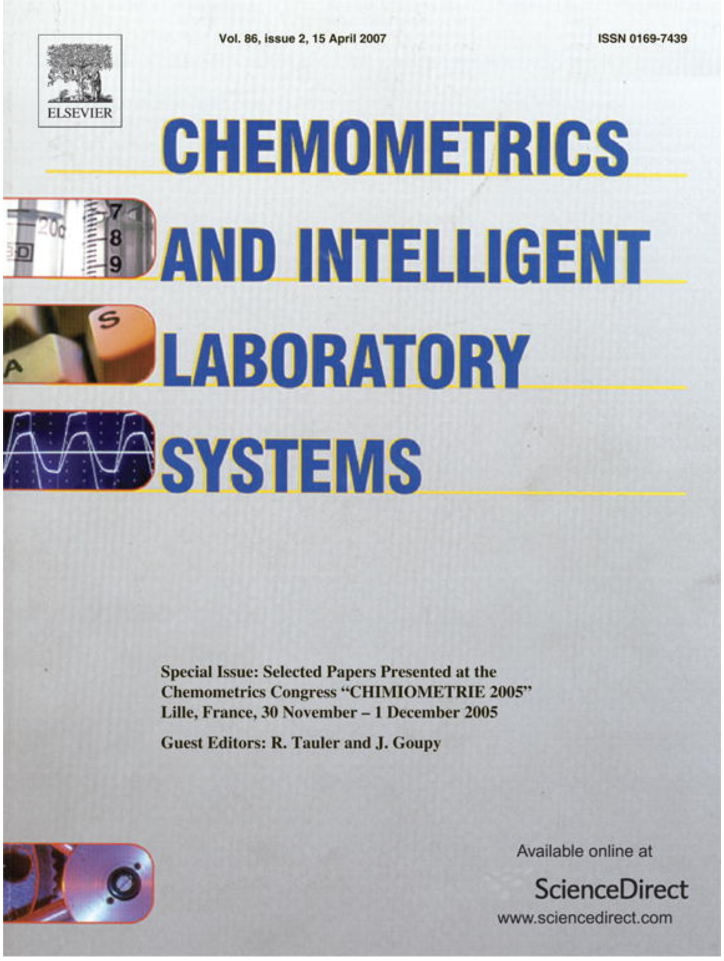

Vol. 86, issue 2, 15 April 2007

ISSN 0169-7439

ELSEVIER

CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS

Special Issue: Selected Papers Presented at the
Chemometrics Congress "CHIMIOMETRIE 2005"
Lille, France, 30 November – 1 December 2005

Guest Editors: R. Tauler and J. Goupy

Available online at
ScienceDirect
www.sciencedirect.com

Available online at www.sciencedirect.com

ScienceDirect

Chemometrics and Intelligent Laboratory Systems 86 (2007) 219 – 223

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

Chemometric contest at '*Chimiométrie 2005*': A discrimination study

Juan Antonio Fernández Pierna *, Pierre Dardenne

Walloon Agricultural Research Centre (CRA-W), Quality of Agricultural Products Department, Chaussée de Namur no. 24, 5030 Gembloux, Belgium

Received 15 January 2006; received in revised form 8 June 2006; accepted 8 June 2006
Available online 14 July 2006

Abstract

Due to the success of the chemometric challenge organized within the framework of the congress 'Chimiométrie 2004' ['A NIR data set is the object of a Chemometric contest at 'Chimiométrie 2004''. P. Dardenne, J.A. Fernández Pierna, Accepted for publication in to Chemometrics and Intelligent Laboratory Systems, special issue Chimiométrie (2004)], the organization committee decided to re-launch the idea for the 2005 edition (http://www.chimiometrie.org/) held in Lille, France (30th November and 1st December) by proposing through its website another data set. This data set was selected in order to test the ability of the participants for using discrimination methods based on IR data for the classification of starches according to the type of chemical modification undergone. As the previous edition, the participants were asked to present during the conference their own approaches. The committee received only two answers, which shows the difficulties that the other participants found by using this data set. This paper summarizes the two approaches proposed by the participants and the proposed approach of the authors.

