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

CLASS-MODELLING APPROACH
I was produced in Tuscany
□ True
□ False
DISCRIMINANT APPROACH
a very common biased strategy...
Two possible provenances:
1) Tuscany
2) Non-Tuscany

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
0003-2670/© 2017 Elsevier B.V. All rights reserved.

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 other variables – which can be easily measured – and class membership. This is – of course – possible if those variables contain useful information and if a number of samples of certain class membership are available to build the classification rule/model.

3. The importance of classification methods in food analytical chemistry

In spite of the widespread tendency to consider any analytical problem as quantitative, analysis on food is often performed to address qualitative issues, the most common of which are identity and quality control tasks.

Typical cases concern identification of raw materials and ingredients, monitoring of ripening [17], investigations on evolution during storage/shelf life [18], verification of authenticity of finished products [19,20], and assessments on the quality [21].

In all of these cases, the answer to the problem of interest can be provided by application of appropriate classification strategies – usually, multivariate – on the analytical data. In particular, when it is possible to define meaningfully and to sample suitably two or more classes, discriminant classification methods may represent a proper solution. The most widespread discriminant methods in chemometric applications are linear discriminant analysis (LDA) [22], quadratic discriminant analysis (QDA) [23], partial least squares discriminant analysis (PLS-DA) [24], and *k*-nearest neighbours (*k*-NN) [25].

Conversely, when the interest is focused on a single target class and the aim is to verify compliance of samples with the features of that class, a class-modelling approach should be adopted. Such methods build an enclosed class space around the class samples. The shape of the class space depends on the particular method applied, while its size is a function of the confidence level that is selected a-priori by the user for the specific case. The principal class-modelling methods used in chemometrics will be described in detail in Section 6.

A key aspect concerns the sampling stage. In fact, functionality of any model, reliability of its validation and its actual applicability strictly depend on the representativeness of the sample sets – a key point whose implications are often underestimated.

4. Differences between class-modelling and discriminant analysis

4.1. One-class and multi-class classification

The first important point to be evaluated when a classification strategy has to be defined is whether the problem under study permits a multi-class or just a one-class choice. In fact, discriminant methods allow to properly address only multi-class situations, while class-modelling can be suitably used to study both one-class and multi-class problems.

Multi-class problems are those in which at least two classes are meaningfully defined – according to the definition given in Section 2 – and can be sampled in a representative way. Examples may include the differentiation between different manufacturing methods (*e.g.*, mechanical *vs*. chemical extraction for vegetable oils [26]), as well as the differentiation between different levels in a process (*e.g.*, different roasting degrees in coffee samples [27]).

One-class problems are instead focused on a single class of interest (the target class), which can be properly defined and sampled, while non-target samples do not constitute a meaningful class and cannot be sampled in a thorough and comprehensive way. A typical example is represented the quality control – where in-specification products define the target class and out-of-specification products constitute a heterogeneous group, which cannot be regarded as an actual class. A very similar situation is that of verification of particular claims (*e.g.*, compliance with the requirements of a protected designation of origin – PDO [28]); in fact, also in this case, the target class can be defined and sampled as required to obtain suitable models, while non-compliant samples usually do not meet the requirements to be considered as a class and cannot be sampled in a representative way [9].

4.2. Ambiguous assignments

When class-modelling methods are applied to multi-class problems, since models for each class are built individually and independently, class spaces may overlap, in the case of classes not completely resolved in the space of the descriptors, as illustrated in Fig. 1 a. Overlapping areas correspond to indecision regions, in which samples are recognised as compatible with models of two or more classes.

Instead, when a pure discriminant method is applied, samples are always unambiguously assigned to a single class, also in the case they are encountered very close to the delimiter (see Fig. 1b), except in the very unlikely case they lie exactly on the delimiter.

Assurance of a null (or quasi-null) rate of ambiguous assignations is one of the reasons for which many users prefer the discriminant strategy instead of the class-modelling one – often perceived as an imperfect strategy, for the same reason.

This represents one of the biggest and most widespread pitfalls in the field of qualitative modelling. In fact, ambiguous assignations potentially represent a very useful outcome, valuable as a diagnostic tool, to indicate that classes are not completely resolved on the basis of the descriptors. Furthermore, indecision regions allow to prevent from wrong (although unambiguous) classifications of samples encountered within overlapping regions.

Considering these actual advantages, discriminant methods can be properly modified to define an indecision region about the delimiter [29], as exemplified in Fig. 1 c.

Unfortunately, these options are rarely implemented in commercial software and, therefore, their applications are quite rarely encountered.

4.3. Outlier detection

In the class-modelling approach, samples that fall outside model boundaries are considered as non-compatible with the class of interest. Also in the case of a multi-class class-modelling, samples may fall outside every class space – a common occurrence for samples pertaining to classes not considered in the study and for anomalous or atypical samples. In other words, capability for outlier detection is an intrinsic feature of class-modelling methods, as it is evident in Fig. 1 a.

On the contrary, the standard discriminant approach always assign a sample to one of the classes pre-defined in the study, according to the assignation rule based on the delimiter (Fig. 1b).

Actually, modified discriminant strategies have been proposed, which define a maximum permitted distance from the delimiter and, on this basis, exclude samples very far from class centroids (*i.e.*, potential outliers, as exemplified in Fig. 1c) [29]. Regrettably, implementations and applications of such criteria, although profitable, are quite limited, similarly to modifications concerning indecision regions about the delimiter, described in the previous section.

4.4. Influence of non-target samples

One of the core features of the class-modelling approach is the

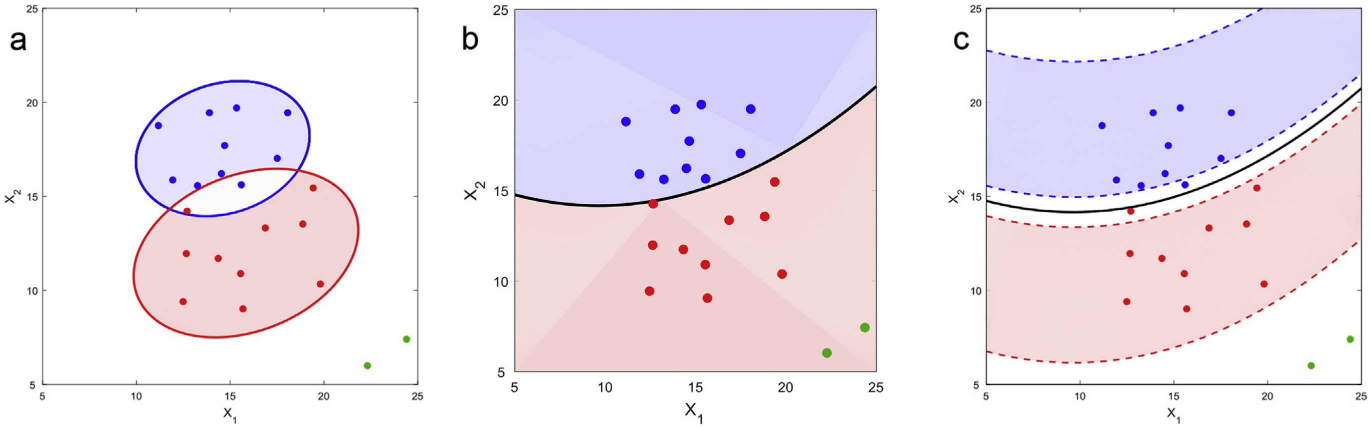

Fig. 1. Bivariate data set: samples belonging to two classes (red and blue circles) plus two outliers (green circles). (a) Class-modelling approach; ellipses = class spaces; white intersection area = indecision region. (b) Pure discriminant approach; black line = delimiter. (c) Modified discriminant approach; black solid line = delimiter; dotted lines = acceptance limits. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

focus on a single class, or on a single class at a time, in the case of multi-class problems. This means that each model is built using solely and exclusively data from the related class, without any influence from other classes. A very important practical consequence of this feature is that a model for a given class can be built also in the case in which no samples from alternative classes have been collected and analysed and, therefore, only data from the target class are available. The same possibility is not allowed by discriminant methods that require, by definition, data from at least two classes to define a delimiter – which is, by definition, influenced by data of both of the classes.

