Manuscript Details

Manuscript number	CHEMOLAB_2016_79
Title	Rigorous and compliant approaches to one-class classification
Article type	Research paper

Abstract

A wide number of real problems requiring qualitative answers should be addressed by one-class classification (OCC), as in the case of authentication studies, verification of particular claims and quality control. The key feature of OCC is that models are developed using only samples from the target class, so that a representative sampling is not strictly required for non-target classes. On the contrary, in the discriminant analysis (DA) approach, all of the classes considered (at least two) have a non-negligible influence in the definition of the delimiter. It follows that faults in the definition of the classes involved and in representative sampling for each of them may determine a bias in the classification rules. A key aspect in one-class classification concerns model optimisation. When the optimal modelling conditions are searched by considering parameters such as type II error or specificity ('compliant' approach), information from the non-target class is being used and may therefore determine a bias in the model. In order to build pure class models ('rigorous' approach), only information from the targeted class should be regarded: in other words, optimisation should be performed only considering type I error, or sensitivity. In the present study, 'compliant' and 'rigorous' approaches are critically compared on real case studies, by applying two novel modelling techniques: partial least squares density modelling (PLS-DM) and data driven soft independent modelling of class analogy (DD-SIMCA).

Keywords	one-class classification; class modelling; discriminant analysis; sensitivity; specificity
Taxonomy	Linear Classifiers, Density Based Approaches, Pattern Recognition
Manuscript category	General Section

Corresponding Author	Paolo Oliveri
Order of Authors	Oxana Rodionova, Paolo Oliveri, Alexey Pomerantsev

Suggested reviewers	Michal Daszykowski, Karoly Heberger, Gerard Downey

Submission Files Included in this PDF

File Name [File Type]

CoverLetter.pdf [Cover Letter]

Answers-to-Reviewers.doc [Response to Reviewers]

Manuscript_revised.doc [Revised Manuscript with Changes Marked]

Manuscript_revised_clear.doc [Manuscript File]

Figure-Captions_revised.docx [Figure]

Figure_01.docx [Figure]

Figure_02.docx [Figure]

Figure_03.docx [Figure]

Figure_04.docx [Figure]

Figure_05_revised.docx [Figure]

Table_01.docx [Table]

Table_02.docx [Table]

Table_03_revised.docx [Table]

Table_04_revised.docx [Table]

Table_05_revised.docx [Table]

Highlights.docx [Highlights]

To view all the submission files, including those not included in the PDF, click on the manuscript title on your EVISE Homepage, then click 'Download zip file'.

Rigorous and compliant approaches to one-class classification

Oxana Ye. Rodionova[1], Paolo Oliveri[2,*], Alexey L. Pomerantsev[1]

[1]*N.N. Semenov Institute of Chemical Physics RAS, Kosygin 4, 119991, Moscow, Russia.*

[2]*Department of Pharmacy, University of Genoa, Viale Cembrano, 4, I-16148 Genoa, Italy.*

*Corresponding Author:

Paolo Oliveri

Department of Pharmacy, University of Genoa

Viale Cembrano, 4

I-16148, Genoa, Italy

Tel. +39 010 3532626

Fax. +39 010 3532684

email: oliveri@difar.unige.it

1 Abstract

A wide number of real problems requiring qualitative answers should be addressed by one-class classification (OCC), as in the case of authentication studies, verification of particular claims and quality control. The key feature of OCC is that models are developed using only samples from the target class, so that a representative sampling is not strictly required for non-target classes. On the contrary, in the discriminant analysis (DA) approach, all of the classes considered (at least two) have a non-negligible influence in the definition of the delimiter. It follows that faults in the definition of the classes involved and in representative sampling for each of them may determine a bias in the classification rules. A key aspect in one-class classification concerns model optimisation. When the optimal modelling conditions are searched by considering parameters such as type II error or specificity ('compliant' approach), information from the non-target class is being used and may therefore determine a bias in the model. In order to build pure class models ('rigorous' approach), only information from the targeted class should be regarded: in other words, optimisation should be performed only considering type I error, or sensitivity. In the present study, 'compliant' and 'rigorous' approaches are critically compared on real case studies, by applying two novel modelling techniques: partial least squares density modelling (PLS-DM) and data driven soft independent modelling of class analogy (DD-SIMCA).

Keywords: one-class classification; class modelling; discriminant analysis; sensitivity; specificity.

1. Introduction

One-class classification (OCC) [1,2] consists in making a description of a target class of objects and in detecting whether a new object resembles this class or not. The term class modelling is often used for denoting OCC methods [3]. In some sense, this approach is opposite to the discrimination problem that is to allocate a new object to one of distinct and exhaustive classes [4].

The critical difference between OCC and discriminant analysis (DA) is that the OCC model is developed using target class samples only.

The work of Harold Hotelling on multivariate quality control (1947) can be considered as the first example of multivariate one-class classification in chemistry [5]. The unequal class models (UNEQ) method was developed by Derde and Massart (1986) as an evolution of these concepts [6]. In fact, such a method – closely related to quadratic discriminant analysis (QDA) – is based on the hypothesis of a multivariate normal distribution in the class to be modelled and defines the width of the class space based on Hotelling's T^2 statistics, at a selected confidence level.

The first method specifically developed for one-class classification in chemometrics was soft independent modelling of class analogy (SIMCA), by Svante Wold [7,8]. This method performs PCA on the samples of the class to be modelled – the SIMCA model being defined as the range of sample scores on the significant PCs. A critical distance, at a given confidence level, is obtained by application of the Fisher F statistics to residuals of each training sample to the model, and is used to define the boundaries of the SIMCA class space around the model.

OCC modelling is a rather new strategy in comparison with DA. The classical OCC version does not utilise any information about non-target (extraneous) classes even when the data regarding such extraneous classes is available. We call such an approach a 'rigorous' one. Contributing to the OCC technique elaboration, we consider the outcomes that can be yielded in case the rigorous concept is violated. The most common violation – which we call a 'compliant' approach – makes use of some relevant non-target information that can influence the results of the OCC modelling.

