BUILDING GENERIC, EASILY-UPDATABLE CHEMOMETRIC MODELS WITH HARMONISATION AND AUGMENTATION FEATURES: THE CASE OF FTIR VEGETABLE OILS CLASSIFICATION

Konstantia Georgouli [1), Katerine Diaz-Chito[2), Jesus Martinez-del Rincon [3), Anastasios Koidis [1)

[1) Institute for Global Food Security, Queens University Belfast, Belfast, United Kingdom, t.koidis@qub.ac.uk
[2) Computer Vision Centre, Universidad Autonoma de Barcelona, Spain, kdiaz@cvc.uab.es
[3) Institute of Electronics, Communications and Information Technology, Queens University Belfast, UK j.martinez-del-rincon@qub.ac.uk

Abstract- Published literature in food authenticity studies is based on multivariate chemometric models that have been calibrated under controlled conditions using a limited dataset and a particular spectral acquisition instrument. There is a challenge to create accurate and robust chemometric models that would be able to perform well when tested with samples that have never been encountered by the calibration data and be applicable when the acquisition instrument is different from the initial instrument. Augmentation of the models with synthetic samples is a fresh approach to overcome these challenges. But even when a chemometric model is modified with the synthetic samples there is always the danger of overfitting it to the calibration set especially because limited new chemical information is added to the model using this technique. The only solution in this case is often the acquisition of more spectra from original authentic samples and retrain the models. The problem starts when original data are not readily available. In all these situations, it is clear that evolving a chemometric model may be a better solution than recreating or retraining it as a full new batch. This will only require access to the existing models and the new samples. In this paper we propose, therefore, two different approaches to tackle the challenges described earlier: a) a novel spectral data augmentation framework (DAF) in order to increase the performance of a typical classification model by generating realistic data augmented samples and b) a simple model updating framework for retraining models from large datasets.

The feasibility of the proposed DAF has been evaluated on three main different experiments where Fourier transform mid infrared (FT-IR) spectroscopic data of vegetable oils were used for the identification of vegetable oil species in oil admixtures.

Results demonstrate a significant ~40% improvement in classification when testing in more than 10 different spectroscopic instruments to the calibration one. On the other hand, the application of our novel model updating technique, called Incremental Generalized Discriminative Common Vectors (IGDCV) based on the same vegetable oil identification scenario, allowed for faster model creation while maintaining the same high accuracy. It is argued that using the combined approach of the DAF and IGDCV techniques can allow the generation of models that are applicable to the real world such as a spectroscopy sensor (NIR or Raman) in the food production floor, a tea processing facility or other examples.

Keywords: *Data augmentation, Incremental model learning, classification, vegetable oils, spectroscopy.*

1. INTRODUCTION

In the last decade the use of chemometrics in food analysis is steadily growing because the output of most analytical methods nowadays that is multivariate data matrices demanding appropriate chemometric analysis in order to capture the most important information in the data. Most studies are often limited to the use of well-known classification methods (e.g. PLS-DA and SIMCA). One of the problems faced by the research community is the generality of the models which limits their applicability in real world scenarios.

At first is the lack of large datasets and the small variation within the dataset (multiple authentic samples from various worldwide locations are required for every authentic food/ingredient type for a truly ‘global’ dataset). There is an increasing demand for larger and varied admixtures/samples

datasets. Acquiring a diverse amount of samples is a time consuming and costly process. In the field of food adulteration detection this challenge is even more obvious. Sourcing pure and authentic commodities as well as adulterants in order to construct the models can be a very challenging task [1] and the official and informal sources of true authentic samples (e.g. a rare spice or an exotic oil) are limited. In addition, to detect adulterants, current practice is to produce an appropriate number of in-house admixtures by mixing several commodity samples with one or more adulterants in different concentration grades. This allows for a robust classification/quantification model, but the number of combinations to be covered may become intractable and the whole process costly.
Data augmentation methods have been applied to multivariate calibration of spectroscopic data inorder to add sample variability only for a single lab validation. These methods were mainly basedon various types of 'noise' addition to the original data set before calibration [2]. Further studies have been conducted where noise augmentation methods were combined with ensemble or bootstrapping methods. All aforementioned methods have exhibited interesting results for multivariate regression and classification problems on spectroscopic data, however, they were closely linked to the application problem and require precise knowledge of the data in order to introduce the specific variation. This impacts their applicability. More sophisticated data augmentation can be achieved by not only manipulating each sample in isolation but also exploiting the relationships among samples such as the generation of artificialsamples.