Exclusive availability of data from the sole target class represents, of course, a limit situation. The problem much more frequently concerns representativeness of samples within alternative classes – and the consequences are much more deceitful.

In fact, when a target class (*e.g.*, a food product labelled with a given authenticity claim) has to be characterised, a common situation is the disposal of a representative sample set for the target class, plus a number of incomplete and poorly representative sample sets for a number of alternative classes. In such a practical occurrence, the modelling approach would provide an unbiased characterisation of the target class, while the discriminant approach would lead to a biased classification rule, due to the incorporation of incomplete information from the alternative classes.

To better illustrate this key issue, the effect of slight variations in the sample set composition of a non-target class is graphically exemplified in Fig. 2, with bivariate data. As it can be noticed, such variations do not affect at all the shape and the size of the class boundary (class-modelling approach), while a considerable influence is observable on the inter-class delimiter (discriminant approach).

4.5. The misuse of class-modelling methods as discriminant tools

From examination of one hundred representative original research papers published in scientific journals over the last ten years in the field of food authentication (Fig. 3), it emerges that in more than one half of the studies, discriminant methods are used, while the class-modelling approach is followed only in a reduced fraction of researches.

An in-depth analysis of the papers that claim application of class-modelling methods reveals, in a relevant number of cases, an anomaly that can be quite easily detected examining the results: the class-modelling method has been modified and forced to perform a discriminant classification.

The most common way to accomplish such an unnatural action consists in building class models for all of the classes of interest and, subsequently, using distances from a given sample to each class model as a criterion to perform a discriminant classification: the sample is assigned to the class for which such a distance is minimum. All of the points equally distant from two class models in the space of the descriptors define the delimiter between this pair of classes, as in the example of Fig. 2.

The main reason that may lead to prefer such a modification is conceivably the apparent advantage of minimising ambiguous assignments, which are a typical occurrence for class-modelling. This actually means the loss of a core feature of the modelling approach (see Section 4.2).

Furthermore, the final classification outcomes are influenced by information from all of the classes, like in any discriminant strategy. Consequently, the main advantage of class-modelling, described in the previous section, is definitely lost.

5. Authenticity verification as a two-class problem: a widespread biased approach

In most of the cases, verification of authenticity of a food product consists in assessing the truth of a given claim (*e.g.*, a specific geographical provenance). If the claim is positively verified, the product is considered as authentic. Conversely, a fraud can be suspected.

In such situations, the focus is on a single target class. For this reason, performing one-class classification by class-modelling methods represents the most appropriate strategy, as already stated. Following this approach, a class model is built using information from the target class, while data from non-target samples may be used, in case, for evaluating model performances.

An alternative – and biased – strategy, which is encountered frequently in published research studies, consists in converting the one-class problem into a two-class problem. In more detail, a second class is defined, besides the target class, as the class of all the samples that do not comply with the authenticity claim to be verified.

To take a practical example, if the classification study was aimed at verifying authenticity of a PDO olive oil produced within a particular geographical region, the discriminant approach would require the collection of training samples for two classes: the target

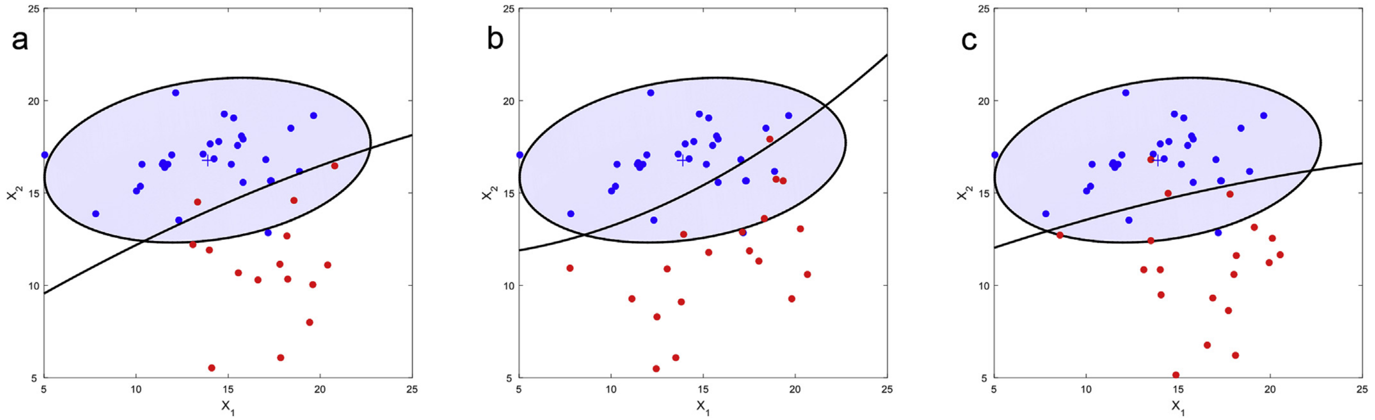

Fig. 2. Bivariate data set: target samples (blue circles) and non-target samples (red circles). Black lines = delimiters (discriminant approach); ellipses = class spaces of the target class (modelling approach). It can be noticed that small differences in the composition of the non-target sample set may determine considerable variations in the features of the discriminant delimiter, while shape and size of the class space are unaltered. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

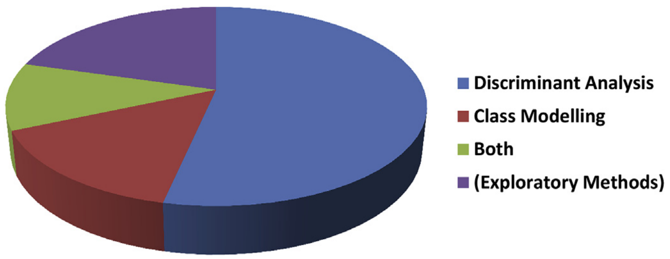

Fig. 3. Pattern recognition tools applied in one hundred representative original research papers in the field of food authenticity verification over the last ten years. Data obtained from a literature search using Scopus, with keywords: “food authenticity” and “chemometrics”.

class, including samples of the PDO oil to be studied, and the non-target, including samples of all of the olive oils produced worldwide outside the geographical region under study. It is quite obvious that collecting a representative set of non-target samples is a rarely realisable task. Indeed, non-target samples do not meet any requirement to be considered as a class, considering the definition of class given in Section 2. As a result, sets of non-target samples are often under-representative, leading inevitably to biased decision rules (as it has been demonstrated in Section 4.4).