The main objective of the present study is the comparison between the outcomes of 'rigorous' and 'compliant' approaches. For this purposes, two different OCC methods, namely, partial least squares density modelling (PLS-DM) [9], and data-driven soft independent modelling of class analogy (DD-SIMCA) [10] are employed. Method descriptions are presented in sections 3.1 and 3.2. An additional goal is to compare these techniques using two real world examples.

2 Theory

2.1 Figures of merit

Performances of one-class classifiers are usually reported using two parameters: sensitivity and specificity. Sensitivity is the fraction of samples of the target class which are correctly recognised as consistent with the model. It can also be defined as the rate of true positives and, therefore, it is complementary to type I error (*i.e.*, the false negative rate). Specificity is the fraction of samples extraneous to the target class which are correctly recognised as inconsistent with the model, corresponding to the rate of true negatives. This parameter is therefore complementary to type II error (*i.e.*, the false positive rate). Efficiency of one-class classifiers is usually defined as the geometric mean of sensitivity and specificity [11].

When these two parameters are considered, it is very important to realise on which sample subset each parameter was calculated. First of all, let us consider the type I error, α. At the stage of model building, some OCC methods enable to set a prior value of α and to use it for establishing the corresponding threshold. After that, it is possible to calculate sensitivity for the training set, referred to as SENS_T in the following sections. This value is a classification analogue of the root mean square error of calibration (RMSEC) for calibration problems. It is important to verify that SENS_T is in agreement with the a-priori α value. Varying the α value, it is possible to control the risk of wrong rejections of target objects. The value of sensitivity that characterises the quality of predictions should be calculated as the fraction of samples from the target *test set* which are correctly recognised (SENS_P). In this case, SENS_P is the classification analogue of the root mean square error of prediction (RMSEP). We can also calculate sensitivity using the cross-validation approach. The corresponding value is referred to as SENS_V. It is worth mentioning that calculation of sensitivity as the fraction of all samples form the target class (training plus test samples) can provide misleading results especially in case the test set is rather small.

For OCC models, specificity is calculated only in the presence of non-target objects. This figure of merit is obtained empirically or, for some OCC method, it can be calculated theoretically [12] as the type II error, β. In case non-target objects are organised in several extraneous classes, specificity should be calculated for each extraneous class separately; otherwise, the reported value of specificity would not reflect the true relationships between the target and alternative classes. Such an example is presented below in Section 4. At the same time, total specificity can be reported, if the customer is not interested in the details.

It should be mentioned that both sensitivity and specificity depend on the selected value of type I error, α. The first parameter has a direct relation: SENS = (1–α). Conversely, dependence of specificity is more complex and will be considered in Section 4.

2.2 'Rigorous' *vs.* 'Compliant' approach

We distinguish two approaches when building OCC models. We call the first one 'rigorous' OCC. This means that a model is developed based merely on the target training dataset, and optimal conditions are obtained employing the type I error, α (the rate of wrong rejections of the target samples), or sensitivity, computed as (1- α). Depending on the method, this α value may be estimated a-priori, and/or calculated a-posteriori. This evaluation is made using the target samples only. Considering that sensitivity is an experimental estimate of type I error, α, of a given model, outcomes whose sensitivity is closest to (1- α) should be considered as optimal, when optimising a model in a 'rigorous' way.

The second approach is called here 'compliant' OCC. Such a very common modelling strategy utilises additional information regarding non-target samples when, except for data from the target class, one/several datasets from extraneous classes are available. For each alternative class, the type II error, β (the rate of wrong acceptances of objects from the alternative class), or the corresponding specificity, computed as (1-β), is estimated. In this case, model optimisation is performed with

respect to the estimates of both α and β and the OCC model that has maximum efficiency (*i.e.*, maximum sensitivity and specificity) is selected.

2.3 Data driven soft independent modelling of class analogy (DD-SIMCA)

The DD-SIMCA method develops a decision rule that delineates the objects from the target class by exploring the corresponding data matrix. The procedure consists of two steps. The first step is the application of principal component analysis (PCA) [13], establishing a model using training samples from the target class. The ($I \times J$) data matrix **X** (duly pre-processed, *e.g.* centred) is decomposed by:

$$\mathbf{X}=\mathbf{T}\mathbf{P}^{\mathrm{t}}+\mathbf{E} \tag{1}$$

where $\mathbf{T}=\{t_{ia}\}$ is the ($I \times A$) scores matrix; $\mathbf{P}=\{p_{ja}\}$ is the ($J \times A$) loadings matrix; $\mathbf{E}=\{e_{ij}\}$ is the ($I \times J$) matrix of residuals; and A is the number of principal components (PC). Matrix $\mathbf{T}^{\mathrm{t}}\mathbf{T}=\mathbf{\Lambda}=\mathrm{diag}(\lambda_{1,\ldots,}\ \lambda_A)$ is a diagonal matrix with elements $\lambda_a=\sum_{i=1}^{I} t_{ia}^2$, which are the eigenvalues of matrix $\mathbf{X}^{\mathrm{t}}\mathbf{X}$ ranked in descending order.

In the second step, we employ the PCA results when calculating two relevant distances for each object i=1,..., I of the training set. They are the score distance (SD), h_i, and the orthogonal distance (OD), v_i:

$$h_i=\mathbf{t}_i^{\mathrm{t}}(\mathbf{T}^{\mathrm{t}}\mathbf{T})^{-1}\mathbf{t}_i=\sum_{a=1}^{A}\frac{t_{ia}^2}{\lambda_a}, \qquad v_i=\sum_{j=1}^{J} e_{ij}^2 \tag{2}$$

SD represents the position of a sample within the score space, while OD characterises a sample distance to the score space.

In a previous study [14], it was shown that distributions of both distances are well approximated by the scaled chi-squared distribution:

$$N_h \frac{h}{h_0} \propto \chi^2(N_h) \qquad N_v \frac{v}{v_0} \propto \chi^2(N_v) \tag{3}$$

where v_0 and h_0 are the scaling factors, N_h and N_v are the numbers of the degrees of freedom (DoF). These parameters are considered unknown and estimated using a data-driven method explained in ref. [10].

