shown in Section 3, it is always important to determine whether a model is of sufficient quality by dividing the data into a training set and a test set, so quite large numbers of samples (relative to the variables measured) are always desirable unless the trends are very pronounced.

3. Variable selection and model validation

A common and related problem in many areas of chemical-pattern recognition is that most variables are usually irrelevant for the particular classification problem (e.g., in a typical metabolomic dataset, 1000 or more unique chromatographic peaks may be identified). The aim is often to see which peaks are markers for specific groups and then determine the optimum subset of peaks for a classification model. Many peaks (or variables) will be due to the background (e.g., the analytical procedure) and factors irrelevant to the problem at hand (e.g., if we want to predict a person's sex from their urine, there will be numerous other compounds detected from different origins, such as their diet, metabolism, age, whether they have a disease, and personal habits). Including these extra variables may add noise to the model. In some cases, the interesting trends might be masked by other trends (e.g., there is no reason why a person's sex should be the main factor influencing the distribution of compounds found in urine, yet we might want to form a model that determines their sex and we suspect that some chromatographic peaks will be related to a person's gender).

Often only a small combination of variables is necessary for classification. Finding the optimum combination of variables is an important task. However, it is rarely possible to determine models on all possible combinations of variables. Consider, for example, a situation where we want to form a model from 10 out of a potential 1000 possible marker peaks. There are 1000!/(10! 900!) combinations or 2.63×10^{23} such combinations. If we were to try to perform a model on all possible combinations of variables, and it took 1 second per model, it would take 8.35×10^{15}years to complete the work. Since the age of the universe is estimated at slightly over 10^{10}years, the calculation would take 10,000 times longer than the time between the Big Bang and now. Hence, it is not possible to test all combinations of variables and it is usual to reduce or to select variables from the original number to a small subset. Generally, these variables are those that are best for the purpose in hand (e.g., if the aim is to discriminate between two classes, those with the highest discriminatory power, measured in a variety of possible ways, ranging from t-statistics to D-PLS weights, are selected).

The problem is that variable selection is often performed on an entire experimental dataset. This can have a serious consequence when assessing the quality of a model. In order to illustrate it, we will use a small simulated dataset consisting of 200 samples, each a member of one of two groups, denoted by +1 (class A) and -1 (class B), with 200 variables. Each variable has a value of 0 (=absent) and 1 (=present). The variables are generated using a random-number generator with a probability of 0.5 that they take the values of 1 or 0. This is illustrated in Fig. 4 and represents the null hypothesis that there is no difference between the two groups,

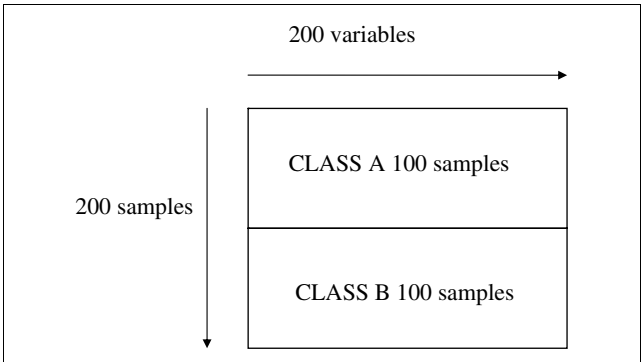

Figure 4. Dataset for example in Section 3. Each variable has a value of 0 (absent) or 1 (present), with 50% probability.

which is a valuable benchmark from which to compare results obtained from real data. Since the numbers were generated randomly, we expect no significant difference between the groups, and would anticipate a %CC of around 50%. The number of times a variable is detected in samples arising from each of the two classes is plotted against each other in Fig. 5, and has a correlation coefficient of -0.0209. The graph of the scores of the first two principal components using centred data on the entire dataset is presented in Fig. 6, suggesting no real separation between the classes. In order to investigate how the quality of a discriminatory model is assessed, D-PLS is performed on these data. In this small example, we consider only a 1 PLS component model, as the aim is not to discuss the choice of models (in fact there is very little difference using different numbers of components in this example), but the effect of variable selection.