6. Main class-modelling methods in chemometrics

6.1. UNEQ

The unequal class models — also referred to as unequal dispersed classes — (UNEQ) modelling method is a based on a parametric probabilistic strategy introduced by Derde and Massart in 1986 [30,31], and closely related to Harold Hotelling's multivariate approaches for quality control [32]. The class of interest is described by an elliptical space built around the barycentre of training data points of the class, namely the centroid vector. In more detail, the UNEQ space for a given class c is defined by Hotelling's T^2 multivariate probability distribution (with v variables), in which location and dispersion are respectively estimated by the centroid vector ($\mathbf{x}_c$) and by the variance-covariance matrix ($\mathbf{V}$) derived from the frequency distribution of training samples within class c:

$$f(\mathbf{x}|c) = \frac{1}{\sqrt{(2\pi)^v|\mathbf{V}_c|}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{x}_c)'\mathbf{V}_c^{-1}(\mathbf{x}-\mathbf{x}_c)} \tag{1}$$

The exponential operator in Eq. (1) (a squared Mahalanobis distance) defines boundaries of the class space as an iso-probability contour ellipse (bivariate case), ellipsoid (trivariate case) or hyper-ellipsoid (multivariate case) [33]. Orientation and eccentricity of the elliptical class space respectively derive from correlation between the variables and their dispersion, accounted for by **V**. The width of the class space is determined according to the critical value of Hotelling's T^2 statistics at a pre-determined confidence level.

An improved version of UNEQ makes use of T^2 statistics for modelling the class space boundaries only of the evaluation sample set, while Beta statistics is used for the training samples [34]. Beta statistics generate a tighter class space. The difference between the width of the two boundaries increases when the number of samples decreases.

6.2. SIMCA

Soft independent modelling of class analogy (SIMCA) is a non-probabilistic distance-based modelling method introduced by Svante Wold [35]. SIMCA models are based on principal components (PCs), which are, by definition, the directions of maximum variance (and, therefore, of maximum information) in a multivariate data space [36]. As indicated by the acronym, PCs are computed independently for each of the classes of interest. To this aim, data are initially transformed by a class-based column autoscaling or column mean centring, either of which shifts the origin of the reference axes to coincide with the class centroid. Then, PCA is performed, with a rotation about the class centroid, and the number of significant PCs is evaluated, usually by means of a double cross-validation procedure [37]. The significant PCs define the so-called SIMCA inner space.

Training samples of the class to be modelled are therefore projected on the significant inner-space PCs, obtaining score values for each samples on each PC. In the original version of SIMCA, the ranges of such PC scores define the class model (normal range model). Such a model has the shape of a segment (one-dimensional inner space), a rectangle (bidimensional inner space), a parallelepiped or hyper-parallelepiped (three or multidimensional inner space), given that PCs are orthogonal by definition [36].

Residuals (namely, the distances between each sample and the model the space defined by the non-significant PCs, called SIMCA outer space) are then computed and used to define a distance from class model (*OD*), in combination with the distance in the score inner space (*ID*), to define the so-called SIMCA augmented distance from sample s to class C ($d_{s,C}$):

$$d_{s,C} = \sqrt{ID_{s,C}^2 + OD_{s,C}^2} \tag{2}$$

The critical value of this distance, which determine acceptation/rejection of a new sample by the model, is defined by the critical value of Fisher statistics at a pre-determined confidence level, given that residuals are assumed to follow a multivariate normal distribution.

Also in this case, it is possible to define two different metrics, for the training and the test samples respectively.

Many versions of the SIMCA algorithm have been proposed, making it a very flexible method. The most important modifications concern the definition of the PC-score based model. For instance, the normal range model can be enlarged so as to avoid the possibility of an under-estimation of the true variability (if few training samples are available), or reduced, to avoid the possibility of an over-estimation (if many training samples are available) [38]. Furthermore, several modifications can be introduced in the computation of SIMCA distance, the most common of which involves calculation of the contribution in the score inner space as a Mahalanobis distance [39].

Data driven SIMCA (DD-SIMCA) [40,41] approximates distribution of both *ID* and *OD* by a scaled chi-squared distribution, whose parameters (scaling factors and degrees of freedom) are estimated using a data-driven method [42]. DD-SIMCA is able to calculate misclassification errors theoretically.

A fuzzy version of SIMCA – referred to as fuzzy grid encoded independent modelling for class analogies (FIMCA) – has been also recently presented [43].

Residuals for a given sample i ($\mathbf{e}_i$) – namely, the vector containing the fraction of information not explained by the components retained in the model – can be studied by Q statistics:

$$Q_i = \mathbf{e}_i\mathbf{e}_i^T \tag{3}$$

whose confidence limit, Q_α, is computed according to Jackson [44]:

$$Q_\alpha = \theta_1\left[\frac{z_\alpha\sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0(1-h_0)}{\theta_1^2}\right]^{\frac{1}{h_0}} \tag{4}$$

where z_α is the value of the standard normal deviate corresponding to the upper (1-α) percentile, and θ_j terms and h_0 are defined as:

$$\theta_j = \sum_{l=L+1}^{\min(I,V)} \lambda_l^j \text{for } j = 1,2,3 \tag{5}$$

and

$$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2} \tag{6}$$

The θ_j terms are the sums of the eigenvalues (λ) raised to the jth power for the components that have not been retained in the model. It has to be remarked that Eq. (4) assumes the residuals to be normally distributed, a condition that is generally verified [44].

6.3. PFM

Potential function methods (PFM) constitute a family of non-parametric probabilistic techniques derived from the work of Coomans and Broeckaert [45]. The first class-modelling version is due to Forina and co-workers [46].

PFM estimate a probability density distribution of a class of interest as a sum of contributions from each single sample of the class in a training set. A variety of functions can be used to define these individual contributions. The name of the techniques reflects the fact that, in the original implementations, functions analogous to the electric potential were used. In the most recent implementations, functions most commonly used are Gaussian-like with a smoothing coefficient that is formally analogous to the standard deviation of the Gaussian probability function and, therefore, concurs in determining the shape of the distribution. Such a coefficient can be the same for all the samples of a given class (fixed potential strategy) or it may be varied as a function of the local density of samples (variable potential strategy). This latter approach is useful especially when the underlying multivariate distribution is very irregular, with regions characterised by non-uniform density of samples. The value of the smoothing coefficient can be optimised by means of a leave-one-out cycle on an optimisation sample set. The resulting estimated overall probability distribution can be sectioned at different confidence levels, with iso-probability contours representing the boundaries of the class space at each confidence level. Two strategies have been proposed for determination of the space boundaries corresponding to a given probability level: the *p*% sample percentile and the equivalent determinant. It has been demonstrated that the latter method is less sensitive to the presence of outliers [46]. Boundaries of the class spaces can be very complex, capable of effectively describing non-normal and non-uniform sample distributions.

6.4. SVM

Support vector machines (SVM) are a family of pattern-recognition methods conceived for efficiently dealing with non-linear data distributions.

The basic feature of SVM is the projection of data points into a space with augmented dimensions, functional to individuate simple (possibly linear) functions able to model the data. Such modelling functions can be projected back into the space of the original predictors, resulting in lower-dimensional but with higher complexity (usually non-linear) functions.

