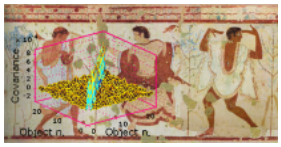
Invited Lecture CMA4CH 2008, Mediterraneum Meeting

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



Classification Methods in Chemometrics

Federico Marini

Department of Chemistry, Sapienza University of Rome, P.le Aldo Moro 5
I-00185, Rome, Italy

Abstract

Pattern recognition methods, i.e. the methods concentrating on the possibility of assigning an object to a class based on the result of a set of measurements are ubiquitous in chemometrics.

In this lecture, the main chemometric classification methods are discussed in terms of their nature, behaviour, advantages and drawbacks. Both parametric and non parametric or discriminant and modelling techniques will be illustrated by means of real examples and simulated data opportunely design to evidence specific features of the individual methods.

The importance of good validation procedures will be eventually discussed.

The 5 Ws of Pattern Recognition

A significant part of the applications of chemometric techniques in analytical chemistry falls in the general framework of pattern recognition, i.e. the classification of objects in groups based on the results of a series of measurements [1]. From a theoretical point of view, several distinction can be made among the different pattern recognition techniques, for example linear/non-linear if our attention is focused on the mathematical form of the decision boundary or parametric/non-parametric if we're more interested in whether a specific underlying probability distribution is assumed or not [2].

Particularly important is the distinction which can be made among pure classification and class-modelling techniques. Pure classification techniques are mainly oriented in discriminating among the different groups and operate dividing the hyperspace in as many regions as the number of classes so that, if a sample falls in the region of space corresponding to a particular category, it is classified as belonging to that category: in this way, each sample is always assigned to one and only one class [2]. This kind of methods include Linear and Quadratic Discriminant Analysis (LDA & QDA) [3], K-Nearest Neighbors (KNN) [4], Partial Least Squares-Discriminant Analysis (PLS-DA) [5], Back-Propagation and Counter-propagation Artificial Neural Networks (BP- & CP-ANN) [6], Support Vector Machines [7] and D-CAIMAN [8]. On the other hand, class-modelling techniques represent a different approach to pattern recognition, as they focus on modelling the analogies among the elements of a class rather than on discriminating among the different categories. In class-modelling each category is modelled separately: objects fitting the model are considered part of the class, while objects which don't fit are rejected as non-members. When more than one class is modelled, three different situations can be encountered: each sample can be assigned to a single category, to more than one category or to no category at all [2]. With respect to pure classification techniques, class-modelling tools offer at least two main advantages: it is in principle possible to identify samples which don't fall in any of the examined categories and which, as a consequence, can be either simply outlying observations or members of a new class not considered during the modelling stage; moreover, as each category is modelled separately, any additional class can be added without recalculating the already existing class models. The most commonly used chemometric class-modelling techniques are SIMCA [9, 10] and UNEQ [11]: the former describes the similarities among the samples of a category using a principal component model, while the latter is instead based on the assumption of multivariate normality for each class population and can be considered

as the modelling analogue of Quadratic Discriminant Analysis. In UNEQ, the class model is represented by the class centroid and the category space is defined on the basis of the Mahalanobis distance from this barycenter, corresponding to a desired confidence level (usually 95%). Since when a category is not homogeneous, the class space can be highly irregular, recently other non-linear class-modelling techniques have been developed, namely modelling CAIMAN [8] and two modelling versions of artificial neural networks [12, 13].

Why this talk?

In this communication the main characteristics of the different available discriminating and modelling pattern recognition methods will be illustrated in terms of their algorithmic model, their requirements in terms of sample to variable ratio, their preferred field of usage and their drawbacks, so to illustrate how the possibility of having a great choice of classification tools can by far be an advantage but can also result in significantly wrong outcomes if the methods are not used properly.

In this framework, the concept of proper validation will be stressed and some basic rules of thumb will be given.

References

1) B. R. Kowalski, C. F. Bender, Pattern Recognition: A powerful Approach to Interpreting Chemical Data, J. Am. Chem. Soc., 94(16) (1972) 5632-5639
2) B. G. M. Vandeginste, D. L. Massart, L. M. C. Buydens, S. De Jong, P. J. Lewi, J. Smeyers-Verbeke, Supervised Pattern Recognition. In: Handbook of Chemometrics and Qualimetrics: Part B. Elsevier, Amsterdam, 1998, 207-241, ISBN: 0-444-82853-2
3) G. McLachlan, Discriminant Analysis and Statistical Pattern Recognition. Wiley, New York, (1992) ISBN: 0471615315.
4) D. Coomans, D. L. Massart, Alternative K-nearest neighbour rules in supervised pattern recognition. Part 1. K-nearest neighbour classification by using alternative voting rules, Anal. Chim. Acta, 136 (1982) 15-27
5) S. Wold, C. Albano, W. J. Dunn III, K. Esbensen, S. Hellberg, E. Johansson, M. Sjöström, Pattern Recognition: finding and using patterns in multivariate data. In H. Martens, H. Russwurm Jr. (eds): Food Research and Data Analysis, Applied Science, Barking, (1983) 147-188, ISBN: 0-85334-206-7
6) J. Zupan, J. Gasteiger, Neural networks in chemistry and drug design, Wiley-VCH, Weinheim, Germany, 2nd ed, 1999, ISBN: 3-527-29779-0
7) Y. Xu, S. Zomer, R. Brereton, Support vector machines: a recent method for classification in chemometrics, Crit. Rev. Anal. Chem., 36(3-4) (2006) 177-188
8) R. Todeschini, D. Ballabio, V. Consonni, A. Mauri, M. Pavan., CAIMAN (Classification and Influence Matrix Analysis): a new approach to the classification based on leverage-scaled functions, Chemometrics Intell. Lab. Syst., 87(1) (2007) 3-17
9) S. Wold, Pattern Recognition by means of disjoint principal components models, Pattern Recognit., 8, (1976) 127-139
10) S. Wold, M. Sjöström, SIMCA: a method for analysing chemical data in terms of similarity and analogy. In B. R. Kowalski (Ed.): Chemometrics, Theory and Application. American Chemical Society Symposium Series No. 52, American Chemical Society, Washington, DC (1977) 243-282, ISBN: 0841203792
11) M. P. Derde, D. L. Massart, UNEQ : a disjoint modelling technique for pattern recognition based on normal distribution, Anal. Chim. Acta, 184, (1986) 33-51
12) F. Marini, J. Zupan, A. L. Magrì, Class-modeling using Kohonen artificial neural networks, Anal. Chim. Acta, 544(1-2) (2005) 306-314
13) F. Marini, A. L. Magrì, R. Bucci, Multilayer feed-forward neural networks for class-modeling, Chemometrics Intell. Lab. Syst., 88(1) (2007) 118-124