DATA FUSION STRATEGIES FOR FOOD AUTHENTICATION

Federico Marini[1]

[1] Dept. of Chemistry, University of Rome "La Sapienza", Rome, Italy, federico.marini@uniroma1.it

Abstract – More and more often food authentication problems require measuring fingerprinting data on the sample by different analytical platforms. To get full advantage of the multi-platform characterization of the products, the possibility of integrating the information provided by the different data blocks into a single, holistic model (data fusion) should be considered. In this present communication, the different strategies for performing data fusion will be presented and their advantages and drawbacks will be discussed and compared. The possible approaches will be illustrated by means of selected examples, involving different analytical platforms and food matrices (craft beer, extra virgin olive oils, almonds).

Keywords: chemometrics, data fusion, classification, partial least squares discriminant analysis (PLS-DA), food authentication

1. INTRODUCTION

it is well known that food is a complex matrix, whose composition may be affected by many different factors: geographical and species (botanical or animal) origin, pedo-climatic conditions, harvest/production year, manufacturing techniques, and so on. At the same time, those factors are also the ones affecting the quality of the food itself, so that quite often one of the goals in food quality control is authentication, i.e. the possibility of checking whether the product is compliant with what reported in the label [1].

In particular, in the context of authentication a particular role is played by the traceability issues, which translate to the quest for methods, which allow to track all movements of product and steps within the production process. This aspect involves also the need of verifying whether the samples were made by the mandatory raw materials and produced in specific geographical areas. So far, food traceability is enacted mainly by inspecting registries, barcodes, RFID tags or other tracking media, so that many efforts are posed in investigating the possibilities of achieving the same goals by means of more objective analytical techniques. In this respect, while only limited success has been obtained from the use of so-called primary indicators, i.e., specific chemical markers whose amount could be directly linked to the product history along its production chain, the most promising results are linked to the use of instrumental fingerprinting techniques coupled to chemometric data processing.

Instrumental fingerprinting techniques, indeed, provide an (often) untargeted characterization of the foodstuff, which in many cases may also be fast, relatively inexpensive and non-destructive/non-invasive. The resulting data may represent a valuable source of information, which however is not immediately available to the operators: there is the need of appropriate multivariate data analytical techniques to extract the meaningful information from the experimental data, which almost always contain a lot of "noise", to be intended as sources of variability not related to the response of interest, i.e. the authenticity of the product.

In this framework, one should stress that when the authentication problem is complex, involving, e.g. the possibility of differentiating samples under the coexistence of various defining factors (raw material/species, geographical origin, production year, etc.), then the use of a single fingerprinting technique may result to be insufficient and the products should be characterized by a multi-platform approach. Accordingly, from a chemometric standpoint, the set of sample is described by different data matrices, one for each technique used, each containing both common (shared with other blocks) and distinctive (unique) information. In this situation, the separate analysis of the individual data matrices is only a suboptimal solution, as it provides a partial and incomplete description of the products of interest: in order to take the maximum advantage from the availability

of different data matrices, the best option is to adopt strategies which integrate all the available blocks into a single, holistic model. These strategies are collectively referred to as "data fusion" [2]. In this communication, the most commonly adopted data fusion approaches will be discussed and illustrated by means of selected examples.

2. FOOD AUTHENTICATION THROUGH THE EYES OF A CHEMOMETRICIAN

When looked from a chemometrician's perspective, food authentication problems almost always fall into the domain of pattern recognition and, in particular, of multivariate classification. Indeed, classification is the procedure through which it is verified what is the category an unknown sample belongs to (discriminant approach) or whether a sample it is likely to belong to a particular group of objects (modeling approach).

In food authentication, one is often faced with the problem of checking if a sample's declared origin (in a broader sense, geographical, varietal, productive, etc.) corresponds to the truth, and therefore one could think of the possible origin as different categories to be considered for the study. For instance, one could describe a problem involving the traceability of extra virgin olive oils from a PDO origin as a classification issue where there are two categories: the PDO of interest and all other oils.