Statistics c, called the *total distance*:

$$c = N_h \frac{h}{h_0} + N_v \frac{v}{v_0} \propto \chi^2(N_h + N_v)\,. \tag{4}$$

is used to generate the decision rules. Any decision rule (*i.e.*, an acceptance area) is determined by an inequality:

$$c \le c_{\text{crit}}. \tag{5}$$

The first decision rule is developed for a given type I error, α:

$$c_{\text{crit}} = \chi^{-2}(1-\alpha, N_h + N_v) \tag{6}$$

where χ^{-2} is the quantile of the chi-squared distribution with N_v+N_h DoF. To calculate the type II error β, we should assume that an alternative class is available:

$$\beta = \Pr\left\{\chi'^2(N_h + N_v, s) < \frac{c_{\text{crit}}}{c_0'}\right\}, \tag{7}$$

where c_{crit} is defined in Eq. (5) and χ'^2 is the non-central chi-squared distribution. Parameters c_0' and s are found by the method explained in ref. [12].

Using this approach, every sample and the acceptance areas can be plotted in the coordinates of h/h_0 against v/v_0. Fig. 2 illustrates an example of this distance plot. Applying theory (Eqs. (6) and (7)) in practice, we can yield two acceptance areas. First, we can develop a 'rigorous' decision rule defined by a given α, and then calculate a subsequent β error. On the other hand, we can employ a 'compliant' rule (area) defined by β, and obtain a subsequent α. The first rule is stronger, because all

samples accepted by the 'rigorous' rule are simultaneously accepted by the 'compliant' rule, but the converse is not true.

2.4 Partial least squares density modelling (PLS-DM)

The method develops a PLS model using dataset **X** as the predictor matrix and a density vector as the **y** response vector. The response value (y_i) – for each sample i of the training set – is computed as an estimation of sample density, based on inter-sample distances in the multivariate space. In more detail, all the Euclidean distances (d) from sample i to each of the other training samples are computed. Such distances are, therefore, ordered, and the density value, is obtained as the sum of the k smallest (*i.e.*, lowest-order) distances:

$$y_i = \sum_{j=1}^{k} d_{i,j} \tag{8}$$

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). The global probability function, $f(x)$, is obtained by summing the individual contributions $f_i(x)$ defined as:

$$f(x) = \sum_{i=1}^{I_C} f_i(x) = \sum_{i=1}^{I_C} \frac{1}{I_c} \cdot \prod_{v=1}^{V} \frac{1}{as_v\sqrt{2\pi}} e^{-\frac{1}{2}\cdot\frac{(x_v - x_{iv})^2}{(as_v)^2}} \tag{9}$$

where a is the smoothing coefficient that applies to all of the variables within the training set, and s_v is the standard deviation of variable v within the training set. The smoothing coefficient cooperates in determining the shape of the distribution, being higher the smoothness when increasing a (which usually ranges between 0.1 and 1.5).

In order to define the class boundary, the critical value (f_α) of the probability density distribution $f(x)$, at a selected type I error, α, is obtained from the critical value of the chi-squared distribution (χ^2_α) with V degrees of freedom by the so-called equivalent determinant method [15], according to:

$$f_\alpha = \frac{1}{(2\pi)^{\frac{V}{2}} \left|\hat{\mathbf{C}}\right|^{\frac{1}{2}}} \cdot e^{-\frac{\chi^2_\alpha}{2}} \tag{10}$$

where V is the number of variables and $\left|\hat{\mathbf{C}}\right|$ is the estimation of the determinant corresponding to the variance-covariance matrix of the multivariate normal distribution equivalent to the probability distribution estimated by PFM, computed as:

$$\left|\hat{\mathbf{C}}\right|^{\frac{1}{2}} = \frac{1}{2^V \pi^{\frac{V}{2}} \frac{\sum f(x)}{I}} \tag{11}$$

The term 'equivalent' is used in this context to indicate distributions with the same mean value[15].

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 [16]. In this way, compliance of each object with the class model is granted when it complies with both f_α and Q_α criteria.

The algorithm calculates models with all of the different parameter combinations – *i.e.* the k distance, the a smoothing coefficient and the L latent variables, as well as the suitable **X**-block *pre-processing*. Then, the procedure selects the optimal parameter combination by: (1) fixing the number of L through the efficiency criterion and (2) evaluating the rest of parameters applying a Pareto's multicriteria decision method ('compliant' strategy) or by evaluating only type I error, α ('rigorous' strategy).

3 Materials

We consider two different types of datasets. One set, *Olives,* is comprised of samples of natural origin, olives in brine. Variability among samples is inevitable. In the present study, variability is taken into account both within a single harvest year and between different harvest years. The second dataset, *Remedy,* consists of samples of artificial origin, uncoated tablets. Certainly, variability between samples is much less and mainly manifests as variation between batches.

3.1 Dataset *Olives*

Dataset *Olives* consists of very close/overlapped classes: one target class, *Taggiasca*, and two non-target classes that are considered as potential adulterants, *Leccino*, and *Coquillo*. Data are organised as follows. Each class consists of three sub-sets: training and internal test samples, both composed by olives from harvests 2010-11 and 2011-12, and external set, composed by olives from harvest 2012-13 [9]. Data summary is presented in Table 1.

Data from subset *Taggiasca* are considered as the target class due to the request of the customer, who is interested in the confirmation of authenticity of this product. Subsets I1 and E1 are not involved in the development of the 'rigorous' *Taggiasca* model; they are used as an independent test set to validate the quality of authentication models. Other subsets are employed for calculation of specificity.