Autoprediction involves modelling the data without taking out samples for testing (i.e. all the samples are included in the model). We use the criterion that, if D-PLS predicts a value of the classifier (c) greater than 0, we assign it to class A and if less than 0 to class B. Using autoprediction and 1 PLS component, we obtain 84.5% correctly classified, which is a high percentage correctly classified and we may incorrectly conclude, from this randomly generated dataset, that there is very good classification ability. What the algorithm has done is to determine whether there are any correlations between variables and the classifier and tried to maximise this (commonly called a covariance) to obtain the best possible separation between the groups. However, as most advocates of chemometrics agree, autoprediction can provide a falsely optimistic view of the quality of the model. A more realistic approach is to remove some samples that are not used for the model, called a test set, as discussed in Section 2. In this simulation, we remove

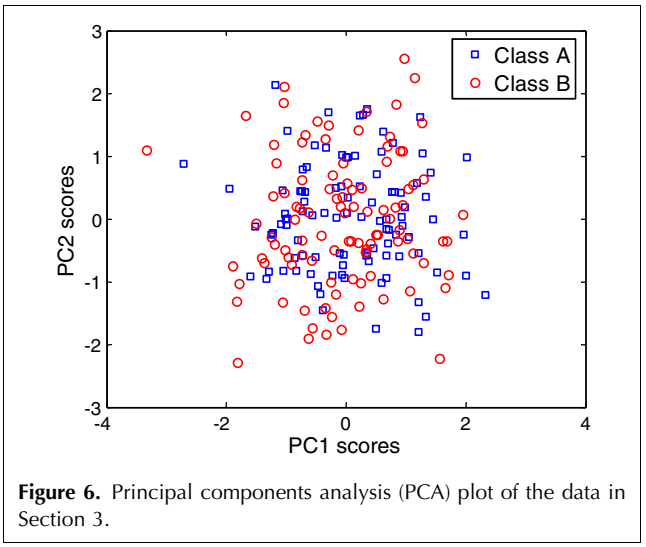

Figure 6. Principal components analysis (PCA) plot of the data in Section 3.

25% of samples, equally distributed among each class, form a model (using 1 PLS component) on the remaining 150 samples (the training set), and then test it on the remainder. Using this new approach we obtain a training set predictive ability of 88% on the 150 samples, but a test set predictive ability of 48% (24 out of 50 samples correctly classified in the test set). This gap between the autoprediction and test set error suggests that the data has been over-fitted. The 48% test set error is around what would be expected, so probably faithfully allows us to assess the quality of the predictive model.

What happens if we reduce the original 200 variables? The best 20 variables can be selected as follows. The 10 variables for which the number of times they are detected in class A minus the number of times they are detected in class B is most are retained, as are the other 10 variables with the opposite property. These variables

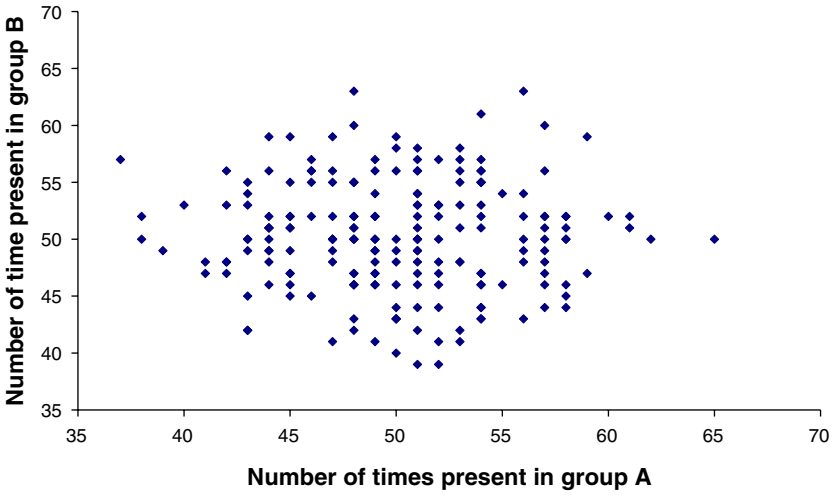

Figure 5. Number of times a variable is present in each class in the simulation in Section 3.