With these premises, discriminant or modeling classification can be accomplished through different approaches, the most suitable for the matrices resulting from high dimensional fingerprinting techniques being, respectively, partial least squares discriminant analysis (PLS-DA) and soft independent modeling of class analogies (SIMCA) [3]. PLS-DA is the classification analogue of PLS calibration and is based on calculating a regression model between the experimental matrix and a dummy Y matrix coding for class belonging. A sample is then assigned to one of the possible categories on the basis of its predicted Y values (in the easiest approach, simply by choosing the class corresponding to the highest component of predicted Y). On the other hand, SIMCA involves the calculation of a separate PCA model for each of the classes of interest. Verification of whether an unknown sample is accepted or not by the model of a particular class is then based on calculation of its distance to the model based on the figures of merit normally used for outlier detection and then setting a decision threshold for this same distance: if the distance of a sample to the model is lower than the selected threshold, it is accepted, otherwise it is rejected.

2.1. Data fusion

When multiple blocks of data are available, the information in the different matrices can be integrated through different strategies which, collectively, fall under the name of "data fusion" [2]. Commonly, data fusion strategies are differentiated according to the level at which fusion occurs into low-level, mid-level and high-level approaches.

In low-level data fusion, integration of the blocks occurs at the level of the pretreated data matrices: a joint data matrix is obtained by concatenating the different blocks after suitable preprocessing and then using this matrix for further data analysis. In general, it has the disadvantage that the gain in information achieved by the fusion is often overcome by the increase in the amount of noisy or irrelevant variables.

In mid level data fusion, features (selected variables or latent vectors) are extracted from the individual blocks and then concatenated to form the final data matrix to be processed by the chosen chemometric technique. Normally, it represents a good compromise in terms of information gain with respect to increase in the noise level.

Lastly, in high level data fusion, each data block is processed separately by the chosen chemometric technique in order to obtain the corresponding prediction. Then, the different prediction for each sample are combined to produce a final response, by means of voting (consensus) schemes or Bayesian techniques.

3. EXAMPLES

As discussed in the previous paragraphs, data fusion strategies are promising approaches to achieve food authentication, especially when the samples under investigation are characterized by different experimental techniques. In this communication, as an example of this paradigm, some selected examples will be presented, where the use of data fusion allowed improving the classification results for different food products.

A first example involved the authentication of an Italian craft beer (Reale by Birra del Borgo) based on

the outcomes of five fingerprinting techniques (UV, visible, mid-infrared and near infrared spectroscopies and thermogravimetry) [4]. In particular, the analyses have been conducted on 60 beer samples (19 "Reale", i.e. the specific beer to be authenticated, 12 beers from the same brewery, Birra del Borgo, and 29 beers of other origin). For the proper validation of the chemometric models developed, the sample set was divided into training (40 individuals) and test (20 objects) matrices, by means of an intelligent splitting algorithm, namely duplex [5]. The analysis was at first carried out on the individual matrices by means of PLS-DA, after suitable pretreatment. Apart from mean centering, which was the last pretreatment step in all cases, the following preprocessing approaches resulted to be optimal: SNV+detrending (MIR), MSC (NIR), none (UV and visible), first derivative (TG). When considering the results obtained by applying PLS-DA to the individual matrices, although a very good classification ability was obtained in the training set (most of the values being above 92%), in the validation stage, no data set allowed an overall predictive accuracy higher than 85%, the best results having been obtained by visible (100% on Reale and 82,7% on the others) and NIR (66.7% on Reale and 100% on the others).

On the other hand, the use of a low-level data fusion approach, after block-scaling, allowed to have a perfect classification of the training data and to increase the prediction rate on the test set to 100% for Reale and 92.3% for the other beers. A correct prediction of all the test samples was, instead, achieved by adopting a mid-level strategy. In this study, the mid-level approach was carried out by concatenating the scores of the individual PLS-DA model, after autoscaling. The classification of the test samples by the mid-level approach is graphically shown in Figure 1.