NIR spectra of dataset *Olives* were recorded by a FT-NIR Thermo Scientific spectrometer (Thermo Scientific, AntarisII™ FT-NIR Analyser), in the 4,000–10,000 cm^{-1} range, at 4 cm^{-1} resolution. Samples were analysed in the reflection mode using standard glass Petri dishes (9 cm diameter). Spectral profiles of each sample were acquired as the mean of 64 scans recorded during rotation of the Petri dish. Systematic differences among Petri dishes – mainly due to small variations in glass thickness – were corrected by dividing point by point the reflectance spectrum of each

sample by the spectrum of a certified reference material (Spectralon®) with 99% reflectance in the entire NIR region, recorded on the same dish. The whole analytical procedure was repeated on three different aliquots of each sample and the resulting average spectrum was submitted to data analysis. Acquired spectra were corrected by the standard normal variate (SNV) transform.

3.2 Dataset *Remedy*

Dataset *Remedy* is a part of a bigger dataset that was analysed in details in ref. [17]. The samples are uncoated tablets of calcium channel blockers, produced by three different manufacturers, denoted as A, B, and C. All the manufactures employed the same quantity, 10 mg, of the active pharmaceutical ingredient (API) originated from the same source. Each manufacture is represented by a set of batches ranging from five to ten. Each batch consists of 10 tablets. Overall, there are 220 objects in dataset *Remedy*, whose summary is presented in Table 2. This set imitates the fakes of various 'quality' in case real counterfeited objects are unavailable. The range of producers considered against a specific genuine class may be used for assessing the target class acceptance areas, which account for a possible future variation of the target class. Each class in turn is considered as a target class.

NIR spectra of dataset *Remedy* were acquired in the interval 4000-12500 cm^{-1} with a resolution of 8 cm^{-1} using the FT-NIR spectrometer MPA (by Bruker Optics) equipped with a handheld fibre-optic probe (FP). Measurements were carried out in the diffuse reflection mode through an optically transparent PVC blister. Each time, triplicate readings are made to control repeatability. Replicas were averaged for data analysis. All spectra were pre-processed by a second-order Savitzky-Golay derivation with a 21 data-point window size and a third-order polynomial. This transformation was used to remove most artefacts caused by the PVC blister and application of the handheld FP.

Aim of the study is to build three independent models for each of the manufactures. It is important both to classify new objects from the same manufacture as objects from the target class and to reveal alien objects.

PCA applied to the whole data sets shows the overall disposition of the subsets (Fig. 1).

4 Results for dataset *Olives*

4.1 DD-SIMCA

As it was mentioned above, two types of models, 'rigorous' and 'compliant', are considered. The results regarding model sensitivity are presented in Table 3. The best results for the 'rigorous' model are obtained with 3 PCs and type I error α=0.01. Both a-priori α values are in good agreement with a-posteriori sensitivity calculated for subsets T1, I1 and E1.

At the same time, specificity is not completely satisfactory (see Fig. 2.a and Table 4). Misclassification results are originated from subsets T3 and I3 (*Coquillo* olives). Taking into account knowledge regarding extraneous samples, the 'compliant' model, which is more complex, 6 PCs, provides better results (see Fig. 2.b and Table 4), providing 0.98 specificity at α=0.05.

By varying α values, we can select the risk of wrong rejection and wrong acceptance: α=0.05 provides better separation between classes on the cost of wrong rejection of 5% of target objects. Instead, for α=0.01, sensitivity is excellent but the risk of acceptance of non-target objects increases.

It is important to notice that, in this example reporting the results as 'total specificity', we hinder the real problem of misclassification, because the source of misclassification is the *Coquillo* class. Thus, a joint calculation of specificity for all alien objects does not reflect the real problem.

4.2 PLS-DM

Application of PLS-DM to the *Olives* data set was performed following both the 'compliant' and the 'rigorous' approaches. Different parameters were varied according to a full factorial design (*i.e.*, all of the possible combinations were tested), within cross-validation (CV) cycles with five deletion groups and Venetian-blind scheme. In more detail, different levels of column pre-processing were tested (namely: raw data, column centring, column scaling, and column autoscaling); k was varied from 1 to 6; a was varied from 0.3 to 0.8, and L was varied from 1 to 10.

Table 3 reports the CV sensitivity values obtained for the target class T1 (*Taggiasca*), at both α=0.05 and α=0.01, for the 'rigorous' and the 'compliant' approaches, respectively. Table 4 reports total and partial specificity results for the *Olives* data set under the optimal conditions selected.

Very satisfactory results are obtained, especially for α=0.05 while, as it can be expected, specificity decreases when decreasing α. Specificity of the target class T1 is higher as long as samples of class *Leccino* (I2 and E2) are considered, while less satisfactory results are obtained when specificity is evaluated by samples of class *Coquillo* (I3). This may suggest that olives belonging to cultivar *Coquillo* are more similar to the *Taggiasca* type than the *Leccino* cultivar, as long as their NIR spectral features are considered.

5 Results for dataset *Remedy*

Unlike the *Olives* case, we consider three peer subsets corresponding to three different manufactures. Samples originated from each manufacture are considered as target class samples and three OCC models are built respectively.

5.1 DD-SIMCA

In the case class A is considered as the target class, attention should be devoted to accurately defining the training conditions. Two PCs are not enough. Three PCs and α=0.01 are chosen to reach sensitivity equal to 1.00 (see Table 5 and Fig. 4.a). Datasets B and C are well-separated from the target objects and specificity is equal to 1.00. Extraneous classes are rather far and the β value for each of the extraneous classes is close to zero.

In the case we consider class C as the target, two PCs and α=0.01 can be chosen as the optimal conditions (see Table 5 and Fig. 3.a). Both of the extraneous classes are well separated from the target class, with specificity =1. From Fig. 3.b, it can be observed that class B is located closer to the border of the acceptance area than class A. The β value for class B is equal to 0.006. This means that, theoretically, one out of 170 objects from class B can be wrongly attributed to the target class C.

Results for class B are rather similar to those for class C. The model with 2 PCs, at α=0.01, is chosen as optimal (see Table 5). Both of the extraneous classes are well separated from the target class, with specificity =1. Though class C, as one might expect, is located closer to the threshold than class A. The β value for class C is equal to 0.001.

