THE CHALLENGES OF VALIDATING MULTIVARIATE METHODS FOR PATTERN RECOGNITION

Richard G Brereton

School of Chemistry, University of Bristol, Cantocks Close, Bristol BS8 1TS, United Kingdom
E-mail : *r.g.brereton@bris.ac.uk*

Abstract

Multivariate methods for pattern recognition are increasingly used in data mining of complex biological processes such as in metabolomics and food science. Whereas the flexibility of such methods for example Support Vector Machines or Self Organising Maps or Partial Least Squares Discriminant Analysis allows for highly sophisticated models, there is a comparable problem of overfitting. Validation is therefore important. It is essential to distinguish between optimisation and validation. It is important to consider the challenges of data with high variable to sample ratios. Variable (or feature) selection can be problematic as if incorrectly done it can accidentally introduce over-optimistic results. Iterative but computationally intense methods are often needed to repeatedly generate training sets and even out the problems of outliers or mislabelled / atypical samples that could unduly influence the training or test sets. Finally performance criteria can be hard to define, as indicators of success depend in part on what is known about the data in advance, so the primary aim of a method may not necessarily be to reduce apparent error rates in test sets : many methods available appear over-attractive because they aim to provide an over-optimistic rather than realistic view especially in internal test sets that may not contain the features of future data.

Keywords: Pattern Recognition, Validation, Multivariate Analysis, Performance Criteria, Classification

1. Introduction

Many years ago comparison and validation of methods was easy. The method that apparently had lowest error rates was considered the best. Typically a portion a dataset was divided into two sections, the training set, used to develop a model, and a test set, used to determine how well the model performs. Several approaches could then be compared, and the one with the lowest error rate (defined in a variety of ways) for the test set was chosen as the most appropriate. Sometimes, this error rate was determined using an internal test set, for example via cross-validation. This so-called optimum method could then be used to determine the provenance of new samples, and assign them, where appropriate, into groups.

A large literature was developed and numerous papers published claiming superior performance for novel methods, and so guiding investigators. In this presentation, we will focus primarily on multivariate classification methods, so an approach that have the lowest classification error is generally considered the most successful.

There are numerous difficulties with this.

- There is no guarantee that the underlying hypothesis, that there is sufficient information to unambiguously separate two or more groups, is correct, and so a method that results in the lowest number of misclassifications may not necessary be the most appropriate, and may not be appropriate to use for future unknown samples.
- The choice of test set may be influenced by a few atypical samples. Unless a dataset is very large, the inclusion of, for example, an outlier in the test set may make a significant difference, especially when comparing closely comparable performances.
- Often variable selection is done on the overall dataset, as many variables can contribute significantly to noise and deterioration of performance, but the criteria for variable selection can bias the model, and should only be done on the training set.
- Frequently optimization and validation are mixed up, for example using cross-validation, resulting in an over-optimistic assessment of the model.
- There is rarely much discussion about how certain knowledge is of the dataset prior to forming the model and also what sort of structure (if any) is expected for future samples.
- Many comparisons focus on one algorithm within a series of steps, for example PLS-DA (Partial Least Squares Discriminant Analysis). However the performance of the algorithm often depends crucially on all the steps in the analysis : as an example under certain circumstances PLS-DA performs identically to LDA (Linear Discriminant Analysis), and others to EDC (Euclidean Distance to Centroids). In practice algorithm performance is dependent on a series of decisions about data preprocessing and success criteria.
- Often different approaches are not comparable. For example SIMCA (Self Independent Modelling of Class Analogy) is in practice a one class classifier and cannot be compared directly to two class classifiers such as PLS-DA or LDA.
- Sometimes classification methods are used as exploratory approach, for example to determine which variables (or features) are most likely to be discriminatory and in other cases they may be used to predict the provenance of future unknown samples. There is no general distinction between these objectives, the literature being primarily algorithmic.

There is, therefore, no general guidance as to what is the best or most appropriate method. Each situation is different. For very simple problems, of course, there is no need to introduce a complicated comparison, but then almost any reasonable method will work and will give good results. Where pattern recognition methods are more tricky is when there are problematic features in a dataset and we will examine some ways of dealing with such situations below.

2. How to assess models to protect against overfitting

There are numerous approaches for safe application of multivariate models, but we will discuss some of the most common.

2.1 Separating validation from optimization

Many years ago, these two quite distinct processes were often confused. Cross-validation was commonly used both to improve the quality of a model and to determine how well it performed. This procedure involves leaving one or more samples out from a dataset, forming a model on the remaining samples, and then determining how well it performs on the left out samples. The procedure is repeated as different samples are left out, until all get removed at some stage. The most common approach is

LOO (leave one out) in which each sample is removed once, but other procedures such as leaving a group out have been reported in the literature : as computer power improves, it is often not necessary to leave groups out, the original procedure was developed to save computer resources.