are best potential candidates as markers for each class according to the presence / absence criterion used in this example. One false, but frequently employed, strategy is to select these variables over the entire dataset. Using these new variables, the autopredictive ability on a model formed using all 200 samples is 71%. This is down from 88%, but reflects primarily that there are fewer variables, although better ones; no variable is overwhelmingly good as a marker, in this specific case. Dividing the same samples into a training set and a test set, as above, the training set predictive ability is 73.33% but the test set predictive ability has increased to 60%, giving a falsely optimistic prediction of how well the model is performing. The correct approach is to select variables from the training set only. Using this approach, the training set predictive ability increases somewhat to 76.67%, but the test set predictive ability reduces to 52%, which is a realistic assessment of the performance of the classifier. This can be understood visually in Fig. 7. If variables are chosen from both the training set and the test set there is a correlation between their distributions (0.38 in the case in question), whereas, if they are chosen just from the training set, there is no correlation (0.00 in this example).

When the ratio of variables to samples is higher, the problem becomes even more severe. A simulation involving 40 samples divided equally into two classes, and 1000 random variables, gave a 100% autopredictive ability using 1 PLS component, but a 40% test set predictive ability on 10 samples, 5 randomly chosen from each class. Selecting the best 10 (out of 1000 variables) from both the training set and the test set increases the predictive ability to 90% on the 10 test set samples. However, in many chromatographic experiments, these are typical of the sorts of numbers of samples and variables that are often employed.

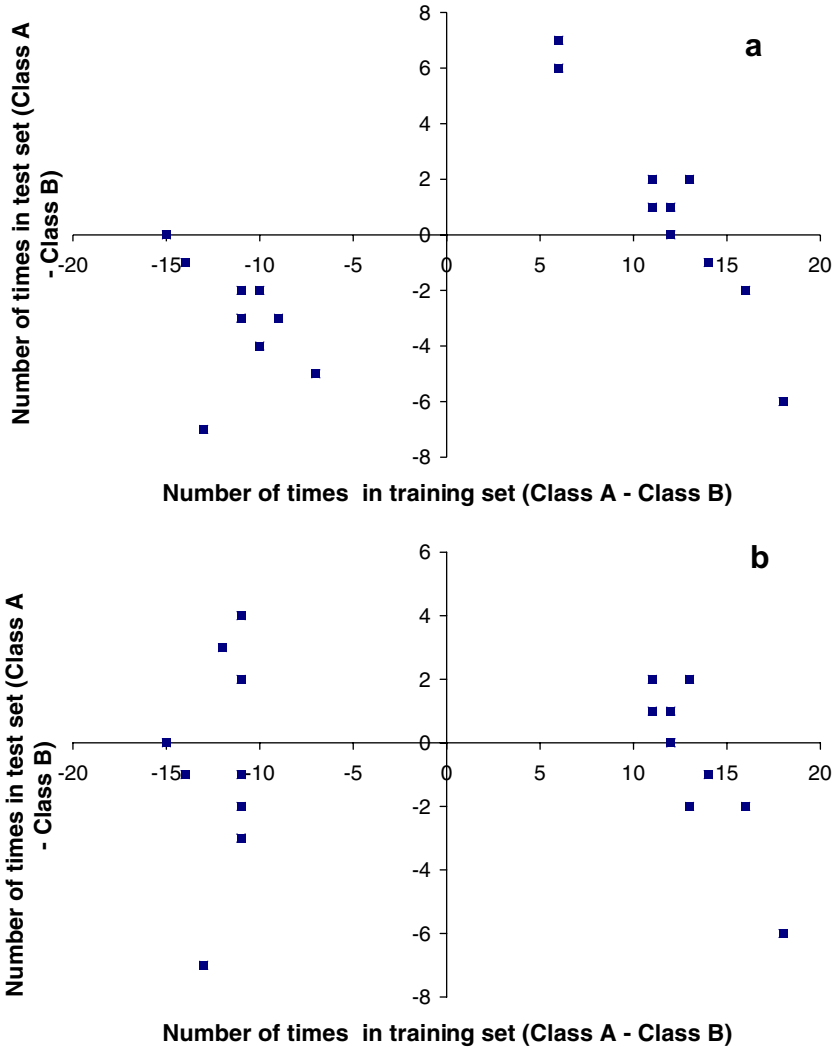

Figure 7. Different strategies for variable selection: (a) selecting variables over both the training set and test set and choosing the 20 best markers; and, (b) selecting variables over the training set and choosing the 20 best markers.

