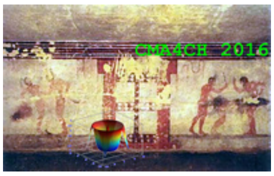
Invicted lecture for
CMA4CH 2016, Mediterraneum Meeting
Straightforward approach in Cultural Heritage and Environment studies - Multivariate Analysis and Chemometry
6th ed., Rome, Italy, Europe, 18-20 December 2016



Multivariate classification methods: A gentle introduction

F. Marini[1]
[1]Dept. of Chemistry, University of Rome “La Sapienza”, P.le Aldo Moro 5, I-00185, Rome, Italy

Abstract

Different analytical chemistry problems aim at predicting one or more qualitative responses on a set of samples (e.g., whether a substance is reactive or not, if a denominated product comes from the declared designated origin or whether a patient is healthy or ill) and therefore may be formally expressed in terms of multivariate classification issues.

In the present lecture, the different chemometric strategies for multivariate classification will be presenting, posing particular attention on the distinction between discriminant and modelling approaches, and with the aid of examples taken from different areas of chemistry.

Introduction

Quite often, in (analytical) chemistry, the multivariate data collected on the samples are used to predict a qualitative characteristics of the sample itself. Examples of such problems involve, for instance the authentication of the origin of a foodstuff, where instrumental fingerprints are collected with the aim of assessing which country or region a particular sample comes from: the different possible origins constitute the discrete values of the sought answer. Another possible case is medical diagnosis or prognosis, where body fluids are collected and analysed in order to verify whether an individual is healthy or ill and, in case, the degree of illness.

From a statistical standpoint, the prediction of one or more discrete answers on a set of individuals falls within the domain of supervised pattern recognition, also called classification [1,2]. Indeed, the aim of classification methods is to build decision strategies which allow to assign an individual (i.e., a sample) to one class (or category) based on its experimental profile: accordingly, a class or category is an abstract collections of objects sharing similar characteristics and corresponds to one of the possible discrete values of the response we want to predict.

In the present communication, the classification approaches most commonly used to solve chemical problems will be presented both from a theoretical and a more application-oriented standpoint. Indeed, the discussion of the characteristics of each method will be accompanied by one or more examples of application to problems pertaining to different chemical fields.

Some relevant concepts

Of all the possible taxonomic differentiations among the available methods presented in the literature, one which is particularly relevant, in the light of the possible applications of classification methods to real world problems, is the one among discriminant and modeling approaches [3].

Discriminant classification methods, as the name suggests, focus their attention on what makes samples from the various categories under study different from one another. In terms of classification rules, this concept translates to the fact that any sample under analysis always assigned to one and only one of the categories represented in the training set. Geometrically, this means that the training samples are used to define the decision boundaries (surfaces) which divide the variable hyperspace in as many regions as the number of classes [4].

On the other hand, there is a different philosophy behind the modeling classification approaches, as they focus their attention on which are the characteristics that are common among samples belonging to the same category [5, 6]. Accordingly, each class is modeled independently on the others and there may exist the case when only a single category is investigated (asymmetric classification). As a consequence, class modeling operates by defining, for each category, multivariate surfaces which enclose a region of space where it is likely to find samples from that particular class: the verification of whether a sample is accepted or rejected by the model of a particular class, then, closely resembles outlier detection. Due to the characteristics reported above, when a sample is analyzed by a class modeling approach, multiple situations can occur: it can be either accepted by a single category, by more than one category ("confused sample"), or by no category at all.

Apart from this fundamental differentiation among the available classification methods, within each of the two families of approaches other further divisions can be made. For instance, the attention can be focused on whether the methods relies explicitly on probabilistic assumptions or not, or on the mathematical complexity of the decision function.

General considerations

Classification problems are ubiquitous in chemistry in general and in analytical chemistry in particular and their solution requires the use of suitable multivariate chemometric classification tools. In this respect, the knowledge of the main characteristics of the available algorithms allows the selection of the approach that can result more suitable to resolve the specific problem under investigation. For instance, due to their nature, class-modelling approaches appear to be ideal to deal with food authentication/traceability, whereas for other situations, discriminant models could result better.

References

1) M. Bevilacqua , R. Nescatelli, R. Bucci, A.D. Magrì, A.L. Magrì, F. Marini, Chemometric Classification techniques as a tool for solving problems in analytical chemistry, *J. AOAC Int.*, 97 (2014), pp. 19-28, https://doi.org/10.5740/jaoacint.SGEBevilacqua.

2) R. Brereton, Chemometrics for pattern recognition, John Wiley and Sons, New York, NY, 2009. ISBN: 978-0-470-98725-4.

3) C. Albano, W. Dunn III, U. Edlund, E. Johansson, B. Nordén, M. Sjöström, S. Wold, Four levels of pattern recognition, *Anal. Chim. Acta*, 103 (1978), pp.429-443. http://dx.doi.org/10.1016/S0003-2670(01)83107-X.

4) G. McLachlan, Discriminant analysis and statistical pattern recognition, Wiley, New York, NY, 2004. ISBN: 978-0-471-69115-0

5) M. Forina, P. Oliveri, S. Lanteri, M. Casale, Class-modeling techniques, classic and new, for old and new problems, *Chemometr. Intell. Lab. Syst.*, 93 (2008) pp.132–148. http://dx.doi.org/10.1016/j.chemolab.2008.05.003

6) P. Oliveri, G. Downey, Multivariate class modeling for the verification of food authenticity claims, *Trends Anal. Chem.*, 35 (2012), pp.74-86. http://dx.doi.org/10.1016/j.trac.2012.02.005.