Analytica Chimica Acta 982 (2017) 9–19

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Tutorial

Class-modelling in food analytical chemistry: Development, sampling, optimisation and validation issues – A tutorial

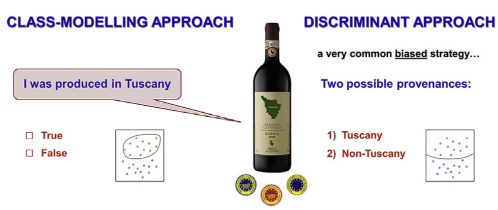
Paolo Oliveri

University of Genova, Department of Pharmacy, Viale Cembrano, 4, I-16148 Genova, Italy

HIGHLIGHTS

- Class-modelling performs verification of compliance by defining multivariate spaces.
- Models built in such a way are free from the distribution of non-target samples.
- Discriminant approaches for one-class problems usually lead to biased solutions.
- Several graphical tools may aid model optimisation and validation stages.
- Rigorous class-modelling should be optimised by considering only sensitivity.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 13 November 2016
Received in revised form
11 May 2017
Accepted 16 May 2017
Available online 29 May 2017

Keywords:
Class-modelling
One-class classification
Discriminant analysis
Food authenticity
Optimisation
Validation

ABSTRACT

Qualitative data modelling is a fundamental branch of pattern recognition, with many applications in analytical chemistry, and embraces two main families: discriminant and class-modelling methods. The first strategy is appropriate when at least two classes are meaningfully defined in the problem under study, while the second strategy is the right choice when the focus is on a single class. For this reason, class-modelling methods are also referred to as one-class classifiers.

Although, in the food analytical field, most of the issues would be properly addressed by class-modelling strategies, the use of such techniques is rather limited and, in many cases, discriminant methods are forcedly used for one-class problems, introducing a bias in the outcomes.

Key aspects related to the development, optimisation and validation of suitable class models for the characterisation of food products are critically analysed and discussed.

Contents

E-mail address: oliveri@difar.unige.it.

http://dx.doi.org/10.1016/j.aca.2017.05.013

1. Introduction

A considerable number of practical cases that require an analytical solution within the food sciences necessitate qualitative answers [1,2]. Typical examples are represented by controls on identity and quality of ingredients or finished products, which may include the verification of fault presence/absence and of agreement with particular claims. These may concern the presence of a particular ingredient, geographical origin, compliance with specific manufacturing rules stated in the product specification, and so on [3,4].

Considering the complexity of such issues and the fact that analytical controls usually provide the assessment of multiple quantities for each of the samples under study, application of multivariate data processing methods is highly profitable [5,6].

In particular, methods that build mathematical rules or models able to characterise a sample with respect to a qualitative property – which can be regarded as the membership to a particular class to be properly defined – are the most appropriate. Two families of multivariate pattern recognition methods satisfy such requirements: discriminant classification and class-modelling [7].

The discriminant approach assigns samples to one among a number of predefined classes (at least two). Instead, class-modelling – also referred to as one-class classification [8] – verifies whether a sample is compatible or not with the characteristics of a single class of interest (or to one single class at a time, in the case of more than one relevant class). These fundamental differences have very important practical implications. For instance, in the discriminant approach, it is fundamental that all of the classes are not only meaningfully defined but also sampled in a fully representative way – a requirement that is hardly fulfilled in many real situations. The typical case is that of verification of compliance with a given specification (*e.g.*, protected designations of origin, geographical indications, quality of ingredients and manufacturing process), which is often addressed as a two-class problem, the two classes being defined as those including compliant and non-compliant samples, respectively. In such cases, while the target class (of compliant samples) can be relevantly defined and sampled, the non-target class (of non-compliant samples) is very often unsuitably defined and poorly sampled [9]. Application of discriminant classification on such malformed data sets is deleterious since it leads to biased classification rules and to similarly biased predictions on new samples. On the contrary, situations like this can be properly addressed by the class-modelling approach, which just needs a representative sample set for the target class to build unbiased verification models.

While one-class classification approaches are commonly used in many fields, from fault detection in industries [10,11] to clinical diagnosis [12,13] and to computer sciences [14,15], their use in chemometrics applied to food sciences is still limited and, in some cases, supplanted by a biased use of discriminant methods [16]. A reason for this is the scarce availability of options for class-modelling – with some exceptions for the SIMCA method – in dedicated chemometric software, which is, in turn, partially ascribable to its scarce usage in the field – a negative chain of factors, indeed.

In the present tutorial, the basic principles of class-modelling are illustrated and critically commented, with a special attention to key aspects of model optimisation and evaluation of the results.

2. Definition of class

A class (or category) is defined as a group of individuals that have one or more properties in common. Usually, these properties can be described by mathematical variables and, therefore, it is possible to state that individuals constituting a class are characterised by the same value of discrete variables, or by similar values (within a defined range) of continuous or pseudo-continuous variables. If such variables that define class membership are easily measurable for every individual, assignation of new individuals to a class is a direct and automatic task. Conversely, if such variables cannot be measured in an easy way, class membership cannot be determined directly. To address this situation, classification methods establish and use mathematical relationships between