There are two principal computational solutions to this dilemma. The simplest is not to select variables at all. In the first example above, the number of variables and samples is equal, and this may be feasible; but, when the number of variables far exceeds the number of samples as in the second example, often this approach is not practicable. Diagnostic variables are likely to be hidden; if we measure 100 variables, perhaps 10 or 20 will be useful, so some form of variable selection is required. The greater the ratio of variables to samples the more necessary it is to consider variable selection: in many cases, this is mandatory, as there will often be variables due to the background or other factors that are irrelevant to the classifier and in future samples may not be present or will be irrelevant due to completely different factors. Consider using extracts from urine to determine the onset of a disease: the majority of compounds will have no diagnostic value and should not be used in the classifier. Under these circumstances, a second alternative is to select variables just from the training set. This may seem the ideal solution given the discussion above, but many methods for model building and validation do not use a single training set but several different training set and test set splits of the data each comprising different subsets of samples (see Section 4), the overall assessment of modelling ability being an average of each iteration. A common approach – leave-one-out cross validation – involves cycling round each sample once, and reforming the model on a training set minus that sample. Hence variable selection should ideally be performed afresh each time a sample is removed. This means that each model is formed on a different subset of variables, but this then causes problems because each model is not completely comparable. How serious this problem is depends partly on the size of the test set. If the test set is quite small, it is not so serious a problem, because the number of samples in the test set is limited, so it will not have a large influence on whether a variable is retained or not. However, if the test set is large relative to the overall sample size, it will have a significant influence, so variable selection must be repeated each time a test set is generated. But many methods in chemometrics, such as the bootstrap, do advocate quite large test sets (typically a third of the dataset), repetitively generated, because obtaining reliable confidence in predictive power requires quite large test sets, as discussed in Section 2.

Of course, the experimental solution is to increase the number of samples assayed, but this in turn may cause problems. Often sampling is expensive, particularly in clinical work, and there may be problems of stability, for example, long-term instrumental and measurement stability, meaning that when a sample is recorded becomes an additional factor in the data interpretation. In addition if analysis times are slow, there may be a differential time between storage and running of samples, which could increase as a backlog builds up. Fast throughput using autoinjectors and constant checking of chromatographic quality is one possible additional experimental answer to the problem, although there is no universal solution and every problem must be considered on its merits. The most important issue is to always consider the motivation of the analyses, and ideally to perform some pilot experiments to determine how serious these problems are, and not to divorce the chemometrics, which may dictate the ideal sample sizes for a specific type of problem, from the experimental sampling design.

4. Model validation and optimisation

A final difficulty is that optimisation of a model can be confused with its validation. Often, methods such as cross-validation are employed to optimise a model (e.g., to determine the number of PLS components required to produce the best model).

Many people calculate the %CC or an equivalent indicator of the quality of the classification (or calibration) model using the samples that are left out during cross-validation (or an equivalent method, such as the bootstrap) using the training set. A problem here is that the model is optimised for these samples that are left out, so, if we use the %CC for the samples used for model optimisation, we are not truly assessing the quality of the model using an independent criterion. The most common way around this is to have a third, and independent, test set, which is used to determine how well the model is obeyed (see Fig. 8). From the training set, one or more groups of samples are removed, often repetitively until all samples are removed at least once, to optimise the model. A separate group of samples is left out from the original data to form the test set. The quality of the model is assessed using this group of samples left out to form the test set rather than the samples used to optimise the model.

Sometimes, this can be done iteratively, with different test sets created in a loop, often many times over, and the average %CC is determined using a model that is optimised using the training sets. Remember that each model will be calculated using a different training set, often, if employing principal components analysis (PCA) or PLS-based methods, using a different number of components. If variable selection is employed, each model may well be formed using different variables also, as discussed in Section 3.