In the literature only minor preliminary attempts have been done in food authentication studies inthis direction with multi-varietal food blends and sample identification.
On the other hand, for existing models as new spectral data come routinely in for subsequent analysis there is no simple way for the calibration models to become updated especially if the original data are not available. In order to enable safe exchange of the models between researchers as well
as avoiding the arduous task of the retraining a model using all the initial and new data a simple model updating framework is necessary. While model updating has been used and proposed in other fields, its intrinsic advantages have been scarcely exploited in the field of food analysis and chemometrics in tea analysis. However, these techniques are scarcely used in food science, reducing the impact of these incremental approaches, and they require huge amounts of calibration samples to generate the calibration models, which is unlikely for most food analysis scenarios.
In this paper we propose two techniques to overcome challenges of that researchers face with chemometric models based on spectroscopic data especially working in the area of food authenticity. Firstly, a novel data augmentation framework (DAF) that generalises previous preliminary attempts in the literature and allows the introduction of not only noise augmentation techniques but also artificial data blends or simulated acquisition instruments. Secondly, a modern technique, called Incremental Generalized Discriminative Common Vectors (IGDCV) that aims to simplify model updating by allowing the addition of new data samples and classes without recalculating the full projection or accessing the previously processed calibration data. Both new techniques are evaluated using vegetable oil type identification as case study brought into attention due to EU Regulation 1169/2011, which requires food manufacturers of processed foods that contain refined vegetable oil blends to name the oil types in the label [3].

1. EXPERIMENTAL

2.1. The proposed data augmentation framework

To cover for the different types of variability encountered (sample preparation, instrumental drifts, we designed and implemented a novel general framework for the application of the data augmentation techniques to spectra (see Fig. 1). This is a carefully designed pipeline of four data independent blocks which can be finely tuned depending on the desired variance for enhancing model's robustness: a) blending
spectra, b) changing the intensity, c) shifting along x axis, and d) adding noise. Each of the four blocks can be enabled either alone or in combination with the others. The blocks also have input parameters that allow to be applied in higher or lower degree depending on the expected variability in testing. When the spectrum of a sample is acquired (e.g. FTIR spectroscopy), it is passed to the data

augmentation framework where one or more samples (augmented samples) are generated from this particular one. The resulting original and augmented samples will then be passed to the chosen classification pipeline, where are preprocessed and used to calibrate the chemometric model.

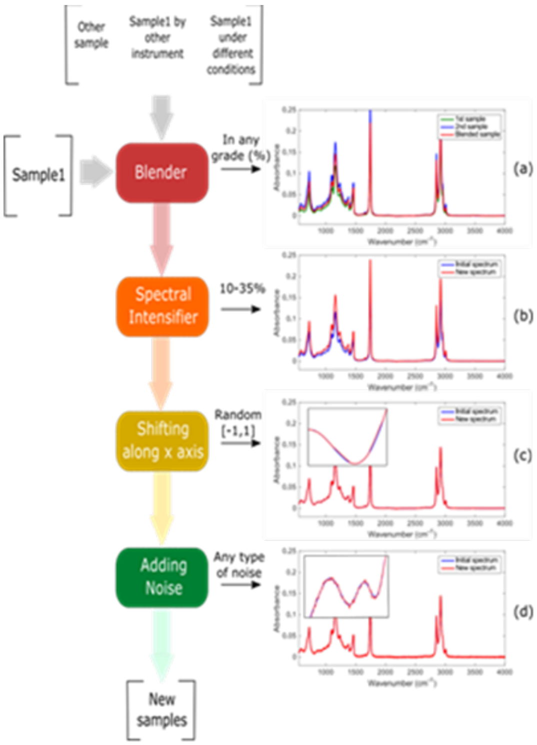


Figure 1. The proposed Data Augmentation Framework (DAF)

2.2. Validation of the DAF: the interlab trial

The rapid identification of vegetable oil species (Osorio et al., 2014) is used here as case of study to prove the potential of DAF. In this application, 6 different classes, comprised of 3 pure oil types and their corresponding binary admixtures, are characterised by their FTIR spectra (630 vegetable oils in total). The acquisition was typically performed in house using a Nicolet iS5 ATR FT-IR spectrometer (Thermo Fisher Scientific, Dublin, Ireland) equipped with a DTGS detector at 4000-550 cm-1 range (32 scans, 4 cm−1 resolution). To test if the DAF improves the model classification this experiment involved the use of 17 additional FTIR instruments from UK based research centres, public services and private food testing labs who agreed to take part in the study. The interlab validation samples sent to participants were independent set of 9 samples of mixed vegetable oil types including blends. The acquisition parameters were harmonised so that they are compatible with every FT-IR instrument. After FTIR collection, typical pre-treatment techniques were applied to all spectra. In total, 3781 variables between 654.23 and 1875.43 cm−1 and between 2520.02 and 3120.74 cm−1 were selected. Soft modelling of class analogy (SIMCA) as the modelling method and partial least squares discriminant analysis (PLS-DA) as a discriminant method were used for identifying vegetable oil admixtures for this experiment.