The main challenge in modelling *Remedy* data is some heterogeneity inside target classes caused by natural variation between batches of one manufacture. Classes are well separated and consideration of the extraneous subsets on the modelling stage does neither bring additional information nor provide any model improvement in comparison with 'rigorous' DD-SIMCA modelling.

Thus results of 'compliant' modelling are the same as for the 'rigorous' one and are not provided here.

5.2 PLS-DM

Interestingly, application of PLS-DM on the *Remedy* dataset led to identical choices in terms of optimal model conditions, for both the 'compliant' and the 'rigorous' strategies. In fact, as summarised in Table 5, models with sensitivity and specificity equal to 1.00 on the test set prediction were obtained for classes A and B, at α=0.01.

Sensitivity values achieved as the prediction on the test set cannot be considered satisfactory for class C. This could be ascribable to data overfitting – a conceivable occurrence when model complexity (*i.e.*, the number of LV) increases.

6 Discussion

Comparing the two OCC methods, we can conclude that DD-SIMCA is a global modelling method, while PLS-DM represents a local approach. At a fixed level of type I error, α, the first method has the only free parameter – the number of PCs – that can be used for tuning in case of 'compliant' approach. When the number of PCs is increased, training sensitivity is varying near to the given sensitivity level (1–α), while validation sensitivity is decreasing. These tendencies are observed due to evident facts. For a given number of PCs, the DD-SIMCA model is developed to the better accounting for all variations in the training set. The test set is not involved in this process, and accounting for its own variations is beyond the scope of the study. A more complex model means that a higher amount of variation in the training set is taken into account, and that more specific properties of the test set come into conflict with this model. In case the test set is a good representation of the target class population, test sensitivity is not decreased when a small number of PCs is used, because the main variations are common for the test and training sets.

This trade-off analysis provides a way for the selection of a proper number of PCs in the pattern recognition context, see e.g. [18], and it was used in the 'rigorous' application of DD-

SIMCA. In the case of the 'compliant' approach, a balanced number of PCs was selected, with the goal of improving specificity, which is also increased while increasing PC number.

A typical example is shown in Fig. 4, which demonstrates the figures of merit for dataset *Olives* obtained at α =0.05 as a function of the number of PCs. In the 'rigorous' case, we obtain curve 1 (training sensitivity, SENS_T) and curve 2 (validation sensitivity, SENS_V). Curve analysis leads to the conclusion that PCs=3 is the optimal choice, with SENS_T=0.95 and SENS_V=0.93.

In the 'compliant' case, we are considering two additional curves: 3 (test set sensitivity, SENS_P), which is going down, and 4 (total specificity, SPEC), which is growing. Our goal is to select the number of PCs at which specificity is large enough and sensitivity is satisfactory. Our choice is PCs=6, with SPEC=0.98, but SENS_T=0.93, SENS_V=0.83, SENS_P=0.79.

The PLS-DM approach demonstrates a similar behaviour. Considering the effect of modelling parameters, in the case of PLS-DM, variations of parameter k do not affect in a systematic way sensitivity and specificity outcomes, while an increment of parameter a usually leads to higher sensitivity and moderately lower specificity. The effect of increasing model complexity (*i.e.*, the number of latent variables L) is usually the opposite.

Fig. 5 graphically reports, as an example, the results which were evaluated in terms of sensitivity ('rigorous' approach, Fig. 5.a), and in terms of sensitivity and specificity ('compliant' approach, Fig. 5.b), for α=0.05 (confidence level = 0.95). In the case of the 'rigorous' approach, the outcomes are evaluated in terms of sensitivity and model complexity (number of latent variables, L). Each model is represented by a scatter point, while the blue line represents the pre-determined confidence level (0.95). Since sensitivity is an experimental estimate of the confidence level, models with sensitivity values closest to the confidence level and lowest model complexity are selected as optimal. In this case, models complying with such a requirement can be obtained for L = 1 to 3 and, following the 'rigorous' approach, the three choices are equivalent. An analogue plot was used for making decisions at α=0.01 (*not shown*).

In the case of the 'compliant' approach, a Pareto chart is used to select the optimal conditions (Fig. 5.b). Blue points represent the individual models, while the red line indicates the so-called Pareto front, which connects the optimal solutions in terms of sensitivity and specificity.

We see that, regardless of the employed OCC method, the same tendency is held. Any improvement of specificity can only be done at the cost of a sensitivity decreasing.

Conclusions

The distinct feature of OCC is possibility to build a model for one class without in-depth information regarding other classes or samples. In the 'rigorous' OCC approach, all model parameters and validation procedures are based only using information regarding the target class. This can be considered as an advantage of OCC, especially for solving authentication problems. At the same time, for overlapping datasets, this is a drawback. When the classes under study are well separated, the 'rigorous' and the 'compliant' approaches may lead to very similar or even identical optimisation outcomes, like in the case of data set *Remedy*. Conversely, when classes are characterised by complex distributions and tend to overlap, the two approaches may lead to different outcomes, like in the case of data set *Olives*.

Depending on the dataset under consideration, the importance of extraneous objects is different. In any case, application of 'compliant' modelling brings additional information to the modelling stage and, depending of the specific OCC method applied, it may provide more reliable results. It is important to underline that, theoretically, non-target objects could be as close as possible to a target class. Even the intensive training and validation measures cannot prevent us from unavoidable misclassification.

Application of OCC is also possible for the goal of discrimination, if each class of interest is modelled using the OCC method and, afterwards, the minimum distance from the different models

is assumed as the discriminant criterion. Though, it should be taken into account that, in the latter case, information regarding extraneous classes is used partly and implicitly. We definitely do not recommend using any OCC method for DA purposes.

Acknowledgments

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).

References

[1] M.M. Moya, M.W. Koch, L.D. Hostetler, One-class classifier networks for target recognition applications, in: I. International Neural Network Society (Ed.), Proc. World Congr. Neural Networks, Portland, OR, 1993: pp. 797–801.

