Conférence CHIMIOMÉTRIE 2019
30 janvier 1er février
Montpellier SupAgro

Tuning the significance level of SIMCA models for reducing the impact of strong class overlap: a novel approach

R. Vitale[1,2], F. Marini[3], C. Ruckebusch[2]

[1] Molecular Imaging and Photonics Unit, Department of Chemistry, Katholieke Universiteit Leuven, Celestijnenlaan 200F, B-3001 Leuven, Belgium

[2] Laboratoire de Spectrochimie Infrarouge et Raman – UMR 8516, Université de Lille, Bâtiment C5, 59655 Villeneuve d'Ascq, France

[3] Department of Chemistry, Università degli Studi di Roma "La Sapienza", Piazzale Aldo Moro 5, 00185 Roma, Italy

Keywords: Class Modelling (CM), Soft Independent Modelling of Class Analogy (SIMCA), Receiver Operating Characteristic (ROC) curves.

1 Introduction

Nowadays, a large number of problems in fields like foodstuff origin authentication, quality control or process monitoring is addressed by Class Modelling (CM) statistical methods. Techniques such as UNEQual class modelling (UNEQ [1]) or Soft Independent Modelling of Class Analogy (SIMCA [2]) have been extensively used in the last decades for similar purposes. Contrarily to the more popular Discriminant Analysis (DA), the basic principle of CM is that classification rules are derived using only samples/objects belonging to a single target category. Faults in the definition of non-target categories, which could bias the classification performance, can thus be avoided.

Nevertheless, it is also well-known that if the classes under study present a high degree of overlap, CM approaches might suffer from severe limitations. In cases like this, properly adjusting the significance level of the resulting models can represent a potential solution to guarantee a better compromise between True Positive and True Negative rate. In this work, a new data-driven methodology that exploits the concept of Receiver Operating Characteristic (ROC) curve [3] is proposed to address such a task. Its only requirement is that measurements for samples belonging to non-target classes are also available. Although this is actually not strictly needed in the CM context, it can be highly beneficial in all situations in which significant overlapping exists between categories. This presentation explores the potential of this procedure as a possible way of tuning SIMCA model parameters in circumstances like this.

2 Theory

Let $\mathbf{X}$ be an N×J dataset constituted by Z submatrices $\mathbf{X}_z$ (N_z×J), each one containing the measurements collected for a single class of samples. In SIMCA, every category of objects is modelled independently from the others based on a Principal Component Analysis (PCA) model of appropriate dimensionality or complexity (say A_z) as:

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

where $\mathbf{T}_z$ (N_z×A_z), $\mathbf{P}_z$ (J×Az) and $\mathbf{E}_z$ (N_z×J) denote the scores, loadings and residuals matrices resulting from the decomposition of $\mathbf{X}_z$, respectively. Once defined the single class subspaces as in Equation 1, the degree of *outlyingness* of new unlabelled samples with respect to all of them can be estimated according to a combined index. For a generic observation $\mathbf{x}^{\mathrm{T}}_{\mathrm{new}}$ (1×J), this combined index is calculated as:

$$d_{\mathrm{new},z} = [(T^2_{\mathrm{new},z}/T^2_{\mathrm{lim},z})^2 + (Q_{\mathrm{new},z}/Q_{\mathrm{lim},z})^2]^{0.5} \tag{2}$$

being $T^2_{\mathrm{new},z}$ a statistic reflecting the (Mahalanobis) distance between the origin of the z-th model hyperplane and the projection of $\mathbf{x}^{\mathrm{T}}_{\mathrm{new}}$ onto it, $Q_{\mathrm{new},z}$ a statistic reflecting the perpendicular (orthogonal) distance between $\mathbf{x}^{\mathrm{T}}$ and the z-th model hyperplane, $T^2_{\mathrm{lim},z}$ an empirical threshold for the T^2_z-statistic (usually corresponding to a significance level of 95%) and $Q_{\mathrm{lim},z}$ an empirical threshold for the Q_z-statistic (usually corresponding to a significance level of 95%). $\mathbf{x}^{\mathrm{T}}_{\mathrm{new}}$ is considered an outlier for the z-th class model and, thus, rejected by it if $d_{\mathrm{new},z}$ is found to be larger than square root of 2. Otherwise, the investigated sample is recognised as part of the z-th category.

3 Material and methods

The algorithm proposed here allows both complexity (number of principal components) and significance level of a SIMCA model to be simultaneously tuned through the construction of cross-validated ROC curves. It will be compared to a more standard procedure for tuning SIMCA model parameters which was described in [4, 5] and that is based on fixing such a significance level *a priori*. The performance of the two methodologies will be assessed in terms of classification sensitivity, specificity and efficiency in external validation in 2 simulated and 4 real case-studies. The results will give a clear view of the robustness of SIMCA models in situations characterized by different degrees of class overlap.

4 Results and conclusions

Two interesting points arose from the analysis of the different handled case-studies:

- in cases of clear and definite separation among classes, the two aforementioned methodologies enabled a similar and equally satisfactory classification of unknown test samples;
- in the presence of strong overlap amongst classes, the implemented approach was found to lead to better classification efficiency in external validation compared to the more standard procedure based on a fixed significance level.

Therefore, it can be said that adequately tuning the significance level guarantees a classification that is more robust towards the dispersion of the target category (i.e. a better compromise between classification sensitivity and specificity may be achieved).

5 References

[1] Derde, M. P. & Massart, D. L. UNEQ: a disjoint modelling technique for pattern recognition based on normal distribution. *Anal. Chim. Acta* 184, 33-51, 1986.

[2] Wold, S. Pattern recognition by means of disjoint principal components models. *Pattern Recogn.* 8, 127-139, 1976.

[3] Vitale, R., Marini, F. & Ruckebusch, C. SIMCA modeling for overlapping classes: fixed or optimized decision threshold? *Anal. Chem.* 90, 10738-10747, 2018.

[4] Bevilacqua, M., Bucci, R., Magrì, A. D., Magrì, A. L. & Marini, F. Tracing the original of extra virgin olive oils by infrared spectroscopy and chemometrics: a case study. *Anal. Chim. Acta* 717, 39-51, 2012.

[5] Vitale, R., Bevilacqua, M., Bucci, R., Magrì, A. D., Magrì, A. L. & Marini, F. A rapid and non-invasive method for authenticating the origin of pistachio samples by NIR spectroscopy and chemometrics. *Chemometr. Intell. Lab.* 121, 90-99, 2013.