SVM are traditionally used for discriminant classification [47]. Nevertheless, some Authors presented modifications functional to class-modelling. Among the most common approaches, it is worth mentioning the support vector domain description (SVDD) method by Tax and Duin [48], which makes use of hyperspheres to define the class spaces. An alternative approach, proposed by Schölkopf and co-workers [49] and claimed to provide one-class models, makes use of hyperplanes.

6.5. PLS-based methods

One of the discriminant techniques most widely applied is the so-called partial least squares discriminant analysis (PLS-DA) or discriminant partial least squares (D-PLS). The method, introduced by Barker and Rayens [24], provides a linear delimiter applying partial least squares (PLS) regression [50] using binary class membership indices (*e.g.*, 0 and 1) for each class as the response variables. When more than two classes are involved, the PLS-2 algorithm is applied which allows the prediction of a matrix of response variables, that is, one for each class. PLS-DA is often used as an alternative to LDA for data sets in which the number of variables is larger than the number of samples. Nonetheless, it can be demonstrated that, when the number of variables considerably exceeds the number of objects, PLS-DA is generally able to find a delimiter that discriminates between two classes, even though such classes are not separated in reality [51]. Therefore, in these cases, a thorough model validation becomes fundamental.

In the recent years, a number of attempts have been addressed to develop class-modelling techniques exploiting the advantages offered by the PLS method.

In particular, a method called one-class PLS (OC-PLS) has been recently presented, in which a PLS model is built using a constant response ($y = 1$), *i.e.*, identical values for all of the training samples belonging to the class of interest [52]. Hotelling's T^2 and Q statistics are used to verify compliance of test samples with the class model, providing a solution that is, in many cases, analogues to that of the SIMCA method.

An alternative method, called PLS density modelling (PLS-DM) [53], develops a PLS model using a density vector as the y response vector, computed – for each sample of the training set – as the sum of the k smallest inter-sample Euclidean distances in the multivariate space. Parameter k influences the smoothness of density function, which evolves from a sharper to a smoother shape while increasing k. The PLS scores on the first L latent variables selected are used as an input to estimate probability density of the class by a potential function method (PFM). Class boundaries are defined according to the critical value (f_α) obtained from the critical value of the chi-squared distribution by the so-called equivalent determinant method [46]. In addition, PLS residuals are used to compute the critical value of Q statistics (Q_α) at the same level of α according to the Jackson-Mudholkar approximation [44]. In this way, compliance of each object with the class model is granted when it complies with both the f_α and Q_α criteria.

7. Evaluation of class-modelling performances

7.1. Validation strategies and model interpretation

Practical usefulness of a class model is strictly related to its reliability in prediction. Model validation, namely the estimation of predictive ability on new samples – not used for building the model – is therefore a key point. Usually, validation strategies divide the available samples into two subsets: a training (or calibration) set used for building the model and a test (or evaluation) set used to assess its validity. Both of the sets must contain samples of known class membership. Furthermore, a reliable validation requires that no information from samples in the test set is used for building the model, so as to avoid overestimations of the prediction ability.

Evaluation of the predictive ability of a model can be performed onto either a single test set – one-step procedure – or different evaluation sets, following an iterative procedure. When a single test set is used, a fraction – usually between 50% and 10% – of the available samples is selected to constitute the test set, with the remaining objects forming the training set. The subdivision may be arbitrary, based on a random choice, or even performed by way uniform sampling designs, such as the Kennard and Stone algorithm and its modifications [54,55], which generate two sample subsets that explore the whole variability domain and are uniformly distributed within it. Cross-validation (CV) is one of the most common choices among the iterative validation procedures. It splits the N rows of the data matrix (samples) into C cancellation groups, following a predetermined scheme (the most common of which are contiguous blocks and Venetian blinds). The model is computed C times, each time using one of the cancellation groups as the test set, and the remaining samples as the training set. C usually ranges from 3 to N – an extreme and, usually, over-optimistic case generally known as the leave-one-out procedure (LOO).

Bootstrap validation is an extensive iterative strategy, in which a high number of models (often more than 1000) are computed, each time randomly extracting (with repetition) different test sets of variable size.

The possibility of interpreting the outcomes of a class-modelling procedure – in terms of assessment of the role of the original variables in class assignment – is fundamental, as it provides the basis for support and endorse the results, and it can be therefore considered as a further implicit validation of the model. Model interpretation is a simple task for the simplest methods (e.g., in the case of UNEQ models on the original variables) and becomes more difficult for more complex methods. More complex methods are often used as black boxes, which do not provide any direct interpretation of the predictions. In such cases, a thorough validation of the results is a key step to avoid overfitting and blunders in predictions on real samples.

7.2. Evaluation parameters

When evaluating the outcomes of class-modelling by means of a sample test set, samples belonging to the class of interest are designated as true positive (*TP*) if they are correctly recognised as compliant by the model, and false negative (*FN*) if they are erroneously rejected. Correspondingly, samples not belonging to the class of interest are labelled as false positive (*FP*) if they are erroneously assigned to the class, and true negative (*TN*) if they are correctly refused.

Fractions of true positive and true negative assignations define two important evaluation parameters: sensitivity and specificity, respectively.

Sensitivity is defined as the fraction of samples of the class of interest which resulted in a true positive assignation:

$$sensitivity = \frac{TP}{TP + FN} \tag{7}$$

It represents an experimental measure of the confidence level of the class space.

Conversely, specificity is that fraction of samples extraneous to the modelled class which is correctly refused by the model:

$$specificity = \frac{TN}{TN + FP} \tag{8}$$

Sensitivity can be evaluated both in fitting (on the samples of the training set of the class of interest) and in prediction (on new samples belonging to the class of interest, but not used in model building). Conversely, specificity cannot be assessed in fitting, since samples not belonging to the class of interest (necessary to evaluate such a parameter) are never used in model definition in none of the class-modelling strategies.

A comprehensive parameter, referred to as efficiency, is defined as the geometric mean of sensitivity and specificity values:

$$efficiency = \sqrt{\frac{TP \cdot TN}{(TP + FN) \cdot (TN + FP)}} \tag{9}$$

It may vary between 0, when either sensitivity or specificity are null, and 1 (the ideal case), when both parameters have the maximum value of 1.

In some applications (mainly in the field of botanical authentications) probability of identification (POI) curves are used – a graphical tool that reports the probability of true positives (a measure of sensitivity) as a function of the value of a given parameter (typically, a concentration) [56,57].

7.3. ROC curves

Receiver operating characteristic (ROC) curves are a graphical tool classically used for evaluating the performances of a given discriminant classifier (generally univariate) in a two-class problem.

The name is related with their first application, which was evaluation of the ability of radar operators in identifying hostile aircrafts, during World War II. Afterwards, ROC curves became popular especially in the biomedical field, both in clinical and forensic investigations [58,59] – one of their principal applications being the evaluation of efficiency of a given parameter (or biomarker) to differentiate between healthy and ill individuals.

ROC curves are built varying the position of the delimiter, which corresponds to the decisional threshold of the measured parameter in the univariate case. True positive rate (TPR) and false positive rate (FPR) are computed at each step and graphically represented in a two-axis Cartesian plot, in which the horizontal axis usually reports FPR, while the vertical axis reports TPR. Experimental outcomes are connected by a line that constitutes the ROC curve.

