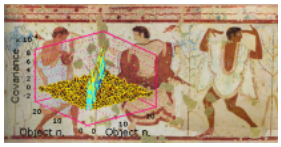
Invited Lecture at CMA4CH 2008, Mediterraneum Meeting

Multivariate Analysis and Chemometry Applied to Environment and Cultural Heritage
2nd ed., Ventotene Island, Italy, Europe, 1-4 June 2008



Application of Modern Supervised Pattern Recognition Methods in Chemometrics

Richard G. Brereton, G. R. Lloyd, S.S. Fong, K. Wongravee, M. Z. Jaafar

Centre for Chemometrics, School of Chemistry, University of Bristol, Bristol BS8 1TS, U.K.

Abstract

Over the past two decades there has been a growth in pattern recognition methods, especially catalysed by machine learning community and the rapid growth in computing power. Methods developed two or three decades ago such as principal components analysis, cross-validation and partial least squares, required limited computing power and are now embedded into modern software packages. Moore's law as variously described suggests a doubling of computer speeds every 2 years, or over 30,000 times increase in 30 years, yet modern pre-packaged chemometric software has not kept pace. Many problems are non-linear especially outside mainstream analytical chemistry and as such are require approaches often more usual in areas such as economics or biology. In addition proper validation and optimisation usually requires significant iterations, for example using a bootstrap and test / training set splits might require a model to be reformed 20,000 times. In addition, self organising maps are a powerful alternative to principal components for the visualisation of relationships between samples. These methods are illustrated on a dataset of ancient Italian pottery coming from different sites.

Introduction

Whilst many of the new approaches to pattern recognition were quite theoretical and of limited applicability several years ago, due to the lack of availability of computing power, these are now feasible as tools for pattern recognition within chemometrics. In many modern studies, such as chemical archaeology, we do not necessarily expect linearly separable data and the datasets are much more complex than for example in traditional analytical chemistry and so require new tools. This presentation is illustrated by a prepublished dataset but the methods in the presentation should be applicable to more complex problems that are now feasible to be studied using chemometrics.

Materials & Methods

The dataset will be illustrated a dataset consisting of measurements of 11 elements on 58 pottery samples from Southern Italy [1]. The aim is to try to classify the pottery into two groups A (black carbon containing bulks) and B (clayey ones) according to their elemental composition.

For simple data display we use Principal Components Analysis. The advantages over PCA of Self Organising Maps (SOMs) [2] for data display and variable selection are demonstrated, using in-house software developed in the authors' laboratory.

For supervised classification or pattern recognition we have developed methods of iteratively splitting the data into test and training sets 100 times over and further using 200 bootstraps in each iteration for optimisation [3] as originally reported for PLS-DA. Various performance indicators including the %correctly classified on the test set, the area under the curve, the predictive ability of individual samples and Receiver Operator Characteristic (ROC) curves are outlined.

Several additional classification approaches have been employed including the Euclidean distance, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Learning Vector Quantisation (LVQ) [4] and Support Vector Machines (SVMs) [5] are employed. These methods involve creating increasingly complex boundaries, starting from linear to quadratic, multilinear and curved, between classes.

Results

SOMs have significant advantages over PCA for data display. They can be used to visualise the grouping into two classes, or by geographical origins. Where there are lots of factors it is hard to find enough symbols in a PC plot, and the PC plots also often waste space and sometimes the largest PCs are not the most significant. SOMs are computationally intense but very effective for exploratory data analysis. They can also be used to determine which elements are most important for separating groups, fig. 1.

