

Using cross-validation to avoid selection bias in the graphical representation of high-dimensional data

J.H. Maindonald and C.J. Burden

Australian National University, Canberra, Australia

Microarray data is typically very high-dimensional, with many fewer features than samples. Here, I will use a data set that has 7000 features and 40 samples (observations), where the aim is to find a classifier that will classify new observations (samples) into one of three known groups. While it can be effective to use those features that are individually the best discriminators, care must be taken to account for such feature selection, both in the assessment of predictive accuracy and in graphs that are designed to show the separation into groups. Naive approaches that ignore the selection bias will inevitably give graphs that, for random data or for data where the group labels have been randomly permuted, show a clear but spurious classification into groups.

A cross-validation procedure that repeats the selection of features at each cross-validation fold gives estimates of predictive accuracy that have a small downward bias, and can be used to determine how many features should be used for optimum predictive accuracy. This paper demonstrates a methodology that uses results from the cross-validation to give a graph that accurately reflects the performance of the classifier. For present purposes, a general linear discriminant classifier is used, as in canonical variate analysis. The method does however have more general application, to any methodology that leads to discriminant scores from which an acceptably accurate low-dimensional representation can be derived.

Scores from the different cross-validation folds are all transformed back on to a “global” set of co-ordinate axes, here chosen to be the discriminant axes obtained when discriminant functions are based on a principal components representation of the data. The computation required for the transformation makes a relatively straightforward use of the singular value decomposition.

Among other possible choices of classifier, Support Vector Machines (SVM) have the advantage that no preliminary selection of features is necessary, but may be incorporated for reasons of parsimony. Decision values that have been obtained from cross-validation can be approximated in a two-dimensional space, and used as the basis for a graph, though in a less natural than for canonical variate analysis.

Similar methodology can be applied to the plotting of principal component scores, which are likewise subject to variable selection effects, though less severe than for discriminant analysis. The methodology has application to other types of high-dimensional data.