O1 CHEM2

A UNIFIED SIMCA FRAMEWORK FOR SINGLE AND MULTI-BLOCK DATA

F. Marini, A. Biancolillo

Department of Chemistry, University of Rome "La Sapienza", Rome, Italy

Modeling classification techniques, sometimes also called one-class classifiers, have several advantages over discriminant ones, especially when dealing with asymmetric problems, where there is only one category of interest [1]. Indeed, in class modeling, attention is focused on a single category at the time, whose class space is built only on the basis of the data collected on samples from that particular group. Classification is then carried out as an outlier detection problem: if a sample is found to be an outlier with respect to the class model (usually, according to a distance to the model criterion), is predicted as not belonging to the category under exam. Among the methods available in the literature for class modeling, soft independent modeling of class analogies (SIMCA) [2] is by far the most commonly used. In SIMCA, the systematic variability among samples belonging to the investigated category is captured by a principal component model of appropriate dimensionality so that the classification of unknown individuals is based on the definition of a distance to the model, which is calculated by combining residuals with a distance in the scores space, which is usually Mahalanobis-like.

However, when dealing with irregularly dispersed or, in general, moderately to highly heterogeneous classes, this may result in a shape of the model class space not corresponding to the actual one, so that high sensitivity can be achieved only at the price of low specificity and vice versa. In such situations, the use of a recently developed ROC-based approach to fine tune the classification thresholds [3] can help in finding the best model efficiency, but further improvements may be expected by redefining the way the class space itself is calculated.

In the present communication, the possibility of defining the scores distribution non-parametrically by means of a gaussian mixture model (potential functions) is presented. Gaussian mixture models approximate the probability density function of a distribution of samples as a linear combination of basis functions (normally triangular or gaussian), centred on the measured points:

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} c_i g(\boldsymbol{x} - \boldsymbol{x}_i)$$

$g(\boldsymbol{x} - \boldsymbol{x}_i)$ being a basis function centered in $\boldsymbol{x}_i$ and c_i the associated weight.

Such approach allows a more-tailored definition of the class space even in the case of severe deviations of the distribution of the class scores from normality, as shown in Figure 1, where the probability density function for the distribution of a set of scores on the first two principal components in a toy examples is shown. Class modeling is accomplished by identifying a threshold value of the potential (probability density function, $f(\boldsymbol{x})$), enclosing a fixed volume of the distribution, which is usually 95%. In practice, the estimation of this

O1 CHEM2

threshold value is usually accomplished by sorting the potentials of the training samples and identifying the desired percentile of the sorted values.

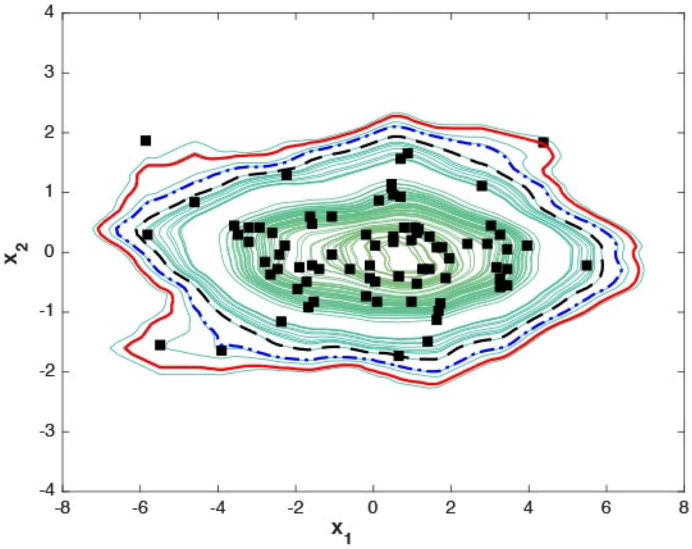

Figure 1. Use of potential functions to estimate the non-parametric scores distribution for a two component PC model. The red line indicates the threshold potential value for class acceptance (95th percentile of the distribution).

Due to its properties, this approach can easily be extended to the multi-block case in a framework which could be defined of mid-level data fusion, and could be applied on the scores calculated by means of different component models, not exclusively PCA.

The potential of the proposed approach will be illustrated by different examples involving food authentication both for the single- and the multi-block implementation.

References:

[1] Albano C., Dunn W. III, Edlund U., Johansson E., Nordén B., Sjöström M. and Wold S., Analytica Chimica Acta, 1978, 103, 429.
[2] Wold S. and Sjöström M. In: Kowalski B.R. (Ed.), Chemometrics: theory and application. ACS Symposium Series, vol. 52. American Chemical Society, Washington, DC, 1977, 243.
[3] Vitale R., Marini F. and Ruckebusch C., Analytical Chemistry, 2018, 90, 10738.