A detailed analysis of ROC curves provides not only the choice of the most appropriate threshold value, as the best compromise between FPR and TPR (whose evaluation obviously depends on the specific case), but also important information about the system under study.

In particular, curves that tend to the diagonal bisector of the plot indicate very poor classifiers, which basically perform like a random class assignation. Conversely, curves that tend to detach from the diagonal bisector towards the upper left corner of the plot are associated to efficient classifiers. A measure of this is often evaluated by quantifying the area under the curve (AOC), which increases (up to the maximum of 1) while increasing the performance of the classifier under evaluation.

A modified version of ROC curves was proposed for evaluating the performance of class-modelling approaches [60]. In this case, curves are built varying the confidence level at which class-modelling is performed, resulting in a variation of the size of the class space. Sensitivity and specificity values, computed at each step, are reported in the plot (100 %-specificity % on the abscissa, and sensitivity % on the ordinate). The way of interpreting such a plot is analogous to that obtained for the discriminant case.

The best compromising between sensitivity and specificity is usually chosen taking into account the final purpose of the specific investigation, as well as costs and risks associated to an incorrect identification of positive and negative events.

Fig. 4 reports an example of comparison between different models, suggesting how ROC curves can be used as a powerful tool for comparing the performances of models obtained either with different parameter settings, or by application of different modelling methods.

A video illustrating the building stages of a classical ROC curve – for the univariate two-class case – and an extended version for class-modelling is available in the Electronic Supporting Material (S1).

7.4. Distance diagrams

A useful tool for a graphical evaluation of class-modelling results, when at least two classes are modelled, is represented by the so-called Coomans' plots [61], which will be illustrated with a practical example – a data set contains FT-NIR spectra recorded (over the spectral range 9000–4200 cm^{-1}, at 4 cm^{-1} resolution) on ground table olives belonging to three different cultivars: *Taggiasca*, *Leccino* and *Coquillo* [62]. Spectral data have been previously submitted to SNV transform and column autoscaled, in order to eliminate unwanted signal variations and, subsequently, submitted to class-modelling with the SIMCA method (5 PCs for the inner space).

In the Coomans' plot represented in Fig. 5 a, the two axes

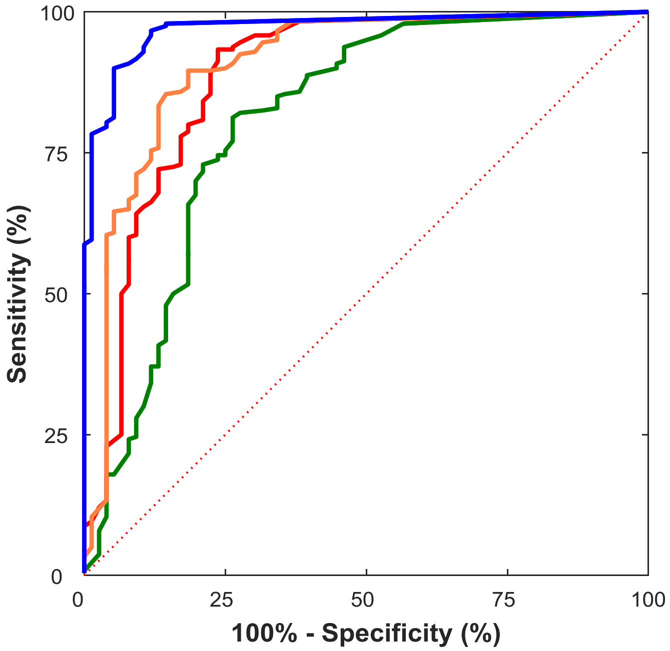

Fig. 4. Example of ROC curves (solid lines) for the evaluation of different class-modelling outcomes. Efficiency of models associated to the curves decreases from blue to orange to red to green. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

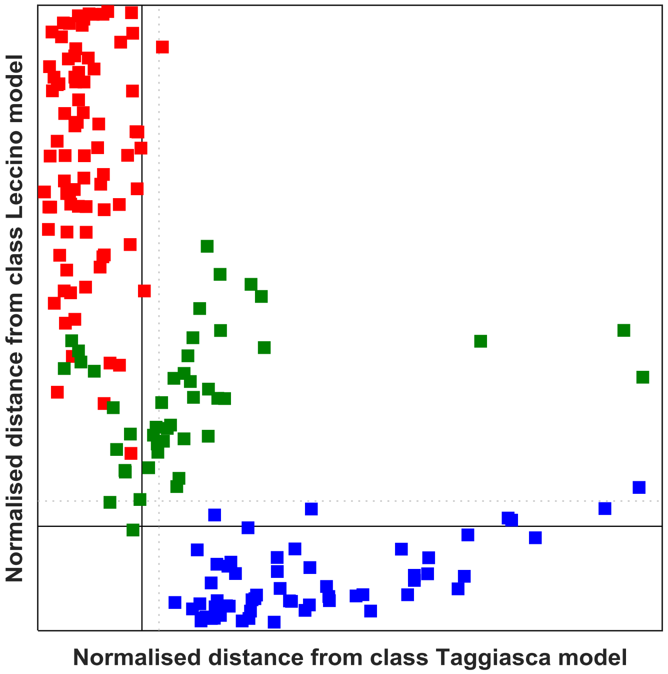

Fig. 5. Example of Coomans' plot for the table olive data set (red squares = *Taggiasca* olives; blue squares = *Leccino* olives; green squares = *Coquillo* olives). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

correspond to the distances of samples from the models of the classes *Taggiasca* and *Leccino*, respectively. Samples of the three classes are represented as scatter points, whose coordinates indicate the relative similarity with the two class models.

The two straight lines parallel to the axes correspond to the critical acceptance/rejection levels for *Taggiasca* and *Leccino* class models, respectively, at the pre-selected confidence level (95%).

The plot area is therefore divided into four sectors, which contain respectively:

- samples accepted by the class *Taggiasca* model (upper left rectangle);
- samples accepted by the class *Leccino* model (lower right rectangle);
- samples accepted by both of the models (lower left square);
- samples rejected by both of the models (upper right square).

For methods whose assignation rule is based on both T^2 and Q values, an informative plot is often obtained by representing T^2 and Q distances from the model of a class. Such a plot is helpful in understanding the nature of outliers, similarly to what is done in the fields of multivariate quality control and multivariate process monitoring.

8. Optimisation of class models

8.1. Validation issues

A class model is useful when it is able to provide reliable predictions on new samples. Predictive ability should be estimated on a set of samples not used for building the models. To this aim, several procedures have been proposed, the most common of which divide the whole set of samples of certain class membership into two subsets: a training (or calibration) set, used for developing the model, and an evaluation (or test) set, used to assess its reliability. Model validation is consistent if samples in the evaluation set have not influenced the model neither in the building nor in the optimisation steps; if such requirements are not met, the prediction ability may be overestimated.

In particular, when some factors (including pre-processing, feature selection and parameters specific of a given modelling method) are optimised in the search for a setting that provides the maximum modelling performance, the risk of overfitting is considerable. Overfitting means that the model fits excessively a given sample subset, using also a considerable fraction of the irrelevant information embodied in the analytical data (*e.g.*, random noise and unwanted sources of variations). This usually leads to poor performances in the prediction in real applications on new samples.