[2] D.M.J. Tax, R.P.W. Duin, Outlier detection using classifier instability, in: Springer Berlin Heidelberg, 1998: pp. 593–601. doi:10.1007/BFb0033283.

[3] M.P. Derde, D.L. Massart, Comparison of the performance of the class modelling techniques UNEQ, SIMCA, and PRIMA, Chemom. Intell. Lab. Syst. 4 (1988) 65–93. doi:10.1016/0169-7439(88)80013-3.

[4] 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. doi:10.1016/j.trac.2016.01.010.

[5] 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.

[6] 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. doi:10.1016/S0003-2670(00)86468-5.

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

[8] M. Wold, Svante; Sjöström, SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy, in: B.R. Kowalski (Ed.), Chemom. Theory Appl., American Chemical Society, Washington, D.C., 1977: pp. 243–282. doi:10.1021/bk-1977-0052.

[9] 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. doi:10.1016/j.aca.2014.09.013.

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

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

[12] A.L. Pomerantsev, O.Y. Rodionova, On the type II error in SIMCA method, J. Chemom. 28 (2014) 518–522. doi:10.1002/cem.2610.

[13] H. Martens, T. Næs, Multivariate calibration, Wiley, 1989.

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

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

[16] G.S. Jackson, J Edward; Mudholkar, Control Procedures for Residuals Associated with Principal Component Analysis, Technometrics. 21 (1979) 341–349.

[17] 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. doi:10.1016/j.jpba.2014.05.033.

[18] B. Krakowska, D. Custers, E. Deconinck, M. Daszykowski, The Monte Carlo validation framework for the discriminant partial least squares model extended with variable selection methods applied to authenticity studies of Viagra® based on chromatographic impurity profiles, Analyst 141 (2016) 1060-1070. doi: 10.1039/C5AN01656H.

Figure Captions

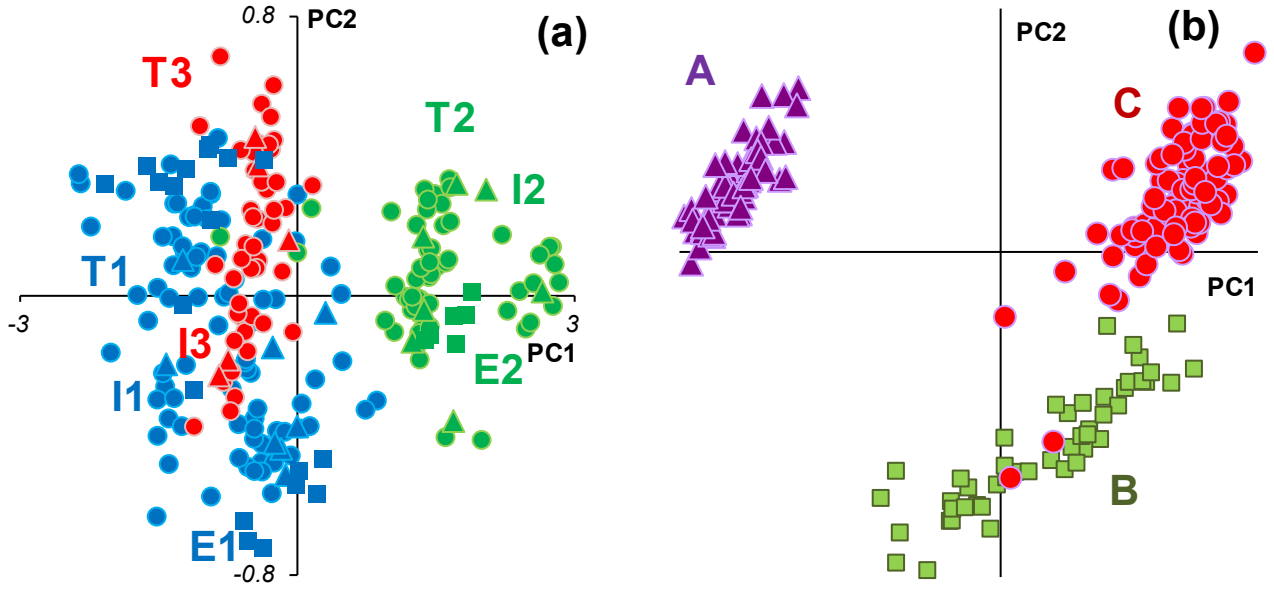
Fig. 1. Joint PCA using all data. Score plots PC1 *vs*. PC2. (a) Dataset *Olives*; (b) dataset *Remedy*.

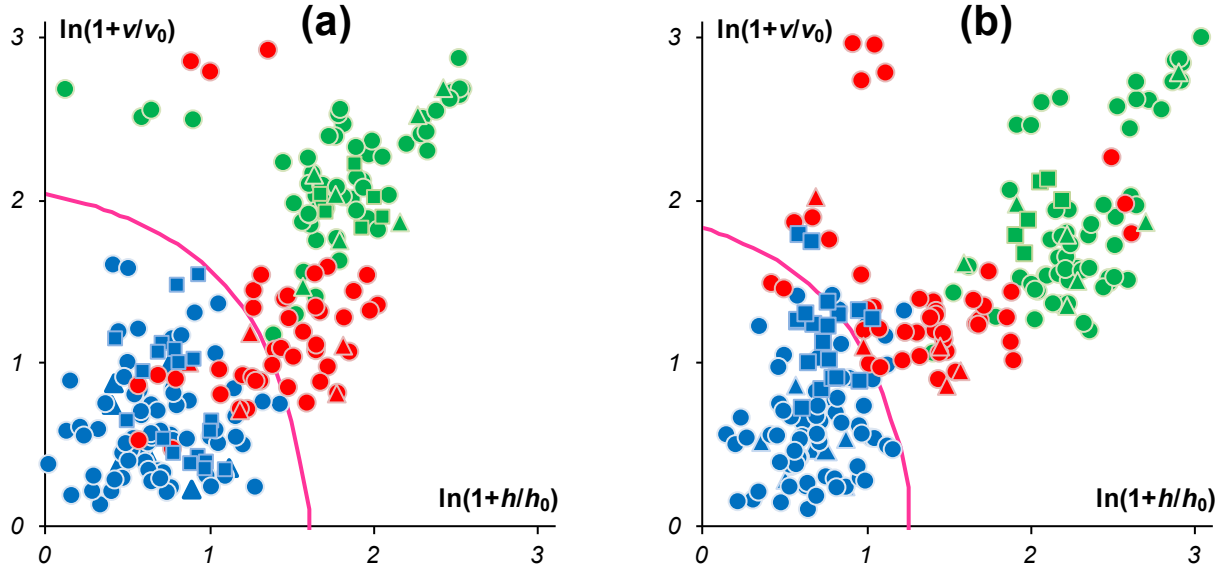
Fig. 2. 'Rigorous' (subplot a, 3 PCs, α=0.01) and 'compliant' (subplot b, 6 PCs, α=0.05) models for *Olives* dataset.