2.3 The IGDCV technique and its validation

Based on the subspace based learning theory, Generalized Discriminative Common Vector technique has been proved to provide discriminative subspaces for classification. Its incremental version (IGDCV) has been used in this study. After applying the IGDCV algorithm, samples are projected into a discriminative subspace where supervised learning based classification can be achieved. Since the subspace has been simplified as part of the algorithm, we have coupled our incremental subspace learning with a k-Nearest Neighbours (kNN) classifier in order to provide this functionality.

In order to evaluate the potential and advantages of IGDCV, the same case study as before was used. Three scenarios where the potential of incremental learning is relevant will be tested. In the first scenario, an oil type identification model is trained with a few calibration samples. After this initial calibration, new samples for each of the oil types to identify become available and are added to the model for improving the initial performance. In the second scenario, a simple model is initially trained to distinguish between just two oil types, and then extended to identify new oil types. In the third scenario, the oil type identification model created by a single lab and using a single FTIR instrument is extended and enhanced to be effective when used in other laboratories and instruments.

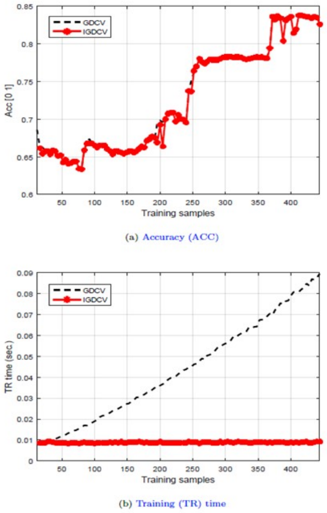
3. RESULTS AND DISCUSSION

3.1. Validation of the DAF

Early experiments demonstrated that the introduction of data augmented spectra following the DAF (Fig. 1) significantly improves the performance of the overall classification model by as high as 20% in a single lab validation (using the FTIR spectra from the in-house instrument in all cases). In the inter-laboratory trial, when the in-house calibration model was augmented by tuning the DAF components, there was a 20% improvement in classification rate when tested with validation set consisting of spectra from 17 different instruments. To move the concept further, the DAF was tested progressively using the same tuning parameters and in addition augmenting the calibration set with the addition of 'a virtual instrument' (blending spectra from two different FTIR instruments of the participants that were removed for the validation set). When all these were applied there was a significant improvement in the classification rate (>40%) achieving results as high as 90% compared with the baseline (without using the DAF). This demonstrated the clear advantages of using DAF to develop a more generic and accurate spectra-basedchemometric model.

3.2 Validation of the model updating framework

Figure 2 shows the results of both incremental and batch model updating methods for the same classification scenario, this time limited to one instrument as the purpose of the experiment is different than before. As expected, models perform better as more calibration samples are available for learning from. Regarding the incremental learning, it can be observed how the accuracy of the incremental approach remains the same (when compared with the batch algorithm) despite not having access to the initial samples but only to the previous model. Moreover, when comparing the computational time required to generate the models (Fig. 2b), one can notice the great difference in efficiency of using an incremental method instead of regenerating larger and larger models from scratch. Further experiments adding new classes and new instruments were followed. The final results highlight the strong potential of the IGDCV technique in the authentication studies.



(a) Accuracy (ACC)

(b) Training (TR) time

Figure 2. Batch GDCV and incremental IGDCV methods regarding new samples.

AKNOWLEDGMENTS

This research was supported with funding from The Department Learning and Employment Northern Ireland (DELNI) and the DEFRA of the UK (Grant no. FAO 157). The authors would like to thanks all UK participants of the interlab trial.

REFERENCES

[1] Zhao, J., Lin, H., Chen, Q., Huang, X., Sun, Z., Zhou, F. (2010). Identification of egg's freshness using NIR and support vector data description.*Journal of food Engineering*, 98(4), 408-414.

[2] Conlin, A. K., Martin, E. B., Morris, A. J. (1998). Data augmentation: an alternative approach to the analysis of spectroscopic data. Chemometrics and intelligent laboratory systems, 44(1), 161-173.

[3] Osorio, S. Haughey, C. Elliott, A. Koidis, Identification of vegetable oil botanical speciation in refined vegetable oil blends using an innovative combination of chromatographic and spectroscopic techniques, Food Chemistry 189 (2015) 67-73.