When an optimisation of parameters is performed, a recommendable strategy is to use three sample subsets: a training set, an optimisation set and an evaluation set. The optimisation set is used to find the optimal settings of the relevant parameters, while the actual reliability of the final model is estimated by way of prediction on the third subset, formed by objects that have never influenced either the model or its optimisation.

8.2. Pareto charts

Pareto charts [63] are a graphical tool that can be effectively applied to compare sensitivities and specificities of class models obtained under different conditions, allowing to identify optimal models evaluating the most profitable balance — which, of course, depends on the particular problem under study — between the evaluation parameters. In fact, being a multicriteria decision strategy, Pareto optimisation considers more than one objective simultaneously, looking for the optimal compromise.

In more detail, Pareto charts for evaluation of class-modelling outcomes are bidimensional Cartesian diagrams, whose axes represent sensitivity and specificity, respectively. Each class model to be compared is therefore represented by a scatter point within the graph. Points are considered Pareto efficient (non-dominated point) if none of the other solutions is better both for sensitivity and specificity simultaneously. The non-dominated points can be connected by a piece-wise line, called Pareto front (or Pareto frontier). The ideal solution can be identified, among the potentially optimal solutions constituting the Pareto front, thoroughly considering the practical implications related to the case studied: in fact, in some situations, it might be preferable just to maximise efficiency while, in other cases, it might be profitable to choose slightly unbalanced solutions that favour a higher sensitivity (or specificity). For instance, a protection consortium of a given food product, might prefer to have models that exclude the lowest number of affiliated producers, that means to select models with the highest sensitivity.

Fig. 6 shows an illustrative Pareto diagram. The black line that connects the set of non-dominated points is the Pareto front. Each point corresponds to a model obtained under a particular setting of the tuneable parameters and can be coded by a colour scale that indicates the levels of a given parameter. In such a way, it is possible to understand the effect of relevant parameters (which can be varied in the optimisation stages according, for instance, to a full factorial design) on the outcomes.

8.3. Compliant and rigorous approaches

A fundamental aspect to suitably implement a class-modelling strategy concerns the evaluation parameters considered in the model optimisation stages.

In many cases, the optimal conditions for a given method/model are looked for by considering both sensitivity and specificity values, like in the Pareto approach described in Section 8.1, or evaluating

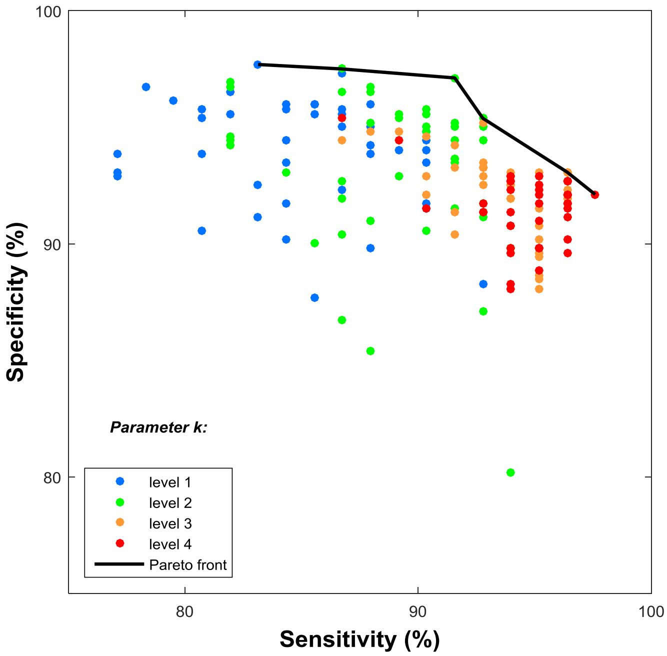

Fig. 6. Example of Pareto diagram. Solid black line = Pareto front connecting optimal solutions. Each point represents a model obtained under different settings of relevant parameters. Colours can be used to code the different levels of a given parameter.

efficiency as a comprehensive parameter. In all of these cases – which probably represent the most common situation – it should be considered that information from non-target samples is influencing the choice of some features of the class model. Such an optimisation strategy – which has been recently defined as compliant approach [64] – may, of course, lead to models characterised by higher efficiency, but at the cost of introducing a potential bias on the model.

The alternative way to optimise a model without introducing such a bias – defined as rigorous approach [64] – is to consider just sensitivity, a parameter that only depends on samples from the target class. In more detail, considering that sensitivity is an experimental estimate of the confidence level that has been set for a given model, following the rigorous approach, models whose sensitivity is closest to the confidence level should be considered as optimal.

9. Conclusions

Class-modelling performs verification of compliance with a specification by defining a multivariate enclosed class space, at a predetermined confidence level, for authentic samples of the class under investigation. Models built in such a way has the advantages of describing the target samples being free from the distribution of non-target samples in the training set.

Conversely, discriminant methods look for a delimiter between two – or more – classes, using a contribution from all of the classes considered. This means that all of the classes must be correctly defined and that samples included must be thoroughly representative of each class since they have a crucial influence on the decision rule to be derived. This is extremely important when the focus is on a single class like, for example, cases involving verification of a food authenticity claim. Several graphical tools may aid model optimisation and validation stages. Nevertheless, it should be reminded that a rigorous class-modelling approach should evaluate only sensitivity when making decisions about the optimal conditions of relevant parameters.

Acknowledgment

Financial support by the Italian Ministry of Education, Universities and Research (MIUR) is acknowledged – Research Project SIR 2014 "Advanced strategies in near infrared spectroscopy and multivariate data analysis for food safety and authentication", RBSI14CJHJ (CUP: D32I15000150008).

Appendix. ASupplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.aca.2017.05.013.

References

[1] M. Valcárcel, S. Cárdenas, Vanguard-rearguard analytical strategies, Trac. Trends Anal. Chem. 24 (2005) 67–74, http://dx.doi.org/10.1016/j.trac.2004.07.016.

[2] Á. Ríos, M. Zougagh, Modern qualitative analysis by miniaturized and microfluidic systems, Trac. Trends Anal. Chem. 69 (2015) 105–113, http://dx.doi.org/10.1016/j.trac.2015.04.003.

[3] G. Guthausen, Analysis of food and emulsions, Trac. Trends Anal. Chem. 83 (2016) 103–106, http://dx.doi.org/10.1016/j.trac.2016.02.011.

[4] G.P. Danezis, A.S. Tsagkaris, F. Camin, V. Brusic, C.A. Georgiou, Food authentication: techniques, trends & emerging approaches, Trac. Trends Anal. Chem. (2016), http://dx.doi.org/10.1016/j.trac.2016.02.026.

[5] Y. Picó, P. Oliveri, M. Forina, Chapter 2 – data analysis and chemometrics, in: Chem. Anal. Food Tech. Appl, 2012, pp. 25–57, http://dx.doi.org/10.1016/B978-0-12-384862-8.00002-9.

[6] B.P. Geurts, J. Engel, B. Rafii, L. Blanchet, A. Suppers, E. Szymańska, J.J. Jansen, L.M.C. Buydens, Improving high-dimensional data fusion by exploiting the multivariate advantage, Chemom. Intell. Lab. Syst. 156 (2016) 231–240, http://dx.doi.org/10.1016/j.chemolab.2016.05.010.