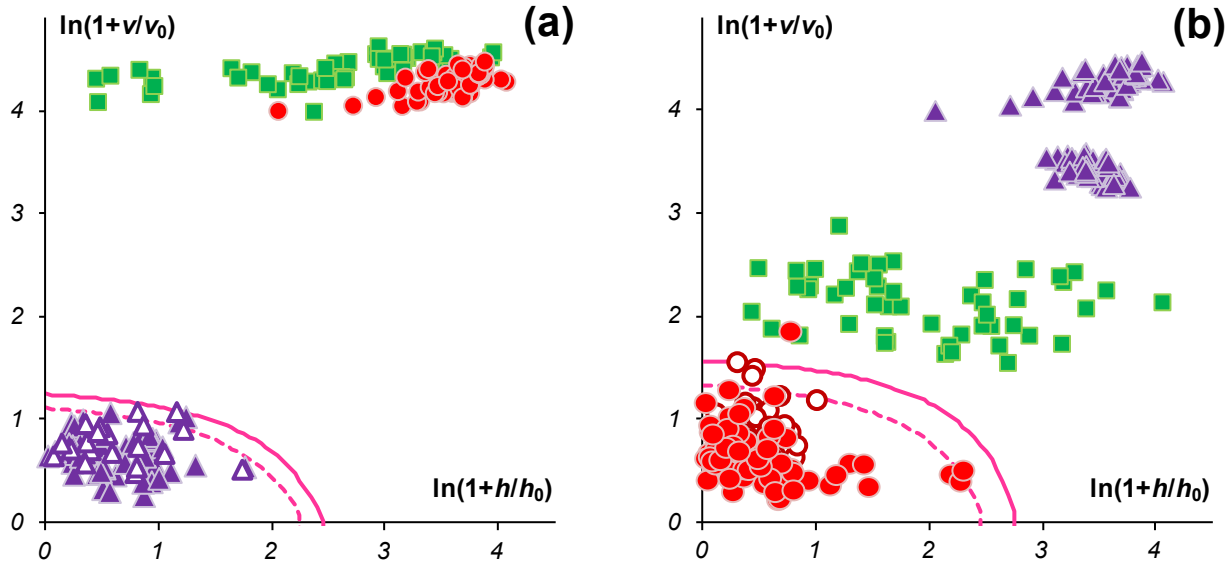
Fig. 3. Modelling in case the target is class A (subplot a, 3 PCs) and class C (subplot b, 2 PCs). Two acceptance thresholds, α=0.05 (dotted line) and α=0.01 (solid line) are shown.

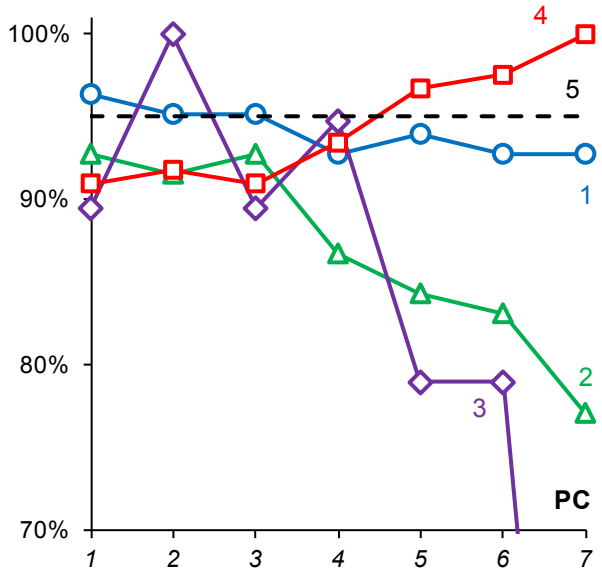
Fig. 4. Figures of merit (α=0.05) for dataset *Olives*. (1, blue circles) training sensitivity, (2, green triangles) cross-validation sensitivity; (3, violet rhombs) test set E1 sensitivity, (4, red squares) total specificity; (5, black dashed line) confidence threshold (1−α)~~100%~~.

Fig. 5. *Olives* data set. Outcomes of 'rigorous' (a) and 'compliant' (b) approaches, α=0.05 (confidence level = 0.95~~%~~).