Cross validation was used to determine, for example, how many components (or latent variables) are needed for optimum model performance. Theoretically as too many components are included, the model degrades due to noise, but with too few it is not adequate. However the problem is that the cross-validated error is also sometimes used to assess model performance. The theory is that the samples left out form a test set and so provide an independent criterion of the quality of the model. The difficulty is that these left out samples are also used to optimize the model.

The solution is to separate model optimization from model testing, by performing model optimization on a training set typically 2/3 of the original samples and model evaluation on a test set, the remaining samples. In addition more computationally intense methods such as the bootstrap are often preferred to cross-validation. A typical procedure is illustrated in Figure 1.

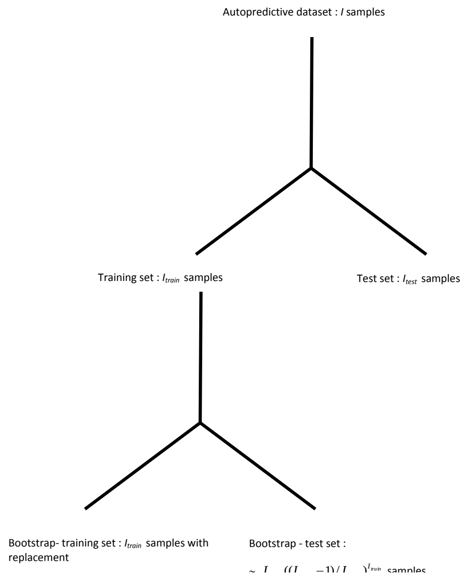

Autopredictive dataset : *I* samples
Training set : I_{train} samples
Test set : I_{test} samples
Bootstrap- training set : I_{train} samples with replacement
Bootstrap - test set :
$\sim I \quad ((I \quad -1)/I \quad)^{I_{train}}$ samples

Figure 1: Typical separation between optimization and validation

2.2 Iterative approaches

The solution described above is to separate model optimization from model testing, by performing model optimization on a training set typically 2/3 of the original samples and model evaluation on the remaining 1/3. But often people compare model evaluation on a quite small subset, for example 30 samples. Classifying 1 sample correctly makes more than 3% difference in the percentage correctly classified, and could be the difference between choosing one model over the other. Hence comparisons between methods and indeed validation of an approach can be crucially influenced by the composition of a test set.

The solution here is to repeatedly generate test and training sets. Typically one could generate a test set 100 times. If 200 bootstrap iterations are performed on the training set, this requires 20,000 models to be built. However modern computers are much more powerful than decades back, when many multivariate methods for classification were first introduced and as such were limited by processor power.

A typical procedure is illustrated in Figure 2.

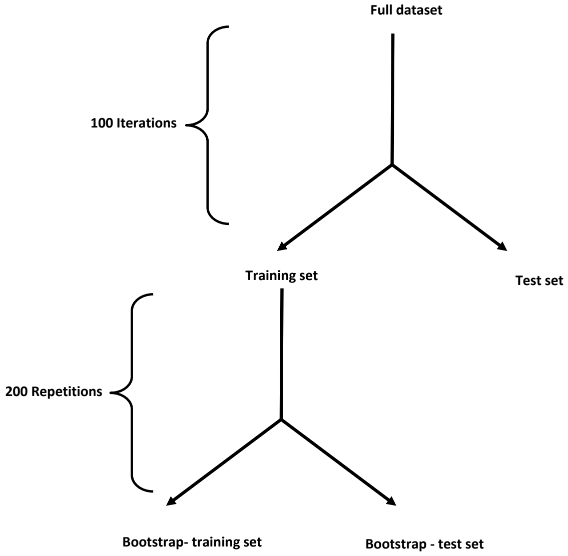


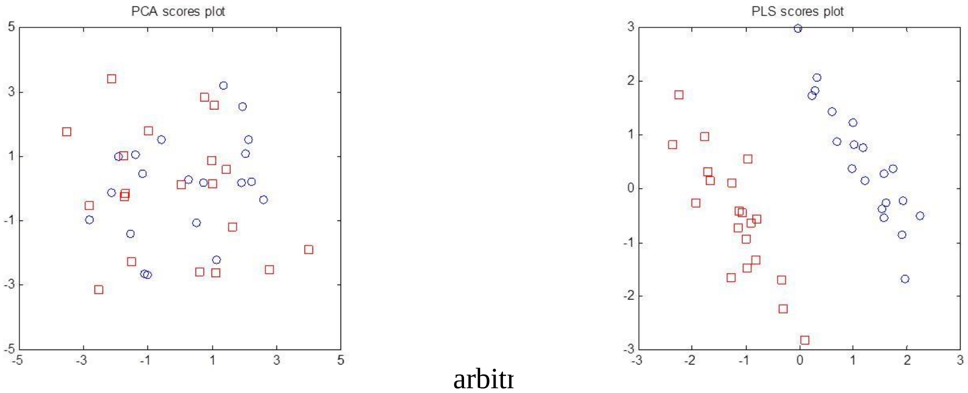
Figure 2: Iterative approach to optimization and validation

2.3 Null datasets and permutations

Some methods attempt to force a solution. One of the problems is that in many modern studies the ratio of variables to samples is usually far greater than one. This means it is easy to overfit a model.