On the other hand, a similar approach was used to build a traceability model for the extra virgin olive oils of the PDO Sabina (Lazio, Italy) based on mid-infrared and near-infrared spectroscopies [6] or on HPLC-DAD profiles recorded at three different wavelengths [7].

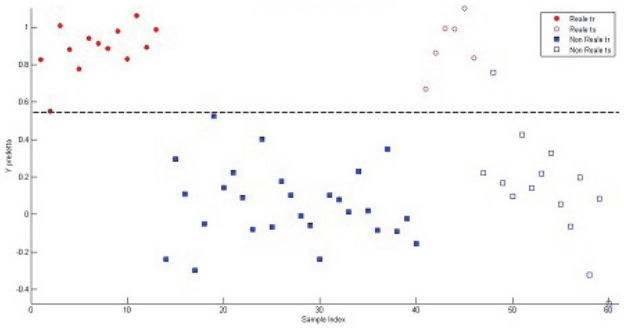

Fig. 1. PLS-DA on the beer data after mid-level data fusion: predicted Y values for the training and test samples.

In the former case, the analysis was carried out on 57 oil samples collected in the harvests 2009 and 2010: 20 samples came from the PDO area of Sabina (Lazio, Italy) while the remaining 37 were from other parts of Italy or from other Mediterranean countries [8]. Also in this case the data were divided into training and test to carry out model validation and the duplex algorithm [5] was used for this purpose. 35 samples were selected as the training set and the remaining 22 were left out to be the test set. All the samples were analyzed in reflectance mode (using an integrating sphere, in the case of near infrared spectroscopy, and by ATR, in the case of MIR). Although the results obtained by the analyses of the individual matrices were already very good (100% correct prediction on Sabina and 95% on other origins, when using NIR spectroscopic data), still it was checked whether they could be improved through a data fusion strategy. Mid-level data fusion was accomplished by concatenating the latent vectors obtained by PLS-DA modeling of the individual data blocks and it allowed the correct classification of all the validation samples.

As far as the HPLC analysis is concerned, the sample set used for the spectroscopic analyses decribed above was augmented with the addition of other 20 oils whose origin was other than Sabina. The samples were analyzed for their polyphenol content and the chromatographic fingerprint was recorded at three different wavelengths (254nm, 280 nm and 340 nm). In order to remove any source of spurious variation, the chromatographic profiles were pretreated by asymmetric least squares for baseline correction [9] and aligned using iCoshift [10]. Accordingly, three blocks of data were available, corresponding to the chromatograms recorded on the samples at the three different

wavlengths. At first, each chromatogram was analyzed independently and PLS-DA models were built and validated on a reduced set of variables, selected by a combination of backward interval PLS and genetic algorithms (biPLS-GA) [11]. When considering the classification ability, both the model built on the chromatograms recorded at 280 nm and the one built on the 340 nm data, performed equally well, leading to a non-error rate of 85.7% and 85.0% on Sabina and other oils, respectively.

In this case, mid-level data fusion was implemented by concatenating the experimental variables (chromatographic intensities) selected by the biPLS-GA approach, and all the possible combinations of two and three blocks were investigated. In particular, the best results were obtained by using only the information from the 280 nm and the 340 nm blocks, whose most relevant variables were fused, leading to an improve in the prediction ability, which reached 90% for the class of other oils.

From an authentication standpoint, the adoption of a data fusion strategy which involved a preliminary variable selection stage, allowed also to identify potential authenticity markers for the PDO Sabina. Indeed, based on the results of biPLS-GA-PLSDA, HPLC/ESI-MS analysis was carried out to interpret the results obtained after variable selection in terms of chemical species which could carry a discriminant information. Indeed, by means of tandem MS (either in positive or negative mode), it was possible to identify the analytes corresponding to the peaks selected by biPLS-GA, and these resulted to be the following substances: vanillic acid, p-coumaric acid, luteolin, pinoresinol, acetoxypinoresinol, apigenin, methoxyluteolin.

