Chemometrics and Intelligent Laboratory Systems 167 (2017) 123–131

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Incremental model learning for spectroscopy-based food analysis

Katerine Diaz-Chito [a], Konstantia Georgouli [b], Anastasios Koidis [b], Jesus Martinez del Rincon [c,*]

[a] Computer Vision Centre, Universidad Autonoma de Barcelona, Spain
[b] Institute for Global Food Security, Queens University Belfast, UK
[c] Institute of Electronics, Communications and Information Technology, Queens University Belfast, UK

ARTICLE INFO

2010 MSC:
00-01
99-00

Keywords:
Incremental model learning
IGDCV technique
Subspace based learning
Identification
Vegetable oils
FT-IR spectroscopy

ABSTRACT

In this paper we propose the use of incremental learning for creating and improving multivariate analysis models in the field of chemometrics of spectral data. As main advantages, our proposed incremental subspace-based learning allows creating models faster, progressively improving previously created models and sharing them between laboratories and institutions without requiring transferring or disclosing individual spectra samples. In particular, our approach allows to improve the generalization and adaptability of previously generated models with a few new spectral samples to be applicable to real-world situations. The potential of our approach is demonstrated using vegetable oil type identification based on spectroscopic data as case study. Results show how incremental models maintain the accuracy of batch learning methodologies while reducing their computational cost and handicaps.

1. Introduction

In the last decade the use of chemometrics in food analysis is steadily growing. This is caused because the output of most analytical methods is nowadays multivariate data matrices (spectroscopic, chromatographic/mass spectrometry data, isotopic, sensorial, etc) which cannot be manually analysed and demand appropriate chemometric analysis in order to process and capture the most important and relevant information in the data. Selection of multivariate methods (e.g. classification methods) however is often limited to a set of well known standard methods (e.g. PLS-DA and SIMCA classification methods) and researchers are faced with some persisting problem with the chemometric models that they generate [1].

Among these problems that must be addressed, the generality of the models created to new conditions is the most important one. While extensive research has been done to create models under controlled conditions, for a small problem or dataset, the applicability of those models in real world -e.g. in food testing in the food industry or in routine analysis in a regulated testing laboratory- is very scarce. This is due to the overfitting of the model to the calibration set when only one instrument, one analytical laboratory or, in general, one set of assumptions are taken into consideration to create the models. Thus, when these models are tested in other slightly different conditions, they report much lower performances than the expected one. Recalibrating or recreating similar models to work in those situations may be an extremely arduous task, with a similar time and effort scale to the design, and tuning of the first model.

To avoid a full recalibration, model updating and calibration transfer techniques have been proposed to cover the transfer of multivariate classification models between different spectrometers [2,3], temperatures [3,4], harvesting seasons [4] and even different geographical regions [5]. Calibration transfer techniques [2] allow mapping the new spectra to the primary model spectra domain by calculating a transformation matrix from one domain to the other. Different calibration transfer techniques have been recently explored in chemical sensor arrays to overcome inherent sensor variability [6–8]. Only a small set of samples are required to be measured in both the primary and secondary conditions. However, in many applications it is not realistic that exactly the same sample can be measured, e.g. the same food sample from two different geographical locations. More interesting are methods based on model updating by augmenting sample spectra from a new condition. While many sample would normally be required to span to the new conditions [4], which amounts to a full recalibration, approaches based on Tikhonov regularisation (TR) [3,5] only needs a few samples to update the model. As disadvantage, TR still requires access to the initial samples to recompute the updated model, with the consequent computational cost of involving all samples in the optimisation, and its

* Corresponding author.
E-mail addresses: kdiaz@cvc.uab.es (K. Diaz-Chito), kgeorgouli01@qub.ac.uk (K. Georgouli), t.koidis@qub.ac.uk (A. Koidis), j.martinez-del-rincon@qub.ac.uk (J. Martinez del Rincon).

http://dx.doi.org/10.1016/j.chemolab.2017.06.002
Received 7 December 2016; Received in revised form 24 May 2017; Accepted 1 June 2017
Available online 9 June 2017

performance heavily relies on a meta-parameter that controls the balance between the initial model and the augmented samples, and which can only be tuned empirically. Finally, some recursive learning approaches [9,10] propose a framework where both incremental and decremental stages are used to improve the initial model. However, to fully exploit their potential and being able to remove old samples, access to the initial samples is also required.

Moreover, new samples are analysed on a routine basis and new data is generated including cases when new component classes are needed to be created (in authentication/adulteration studies, in traceability, proximate analysis prediction etc). As a result, existing and validated models may stop being useful and/or applicable. It is then necessary to retrain them. However, this requires access to the original samples, which may be lost or unavailable. Similarly, if an external laboratory, or other third party such as a company or an institution wishes to improve an existing model, the access to the original samples may be tricky or impossible, with privacy or confidentiality issues playing a role. In all these previously described 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. It will also be a more efficient manner to store the information, reducing the memory and physical space required and it can potentially decrease the time to create an improved model.