Consider, for example, the toss of an unbiased coin, representing 10 samples, and then repeat this 1000 times, representing 1000 variables. A variable to sample ratio of 100 (the variables being for example GCMS elution times) is not unusual, something early statisticians rarely encountered. There will however be several incidences in which 8 or 9 coins are Heads. If we repeated 10 tosses of an unbiased coin 1000 times over, we expect there to be around 45 times in which the coin comes up Heads 8 times or more. We could then select these sets of tosses, rather like selecting features that appear most diagnostic of separation and form a model using these "informative variables". We will find that the coin incorrectly appears biased.

Some methods such as PLS-DA are in practice designed to find variables that have the highest covariance with a classifier, usually a numerical value such as +/- 1 whether a sample is a member of a predefined class or not. These variables are then used to form a model between the analytical data and the classifier. The problem with this is that it is very easy to find false positives. Randomness is not the same as uniformity, and there will always be some variables that show correlation even if there is no underlying population trend. This problem is particularly severe when there are many variables. Figure 3 shows the PLS scores plot of two randomly generated groups, and it appears there is good separation. Often this is used to justify experiments have worked when in practice this is just a consequence of low sample : variable ratios.



Fi arbitr
scores of the first two principal components and partial least squares components

To protect against this type of problem, it is desirable to have a background or control dataset. The simplest is a null dataset, which is simply a randomly generated dataset. If a method has been correctly implemented, the classification per ability to group samples into two classes, for example, should be close to 50%. If it is much higher the method chosen is prone to overfitting. A second approach involves permuting the classifier randomly. This is sometimes called a Monte Carlo method. Again in the absence of overfitting, the classification ability should be close to 50% for two groups. Sometimes there could be some deviation from these ideal situations, even if methods chosen are appropriate. To protect against this, repeat the permutation (or regenerate the null dataset) several times, to get an average.

2.4 Comparing Methods

One of the commonest difficulties is in method comparison. It is quite common to try to compare the performance of one approach against another. This is dangerous procedure.

The first problem is that whether one method is suitable or not depends on the nature of the data. A test dataset may, for example, be simulated, so does this represent real data? In the real data are there likely to be outliers or misdiagnosed samples? How representative is the real case study, and can we guarantee it will have the same structure in the future? For agricultural studies this can be particularly serious as soil conditions, weather, producers and so on can change will time. Often methods are tested on such artificial datasets that they have little relevance to the problem in hand.

Sometimes methods are not comparable. A one class classifier, such as one class QDA (quadratic discriminant analysis) or SIMCA results in quite different verdicts to a two class classifier such as PLS-DA or LDA. In the former we try to determine whether a sample belongs to a predefined class or not. If there are two classes, there are four verdicts. 1. Class A. 2. Class B. 3. Neither. 4. Both. However a two class classifier has just two verdicts, Class A or Class B. It is not possible to sensibly compare these approaches.

The third problems is that a classification algorithm is only one of several steps in pattern recognition, for example, deciding on test / training sets, preprocessing, classification criteria, selecting the number of components as appropriate. As a simple example PLS-DA can provide identical results to LDA if all non-zero components are chosen, and columns are transformed in a specific way, so how can we compare PLS-DA to LDA unless we know how the data was preprocessed? The classification algorithm should be considered just as a single step among others, and may not be the most critical and all steps in a classification protocol should be reported in detail.

Fourth, assumptions about the data are critical, as is the composition of the training and test sets. It is impossible to control all the features in future, often samples of unknown provenance, and these may often not be represented in the original test data. Often methods that appear to perform better on a test set are more strict in their requirements for the correlation structure of future datasets, so should a method be more robust to future outliers but perform less well on a test set, or perform better on the test set but result in quite poor prediction if future data contains outliers?

There is no universal solution to these dilemmas. However each classification problem should be treated as a unique challenge and there should be no universal solution, often requiring time to find an appropriate protocol. The idea that one method is “better” than another is dangerous. It is a good idea to use two or more totally different protocols for assessing the provenance of samples and come to a consensus conclusion.

3. Conclusion

This presentation has summarized only a few approaches designed to protect against over-optimistic models, and so false positives both when classifying samples or determining whether variables are influential and so potential diagnostic markers.

As variable to sample ratios increase, many classical methods, often designed when features or variables were hard to measure, break down. This means that overfitting becomes a serious problem, and also the difficulties of datasets that may contain outliers or misclassified samples or even unknown hidden groups or substructures.

However, in contrast, computing power has increased vastly. So much more computationally intense methods such as the bootstrap and Monte Carlo methods can be used to protect against these problems. There is no universal answer to these dilemmas except to be aware of the limitation and to take every problem as a unique one.

References

Brereton, R.G. (2009). Chemometrics for Pattern Recognition, Chichester : Wiley

Brereton, R.G., & Lloyd G.R. (2014). Partial Least Squares Discriminant Analysis. Taking the Magic Away, *Journal of Chemometrics*, 28, 213-225