[7] P. Oliveri, G. Downey, Multivariate class modeling for the verification of food-authenticity claims, Trac. - Trends Anal. Chem. 35 (2012) 74–86.

[8] R.G. Brereton, One-class classifiers, J. Chemom. 25 (2011) 225–246, http://dx.doi.org/10.1002/cem.1397.

[9] S.S. Khan, M.G. Madden, One-class classification: taxonomy of study and review of techniques, Knowl. Eng. Rev. 29 (2014) 345–374, http://dx.doi.org/10.1017/S026988891300043X.

[10] D. Martínez-Rego, O. Fontenla-Romero, A. Alonso-Betanzos, J.C. Principe, Fault Detection via Recurrence Time Statistics and One-class Classification, 2016, http://dx.doi.org/10.1016/j.patrec.2016.07.019.

[11] X. Gao, R. Ma, Fault detection of batch process based on MSICA-OCSVM, in: 2016 Chinese Control Decis. Conf, IEEE, 2016, pp. 3461–3465, http://dx.doi.org/10.1109/CCDC.2016.7531581.

[12] I. Irigoien, B. Sierra, C. Arenas, I. Irigoien, B. Sierra, C. Arenas, C. Arenas, Towards application of one-class classification methods to medical data, ScientificWorldJournal 2014 (2014) 730712, http://dx.doi.org/10.1155/2014/730712.

[13] A. Retico, I. Gori, A. Giuliano, F. Muratori, S. Calderoni, One-class support vector machines identify the language and default mode regions as common patterns of structural alterations in young children with autism spectrum disorders, Front. Neurosci. 10 (2016) 306, http://dx.doi.org/10.3389/fnins.2016.00306.

[14] Q. Miao, J. Liu, Y. Cao, J. Song, Malware detection using bilayer behavior abstraction and improved one-class support vector machines, Int. J. Inf. Secur 15 (2016) 361–379, http://dx.doi.org/10.1007/s10207-015-0297-6.

[15] S. Agarwal, A. Sureka, Using KNN and SVM Based One-class Classifier for Detecting Online Radicalization on Twitter, Springer International Publishing, 2015, pp. 431–442, http://dx.doi.org/10.1007/978-3-319-14977-6_47.

[16] O.Y. Rodionova, A.V. Titova, A.L. Pomerantsev, Discriminant analysis is an inappropriate method of authentication, Trac. Trends Anal. Chem. 78 (2016) 17–22, http://dx.doi.org/10.1016/j.trac.2016.01.010.

[17] V. Giovenzana, R. Beghi, C. Malegori, R. Civelli, R. Guidetti, Wavelength selection with a view to a simplified handheld optical system to estimate grape ripeness, Am. J. Enol. Vitic. 65 (2014).

[18] R. Beghi, G. Giovanelli, C. Malegori, V. Giovenzana, R. Guidetti, Testing of a VIS-NIR system for the monitoring of long-term apple storage, Food Bioprocess Technol. 7 (2014) 2134–2143, http://dx.doi.org/10.1007/s11947-014-1294-x.

[19] P. Oliveri, V. Di Egidio, T. Woodcock, G. Downey, Application of class-modelling techniques to near infrared data for food authentication purposes, Food Chem. 125 (2011) 1450–1456.

[20] C.V. Di Anibal, S. Rodríguez, L. Albertengo, M.S. Rodríguez, Uv-visible spectroscopy and multivariate classification as a screening tool for determining the adulteration of sauces, Food Anal. Methods 9 (2016) 3117–3124, http://dx.doi.org/10.1007/s12161-016-0485-7.

[21] F.S. Uslu, H. Binol, A. Bal, Food inspection using hyperspectral imaging and SVDD, in: M.S. Kim, K. Chao, B.A. Chin (Eds.), International Society for Optics and Photonics, 2016, p. 98640N, http://dx.doi.org/10.1117/12.2223938.

[22] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (1936) 179–188, http://dx.doi.org/10.1111/j.1469-1809.1936.tb02137.x.

[23] S. Geisser, Posterior odds for multivariate normal classifications, J. R. Stat. Soc. Ser. B 26 (1964) 69–76. https://www.jstor.org/stable/2984606?seq=1#page_scan_tab_contents.

[24] M. Barker, W. Rayens, Partial least squares for discrimination, J. Chemom. 17 (2003) 166–173, http://dx.doi.org/10.1002/cem.785.

[25] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theory 13 (1967) 21–27, http://dx.doi.org/10.1109/TIT.1967.1053964.

[26] C.E. Gumus, A. Yorulmaz, A. Tekin, Differentiation of mechanically and chemically extracted hazelnut oils based on their sterol and wax profiles, J. Am. Oil Chem. Soc. (2016) 1–11, http://dx.doi.org/10.1007/s11746-016-2882-x.

[27] P.R.A.B. de Toledo, M.M.R. de Melo, H.R. Pezza, L. Pezza, A.T. Toci, C.M. Silva, Reliable discriminant analysis tool for controlling the roast degree of coffee samples through chemical markers approach, Eur. Food Res. Technol. (2016) 1–8, http://dx.doi.org/10.1007/s00217-016-2790-1.

[28] P. Oliveri, M. Casale, M.C. Casolino, M.A. Baldo, F. Nizzi Grifi, M. Forina, Comparison between classical and innovative class-modelling techniques for the characterisation of a PDO olive oil, Anal. Bioanal. Chem. 399 (2011) 2105–2113, http://dx.doi.org/10.1007/s00216-010-4377-1.

[29] O.M. Kvalheim, T.V. Karstang, SIMCA - classification by means of disjoint cross validated principal components models, in: R.G. Brereton (Ed.), Multivar. Pattern Recognit. Chemom. Illus. By Case Stud, Elsevier Science, 1992, p. 232.

[30] 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, http://dx.doi.org/10.1016/S0003-2670(00)86468-5.

[31] M.P. Derde, D.L. Massart, UNEQ: a class modelling supervised pattern recognition technique, Mikrochim. Acta 89 (1986) 139–152, http://dx.doi.org/10.1007/BF01207313.

[32] H. Hotelling, Multivariate quality control illustrated by air testing of sample bombsights, in: C. Eisenhart, M.W. Hastay, W.A. Wallis (Eds.), Sel. Tech. Stat. Anal. Sci. Ind. Res. Prod. Manag. Eng, McGraw-Hill Book Company, Inc., New York and London, 1947, pp. 111–184.

[33] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, The Mahalanobis distance, Chemom. Intell. Lab. Syst. 50 (2000) 1–18, http://dx.doi.org/10.1016/S0169-7439(99)00047-7.

[34] M. Forina, S. Lanteri, L. Sarabia, Distance and class space in the UNEQ class-modeling technique, J. Chemom. 9 (1995) 69–89, http://dx.doi.org/10.1002/cem.1180090202.

[35] S. Wold, Pattern recognition by means of disjoint principal components models, Pattern Recognit. 8 (1976) 127–139, http://dx.doi.org/10.1016/0031-3203(76)90014-5.

[36] I.T. Jolliffe, Principal Component Analysis, Springer-Verlag, New York, 2002.