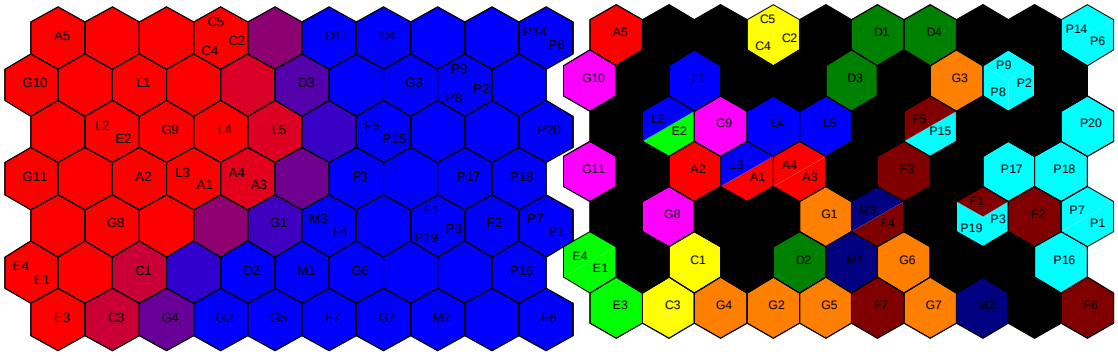


Fig. 1: Representation of objects by class and by origin using SOMs

The performance of the different classifiers can be compared by visualising the boundaries between different groups, as represented on the PC scores projection. Euclidean and LDA classifiers draw linear boundaries, QDA quadratic, LVQ multilinear and SOM quite complex curved boundaries. The example in this study is fairly simple but for complex datasets often with many classes with non-Gaussian distributions the choice of classifier can be quite important and depends on data structure, fig. 2.

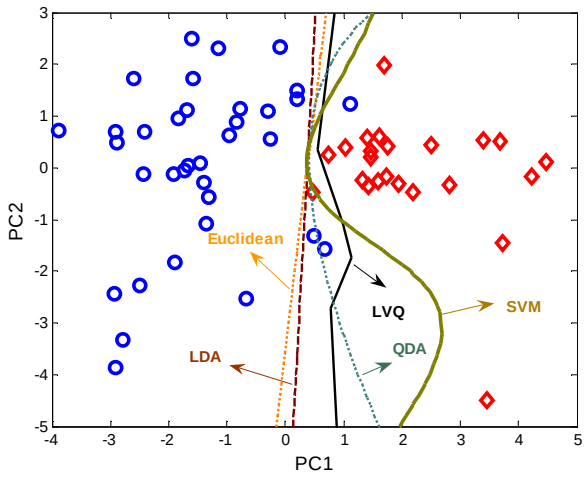


Fig. 2: Different types of boundaries between two classes

Conclusions

There are a large number of methods now available for handling complex real world datasets such as are found in the area of chemical archaeology and other cultural studies. These are feasible using more computing.

References

1) P. Bruno, M. Caselli, M. L. Curri, A. Genga, R. Striccoli, A. Traini, Chemical characterisation of ancient pottery from south of Italy by Inductively Coupled Plasma Atomic Emission Spectroscopy (ICP-AES) Statistical multivariate analysis of data, Anal. Chim .Acta., 410(1-2) (2000) 193-202

2) G.R. Lloyd, R.G. Brereton, J.C. Duncan, Self Organising Maps for distinguishing polymer groups using thermal response curves obtained by dynamic mechanical analysis, Analyst, in press (2008)

3) Y. Xu, R.G. Brereton, K. Trebesius, I. Bergmaier, E. Oberzaucher, K. Grammer, D.J. Penn, A fuzzy distance metric for measuring the dissimilarity of planar chromatographic profiles with application to denaturing gradient gel electrophoresis data from human skin microbes: demonstration of an individual and gender-based fingerprint, Analyst, 132(7) (2007) 638-646

4) G.R. Lloyd, R.G. Brereton, R. Faria, J.C. Duncan, Learning Vector Quantization for Multiclass Classification: Application to Characterization of Plastics, J. Chem. Inf. Model., 47(7) (2007) 1553-1563

5) Y. Xu, S. Zomer, R.G. Brereton, Support Vector Machines: A Recent Method for Classification in Chemometrics, Crit. Rev. Anal. Chem., 36(3-4) (2006) 177-188