Lastly, data fusion was also used to authenticate PDO almonds from Avola coupling the information from mid and near infrared spectroscopies and thermogravimetry. In this context, the analysis of samples with three different instrumental techniques has allowed to investigate in detail the characteristics of the product and to characterize its variability compared to other almonds commercially available.

Almonds were analyzed as such and, subsequently, ground, with the aim of developing a method for the recognition of the type "Avola" and "Not Avola". The individual discriminant models, built with PLS-DA, allowed to obtain good recognition rates of Avola almonds, and a low number of false positives. In particular, by analyzing the results obtained with the individual techniques, the best models were obtained starting from the NIR spectra recorded on the samples after grinding. On the other hand, the use of a mid-level strategy, which allows to better integrate the information present in different blocks, has allowed to improve the separation between the classes (or their differentiation) and, on the training samples, to obtain 100% of correct classifications.

4. CONCLUSIONS

Data fusion strategies are powerful tools to integrate the information from different data blocks into a holistic model for food authentication. The different existing approaches allow versatility in the definition of the way the various blocks are fused and in trading off between the increase in the information content and the possibility of including higher amounts of irrelevant predictors.

REFERENCES

[1] G. P. Danezis, A. S. Tsagkaris, F. Camin, V. Brusic, C. A. Georgiou, “Food authentication: Techniques, trends & emerging approaches”, *TrAC Trends in Analytical Chemistry*, in press, March 2016, http://dx.doi.org/10.1016/j.trac.2016.02.026.

[2] E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, O. Busto, “Data fusion methodologies for food and beverage authentication and quality assessment – A review”, *Analytica Chimica Acta*, vol. 891, pp. 1-14, September 2015.

[3] M. Bevilacqua, R. Nescatelli, R. Bucci, A. D. Magrì, A. L. Magrì, F. Marini, “Chemometric Classification Techniques as a Tool for Solving Problems in Analytical Chemistry”, *Journal of the AOAC International*, vol. 97, nº. 1, pp. 19-28, January-February 2014.

[4] A. Biancolillo, R. Bucci, A. L. Magrì, A. D. Magrì, F. Marini, “Data-fusion for multiplatform characterization of an italian craft beer aimed at its authentication”, *Analytica Chimica Acta*, vol. 820, pp. 23-31, April 2014.

[5] R. D. Snee, “Validation of regression models: methods and examples”, *Technometrics*, vol. 19, pp. 415-428, 1977.

[6] M. Bevilacqua, R. Bucci, A.D. Magrì, A. L. Magrì, F. Marini, “Data fusion for food authentication. Combining near and mid infrared to trace the origin of extra virgin olive oils”, *NIR News*, vol. 24, nº. 2, pp. 12-15, March 2013.

[7] R. Nescatelli, R. C. Bonanni, R. Bucci, A.L. Magrì, A. D. Magrì, F. Marini, "Geographical traceability of extra virgin olive oils from Sabina PDO by chromatographic fingerprinting of the phenolic fraction coupled to chemometrics", *Chemometrics and Intelligent Laboratory Systems*, vol. 139, pp. 175-180, December 2014.

[8] M. Bevilacqua, R. Bucci, A. D. Magrì, A. L. Magrì, F. Marini, "Tracing the origin of extra virgin olive oils by infrared spectroscopy and chemometrics: A case study", *Analytica Chimica Acta*, vol. 717, pp. 39-51, March 2012.

[9] P.H.C. Eilers, "Parametric time warping", *Analytical Chemistry*, vol. 76, pp. 404–411, 2004.

[10] G. Tomasi, F. Savorani, S.B. Engelsen, "iCoshift: an effective tools for the alignment of chromatographic data", *Journal of Chromatography A*, vol. 1218, pp. 7832–7840, 2011.

[11] R. Leardi, L. Nørgård, "Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions", Journal of Chemometrics, vol. 18, pp. 486–497, 2004.