[37] S. Wold, Cross-validatory estimation of the number of components in factor and principal components models, Technometrics 20 (1978) 397–405, http://dx.doi.org/10.1080/00401706.1978.10489693.

[38] M. Forina, Fifty Years of Chemometrics, Fifty Years with Chemometrics, 2015, pp. 1–26, http://dx.doi.org/10.13140/RG.2.1.2199.3445.

[39] R. De Maesschalck, A. Candolfi, D.L. Massart, S. Heuerding, Decision criteria for soft independent modelling of class analogy applied to near infrared data, Chemom. Intell. Lab. Syst. 47 (1999) 65–77, http://dx.doi.org/10.1016/S0169-7439(98)00159-2.

[40] O.Y. Rodionova, K.S. Balyklova, A.V. Titova, A.L. Pomerantsev, Quantitative risk assessment in classification of drugs with identical API content, J. Pharm. Biomed. Anal. 98 (2014) 186–192, http://dx.doi.org/10.1016/j.jpba.2014.05.033.

[41] A.L. Pomerantsev, O.Y. Rodionova, Concept and role of extreme objects in PCA/SIMCA, J. Chemom. 28 (2014) 429–438, http://dx.doi.org/10.1002/cem.2506.

[42] A.L. Pomerantsev, Acceptance areas for multivariate classification derived by projection methods, J. Chemom. 22 (2008) 601–609, http://dx.doi.org/10.1002/cem.1147.

[43] P. de B. Harrington, Fuzzy grid encoded independent modeling for class analogies (FIMCA), Anal. Chem. 86 (2014) 4883–4892, http://dx.doi.org/10.1021/ac5001543.

[44] G.S. Jackson, J. Edward, Mudholkar, Control procedures for residuals associated with principal component analysis, Technometrics 21 (1979) 341–349. http://www.jstor.org/sici?sici=00401706(1979)21:3%3C341:%3E2.0.CO;2-J (Accessed 6 April 2014).

[45] D. Coomans, D.L. Massart, I. Broeckaert, A. Tassin, Potential methods in pattern recognition, Anal. Chim. Acta 133 (1981) 215–224, http://dx.doi.org/10.1016/S0003-2670(01)83196-2.

[46] M. Forina, C. Armanino, R. Leardi, G. Drava, A class-modelling technique based on potential functions, J. Chemom. 5 (1991) 435–453, http://dx.doi.org/10.1002/cem.1180050504.

[47] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn 20 (1995) 273–297, http://dx.doi.org/10.1007/BF00994018.

[48] D.M. Tax, R.P. Duin, Support Vector Domain Description, 1999, http://dx.doi.org/10.1016/S0167-8655(99)00087-2.

[49] B. Schölkopf, R.C. Williamson, A.J. Smola, J. Shawe-Taylor, J.C. Platt, Support Vector Method for Novelty Detection vol. 12, 1999, pp. 582–588.

[50] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130, http://dx.doi.org/10.1016/S0169-7439(01)00155-1.

[51] R.G. Brereton, G.R. Lloyd, Partial least squares discriminant analysis: taking the magic away, J. Chemom. 28 (2014) 213–225, http://dx.doi.org/10.1002/cem.2609.

[52] L. Xu, S.-M. Yan, C.-B. Cai, X.-P. Yu, One-class partial least squares (OCPLS) classifier, Chemom. Intell. Lab. Syst. 126 (2013) 1–5, http://dx.doi.org/10.1016/j.chemolab.2013.04.008.

[53] P. Oliveri, M.I. López, M.C. Casolino, I. Ruisánchez, M.P. Callao, L. Medini, S. Lanteri, Partial least squares density modeling (PLS-DM) - a new class-modeling strategy applied to the authentication of olives in brine by near-infrared spectroscopy, Anal. Chim. Acta 851 (2014) 30–36, http://dx.doi.org/10.1016/j.aca.2014.09.013.

[54] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148, http://dx.doi.org/10.1080/00401706.1969.10490666.

[55] R.D. Snee, Validation of regression models: methods and examples, Technometrics 19 (1977) 415–428, http://dx.doi.org/10.1080/00401706.1977.10489581.

[56] R.A. LaBudde, J.M. Harnly, Probability of identification: a statistical model for the validation of qualitative botanical identification methods, J. AOAC Int. 95 (n.d.) 273–285. http://www.ncbi.nlm.nih.gov/pubmed/22468371 (Accessed 11 May 2017).

[57] J. Harnly, P. Chen, P.D.B. Harrington, Probability of identification: adulteration of american ginseng with asian ginseng, J. AOAC Int. 96, (n.d.) 1258–1265. http://www.ncbi.nlm.nih.gov/pubmed/24645502 (Accessed 11 May 2017).

[58] M.H. Zweig, G. Campbell, Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine, Clin. Chem. 39 (1993).

[59] N.A. Obuchowski, Receiver operating characteristic curves and their use in radiology, Radiology 229 (2003) 3–8, http://dx.doi.org/10.1148/radiol.2291010898.

[60] V. Pirro, P. Oliveri, B. Sciutteri, R. Salvo, A. Salomone, S. Lanteri, M. Vincenti, Multivariate strategies for screening evaluation of harmful drinking, Bioanalysis 5 (2013) 687–699, http://dx.doi.org/10.4155/bio.13.12.

[61] D. Coomans, I. Broeckaert, M.P. Derde, A. Tassin, D.L. Massart, S. Wold, Use of a microcomputer for the definition of multivariate confidence regions in medical diagnosis based on clinical laboratory profiles, Comput. Biomed. Res. 17 (1984) 1–14, http://dx.doi.org/10.1016/0010-4809(84)90002-8.

[62] P. Oliveri, M.C. Casolino, M. Casale, L. Medini, F. Mare, S. Lanteri, A spectral transfer procedure for application of a single class-model to spectra recorded by different near-infrared spectrometers for authentication of olives in brine, Anal. Chim. Acta 761 (2013) 46–52, http://dx.doi.org/10.1016/j.aca.2012.11.020.

[63] A.K. Smilde, A. Knevelman, P.M.J. Coenegracht, Introduction of multi-criteria decision making in optimization procedures for high-performance liquid chromatographic separations, J. Chromatogr. A 369 (1986) 1–10, http://dx.doi.org/10.1016/S0021-9673(00)90093-1.

[64] O.Y. Rodionova, P. Oliveri, A.L. Pomerantsev, Rigorous and compliant approaches to one-class classification, Chemom. Intell. Lab. Syst. 159 (2016) 89–96, http://dx.doi.org/10.1016/j.chemolab.2016.10.002.

Paolo Oliveri (PhD in Sciences and Technologies of Chemistry and Materials, 2010) currently works as assistant professor at the Department of Pharmacy of the University of Genova. His research is mainly focused on spectroscopic analytical methods and chemometrics, with particular regard to signal processing and pattern recognition techniques. Awarded as “Best Young Researcher” by the Division of Analytical Chemistry of the Italian Society of Chemistry (2010). Co-author than more than 50 scientific journal papers and book chapters and more than 100 communications at national and international conferences. Principal investigator of a three-year research project (Scientific Independence of young Researchers – SIR) founded by the Italian Ministry of Education, Universities and Research (MIUR). Member of the board of the Division of Analytical Chemistry of the Italian Society of Chemistry (SCI).