While incremental learning has been used and proposed in other fields [9–13], its intrinsic advantages have been scarcely exploited in the field of food analysis and chemometrics [14–19]. Bhattacharyya et al. [14,15] applied neural networks for identification of seven different black tea classes. Their incremental approach allow to add new classes of black tea to the original set. In Tudu et al. (2009) [16], the same researchers applied incremental fuzzy logic to the black tea identification. Cernuda et al. [17–19] proposed a flexible fuzzy inference system for the monitor of the concentration of sulphuric acid (H_2SO_4), sodium sulfate (Na_2SO_4) and zinc sulfate ($ZnSO_4$) in viscose production and in the melamine resin production process, which allows online adaptation of parameters and structural changes in the model. However, techniques based on neural networks and fuzzy logic 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 aim to extend the use of incremental learning in the field of food analysis and chemometrics. Among the variety of incremental learning techniques, we have chosen subspace based learning as the family of machine learning to apply due to their proved ability to evolve online [13], the ability to generate efficient models using a reduced number of calibration samples, and the extensive use of some of the basic subspace based methods such as Principal Component Analysis (PCA), and Soft independent modelling of class analogies (SIMCA)- in food science [20,21], both for exploratory analysis [22] and classification [23–25]. Thus, the present work introduces the use of an incremental subspace based learning technique, called Incremental Generalized Discriminative Common Vectors (IGDCV), which allows efficiently adding new data samples and classes to a knowledge base. In this way, our methodology is able to update the model to the new scenario without recalculating the full projection or accessing the previously processed calibration data, while retaining the previously acquired knowledge. Our approach is evaluated using vegetable oil type identification [22,26–28] as case study and results are compared against a non incremental learning technique, i.e. an equivalent batch method. Three different incremental scenarios are tested in this application area: when new samples are available to improve the model, when new classes must be identified by the model, and when new instruments are used in the identification process.

2. Incremental learning framework

Several incremental feature extraction based on linear subspace methods have been proposed and used on many practical applications. Among them, we find the Incremental approaches of the PCA [29], Linear Discriminant Analysis (LDA) [30] and DCV [31]. While PCA-based incremental approaches are simple and versatile, they are not optimal for discrimination and classification purposes since no class information is used to obtain principal components which may lead to unsuited subspaces. On the contrary, LDA is a supervised technique which makes use of the class information to obtain the most discriminative space by maximizing the distance between classes while minimizing the distance between the samples within the same class. However, LDA-based approaches cannot be applied when the dimension of the sample space is larger than the number of samples in the calibration set, since the within-class scatter matrix will be singular. This problem is known as the *Small Sample Size* SSS problem [32], and it is frequent in spectroscopic and chromatographic application, where the number of variables per sample is in the order of thousands while the total number of samples used for calibration rarely goes above the hundreds [22].

Among the approaches that have been proposed to solve the SSS problem, the Generalized Discriminative Common Vectors (GDCV) has been proved [13] to provide discriminative subspaces for classification regardless of the SSS assumption. GDCV is a variation of LDA [33,34] which introduces the idea of approximate extended null and reduced range subspaces of the within-class scatter matrix. Given the good performance of GDCV batch approaches, we proposed the use of Incremental GDCV [13] as the base of our online learning framework for food analysis, where new information is added while retaining the previously acquired knowledge, without accessing the previously processed calibration data.

2.1. IGDCV

Formally, let the calibration set X be composed of c classes, where every class j has m_j samples. The total number of samples in the calibration set is $M = \sum_{j=1}^{c} m_j$. Let x_j^i be a d-dimensional column vector which denotes the i^{th} sample from the j^{th} class. The within-class scatter matrix, S_w^X, is defined as,

$$S_w^X = \sum_{j=1}^{c} \sum_{i=1}^{m_j} \left(x_j^i - \bar{x}_j\right)\left(x_j^i - \bar{x}_j\right)^T = X_c X_c^T \tag{1}$$

where $\bar{x}_j$ is the average of the samples in the j^{th} class, and the centered data matrix, X_c consists of column vectors $(x_j^i - \bar{x}_j)$ for all $j = 1 \ldots c$ and $i = 1 \ldots m_j$.

The extension of the null space of S_w^X (which implies restricting the corresponding range space) is done from the Eigen-Value Decomposition (EVD) of S_w^X.

$$EVD\left(S_w^X\right) : U_r \Lambda_r U_r^T \tag{2}$$

where $U_r \in \mathbb{R}^{d \times r}$ are the eigenvectors associated to the nonzero eigenvalues Λ_r. The scattering added to the null space can be measured as the trace $tr(U_\alpha^T S_w^X U_\alpha)$. This quantity is up to $tr(S_w^X)$ when no directions are removed, $U_\alpha = U_r$, and decreases as more and more important directions disappear from U_r. Consequently, the scattering preserved after a projection, U_α, can be written as follows

$$\alpha = 1 - \frac{tr\left(U_\alpha^T S_w^X U_\alpha\right)}{tr\left(S_w^X\right)} \tag{3}$$

The projection basis fulfilling the above conditions for a given value of α can be obtained through U_r, such that r is reassigned. The α value is the main parameter of GDCV, which can be tuned by using cross-validation over the training set. The GDCV method can be the summarized as