(a)
PC2
0.8
-0.8
-3
3
PC1
T3
T2
I2
T1
I3
I1
E2
E1
(b)
PC2
PC1
A
C
B


(a)
ln(1+v/v0)
ln(1+h/h0)
0
1
2
3
(b)
ln(1+v/v0)
ln(1+h/h0)
0
1
2
3


(a)
ln(1+v/v0)
ln(1+h/h0)
0
1
2
3
4
(b)
ln(1+v/v0)
ln(1+h/h0)
0
1
2
3
4


100%
90%
80%
70%
1
2
3
4
5
6
7
PC
1
2
3
4
5

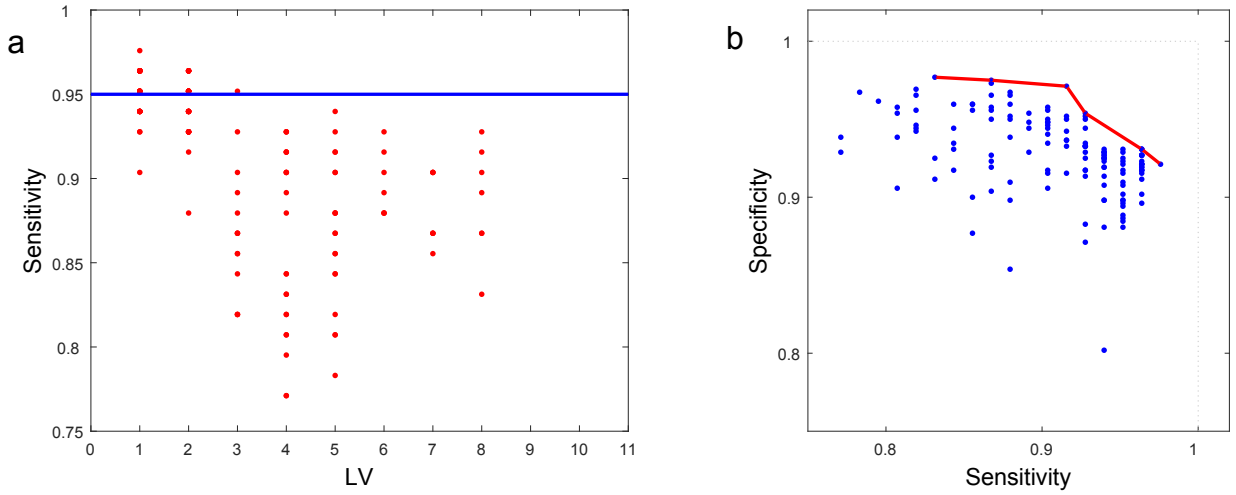

a
Sensitivity
LV
1
0.95
0.9
0.85
0.8
0.75
0
1
2
3
4
5
6
7
8
9
10
11
b
Specificity
Sensitivity
1
0.9
0.8
0.8
0.9
1

Table 1. Dataset *Olives*. Subset names, marks, and corresponding number of samples.

	Training class	**Internal test set**	**External test set**
Taggiasca	T1 ● 83	I1 ▲ 9	E1 ■ 19
Leccino	T2 ● 59	I2 ▲ 7	E2 ■ 6
Coquillo	T3 ● 45	I3 ▲ 5	–

Table 1. Dataset *Remedy* Subset names, marks, and corresponding number of samples.

Name	Training set	Test set
A	50 ▲	△ 20
B	40 ■	□ 10
C	70 ●	○ 30

Table 1. Sensitivity for *Olives* data

		α=0.05				α=0.01			
Method	**Strategy**	**Model parameters**	**T1**	**I1**	**E1**	**Model parameters**	**T1**	**I1**	**E1**
DD-SIMCA	Rigorous	*3 PCs*	0.95	1.00	0.89	*3 PCs*	1.00	1.00	1.00
	Compliant	*6 PCs*	0.93	1.00	0.79	*6 PCs*	0.99	1.00	0.89
PLS-DM	Rigorous	*L=1, a=0.3, k=2, autoscaled data*	0.95	1.00	0.95	*L=4, a=0.5, k=5, autoscaled data*	0.99	1.00	1.00
	Compliant	*L=1, a=0.8, k=2, autoscaled data*	0.98	1.00	1.00	*L=6, a=0.5, k=1, centred data*	0.94	0.89	0.82

Table 1. Total and partial specificities for *Olives* data.

		Total specificity				Partial specificity				
		α=0.01		α=0.05		α=0.01				
Method	**Strategy**	**Model parameters**	**Spec.**	**Model parameters**	**Spec.**	**T2**	**I2**	**E2**	**T3**	**I3**
DD-SIMCA	Rigorous	*3 PCs*	0.91	*3 PCs*	0.86	1.00	1.00	1.00	0.69	0.40
	Compliant	*6 PCs*	0.98	*6 PCs*	0.92	1.00	1.00	1.00	0.80	0.80
PLS-DM	Rigorous	*L=4, a=0.5, k=5, autoscaled data*	0.94	*L=1, a=0.3, k=2, autoscaled data*	0.89	1.00	1.00	1.00	0.74	0.60
	Compliant	*L=6, a=0.5, k=1, centred data*	0.94	*L=1, a=0.8, k=2, autoscaled data*	0.89	1.00	1.00	1.00	0.88	0.60

Table 1. Dataset *Remedy*. Model sensitivity ~~(%)~~ for the three classes. Specificity for all models is equal to 1.00 and, therefore, those values are not presented in the table.

	Class A			Class B			Class C		
	Model parameters	**Training**	**Test**	**Model parameters**	**Training**	**Test**	**Model parameters**	**Training**	**Test**
Method	**α=0.05**								
DD-SIMCA	*3 PCs*	0.94	0.90	*2 PCs*	0.98	0.90	*2 PCs*	0.97	0.90
PLS-DM	*L=2, k=1, a=0.6, scaled data*	0.96	1.00	*L=4, k=1, a=0.7, scaled data*	0.98	1.00	*L=3, k=4, a=0.6, scaled data*	0.96	0.60
	α=0.01								
DD-SIMCA	*3 PCs*	1.00	1.00	*2 PCs*	1.00	1.00	*2 PCs*	0.99	0.97
PLS-DM	*L=2, k=2, a=0.8, centred data*	1.00	1.00	*L=3, k=1, a=0.4, autoscaled data*	1.00	1.00	*L=5, k=2, a=0.6, scaled data*	0.98	0.60

Highlights

- Many qualitative analytical issues should be addressed by one-class classification
- We present ‘rigorous’ and ‘compliant’ approaches for model development
- In the ‘rigorous’ approach, only information from the target class is employed
- In the ‘compliant’ approach, the non-target class data is also utilised
- The pros and cons of both methods are critically discussed