L01. One class classifiers in chemometrics

Richard G. Brereton
Centre for Chemometrics, School of Chemistry, University of Bristol, UK

One class classifiers involve modelling each class in a dataset independently, in contrast to two class classifiers that divide dataspace into two regions. The difference between one and two (multi) class classifers, between hard and soft models and between disjoint and conjoint PC models will be described.

Many one class classifiers in chemometrics derive from QDA (quadratic discriminant analysis), and attempt to form boundaries at a given level of confidence from the centroid of a class, If PC reduction is performed first then this can be either conjoint (on all data) or disjoint (on each class separately — the principle of SIMCA). For disjoint models it is advisable first to see how well the data fit into the PC model, often using the Q statistic. The D statistic, based on Hotelling's T2 can be employed subsequently. It is possible to derive joint Q and D confidence limits. We illustrate the use of these statistics using iterative methods for formulating models to define Predictive Ability and Model Stability.

If data do not well fit into a normal distrbution, these methods may not be appropriate, and an alternative, Support Vector Domain Description (SVDD) allows for more complex boundaries, and will be illustrated. In addition to the advantage of not requiring multinormality, SVDD do not require all models to be concentric, overcoming a problem of least squares solutions that can be influenced unduly by outliers. Movies will be shown to illustrate the change in boundaries according to the values of the penalty error and gaussian radius (for Radial Basis Functions).

Problems of optimising and validation of one class classifiers will be discussed and strategies for overcoming these. The use of ROC curves and Class Membership plots will be introduced.

The methods will be applied to a variety of datasets including simulations, metabolomics, polymer characterisation, forensics and environmental. They will be compared to two class and multiclass models. One class approaches are especially useful where there are many groups in the data or alternatively e.g. in multivariate process control where only one class (e.g. the Normal Operating Conditions region) can be adequately characterised.