The dilemma is that, if a successful predictive model is obtained, it is then necessary to choose a model that will be applicable to the overall dataset and to new samples, yet the assessment of the model has been performed on one or more subsets of samples, each with its own number of components, data scaling and variable selection. There are no accepted criteria for the way to

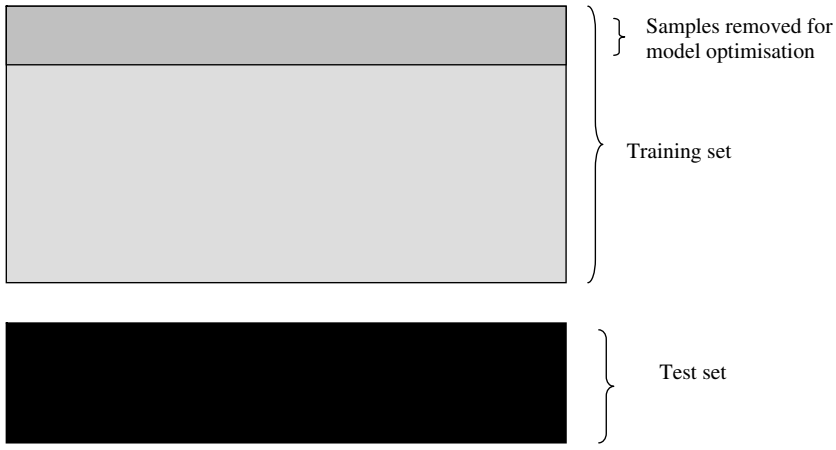

Figure 8. Dividing data into three blocks for optimisation, autoprediction and testing.

choose the overall model, but often one may then return to the full dataset, using the overall dataset as a training set, and form a model afresh using this.

5. Conclusions

This article has focussed on assessing the quality of methods for classification. With the wide availability of chemometrics software, there are numerous reports in the literature. An important growth area concerns datasets where there is a large number of variables, especially in coupled chromatography of biomedical data. A problem emerges in that large datasets are expensive to obtain, but that the number of variables recorded is often quite large. It is easy to obtain falsely optimistic estimates of predictive ability under such circumstances and great care is required to determine the correct interpretation of the results of classification. Yet, with the expansion especially of the application of chemometrics in biology, such problems are becoming more common and it is important to understand the consequences of handling such large datasets, common in metabolomics and proteomics, to avoid coming to false conclusions.

The examples and numbers presented in this article should provide useful starting points that can be extended in individual circumstances. The article also provides an example of how simulations of the null hypothesis (i.e. there is no difference between groups) can be used to validate methods and interpret results of pattern recognition. Table 1 summarises some common problems discussed in this article and their solutions.

With the widespread availability of classification methods in modern analytical chemistry, especially at

Table 1. Some common problems associated with pattern recognition and their solutions

Common problems	Solutions
Quoting %CC without providing information on the sample size used for testing the validity of the model	Quote the test set sample size together with the %CC
Using sample sizes that are too small to determine whether a high %CC could have arisen by chance	If the confidence level is low, use a larger sample size
Not considering the null hypothesis	Under certain circumstances, it is helpful to use null simulations that are as close as possible in size to the observed dataset and compare %CC
Selecting variables from the overall dataset including test samples	Select variables only from the training set
Obtaining test set and training set predictions that differ significantly	Often the data are not sufficient for the determination of an adequate predictive model: increase sample size or modify analytical technique
Confusing optimisation with testing	Divide samples into three groups, one for independent testing, one for determining the optimal model and one for autoprediction, often using iterative methods

the biological interface, having an understanding of relatively simple principles is crucial for their use to be valid.

References

[1] R.G. Brereton, Chemometrics: Data Analysis for the Laboratory and Chemical Plant, Wiley, Chichester, West Sussex, UK, 2003 Chapter 4.
[2] R.G. Brereton, Applied Chemometrics for Scientists, Wiley, Chichester West Sussex UK (In press, publication due February 2007).
[3] R.G. Brereton (Editor), Multivariate Pattern Recognition in Chemometrics (illustrated by case studies), Elsevier, Amsterdam, The Netherlands, 1992.
[4] S.J. Dixon, Y. Xu, R.G. Brereton, H.A. Soini, M.V. Novotny, E. Oberzaucher, K. Grammer, D.J. Penn, Chemom. Intell. Lab. Syst., submitted for publication.
[5] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, Chemom. Intell. Lab. Syst. 50 (2000) 1.
[6] S. Wold, Patt. Recognit. 8 (1976) 127.
[7] M. Defernez, K. Kemsley, Trends Anal. Chem. 16 (1997) 216.
[8] Y. Xu, S. Zomer, R.G. Brereton, Crit. Rev. Anal. Chem. (In press).
[9] S. Wold, Technometrics 20 (1978) 397.
[10] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman and Hall, New York, USA, 1993.