Keywords: Discrimination; SVM; Mid-IR; Starch; Chemometrics

1. The data of the challenge

Two data sets were available on the website of the conference (http://www.chimiometrie.org/), which includes a calibration data set (Chemo_cal) and a test set (Chemo_test). The Chemo_cal data included 215 MIR spectra of starches of four different classes taken on a Perkin-Elmer Spectrum 2000 FTIR spectrometer (Perkin Elmer Corporation, Norwalk, CT, USA) between 4000 and 600 cm^{-1} at 1 cm^{-1} data interval. The Chemo_test set contained 43 samples measured under the same conditions as the calibration set.

The data sets used in this challenge, and therefore in this paper, are published and explained in [1]. Some samples from the original data set were removed in order to facilitate the treatment of the data. Moreover, for the challenge some modifications were performed in the data in order to increase the difficulty of the discrimination. These modifications consisted in the creation of some outliers as it was performed in the previous edition of the congress [2].

The aim for the participants was to discriminate as accurate as possible between the four different classes included in the *y* matrix and then to predict the test set blind spectra using any kind of method and send to the jury a text file or a slide presentation with the proposed methodology and the predicted results.

2. Deliberate modifications in the data sets

The calibration set stayed unmodified and in the test set some spectral data were slightly modified. The modifications were included in order to prove the ability of the participants to find these outliers. In total 4 outliers were created.

2.1. First modification

The first spectral modification was to add a shift to one existing sample and to create a new one. The first 6 data points of spectrum 1 were removed and six new variables corresponding to the last data point were added at the end of the spectrum producing a shift. This new spectrum corresponds to sample 2 and it can be seen in Fig. 1 where a zoom of the spectrum is shown.

2.2. Second modification

This modification was performed in the original spectrum 2 and a new sample 4 was created. Heteroscedastic noise is

[*] Corresponding author.
E-mail address: fernandez@cra.wallonie.be (J.A. Fernández Pierna).

0169-7439/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2006.06.009

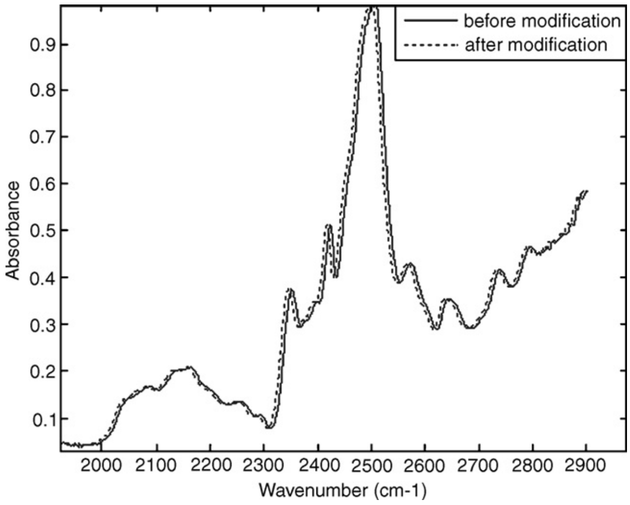

Fig. 1. Zoom of sample 2 before and after the first modification.

simulated by first generating gaussian white noise, and then multiplying this noise by the absorbance values of sample 2. Noise spectra are reconstructed by adding the heteroscedastic noise to the original signal. Figs. 2 and 3 show the spectra of sample 4 before and after this modification respectively.

This outlier can be found by applying a first derivative. The derivative must be done with a gap derivative algorithm with no smoothing and not by a Savitsky-Golay function. The trick is to subtract the consecutive data points. Fig. 4 shows the spectra after the derivative and it is evident that sample 4 (grey) becomes easily detected.

2.3. Third modification

This spectral modification was to change a data point of one spectrum to simulate a spike. Data point 2456 of spectrum 39 was changed from 0.5474 to 0.65 and it became sample number 43 (Fig. 5).

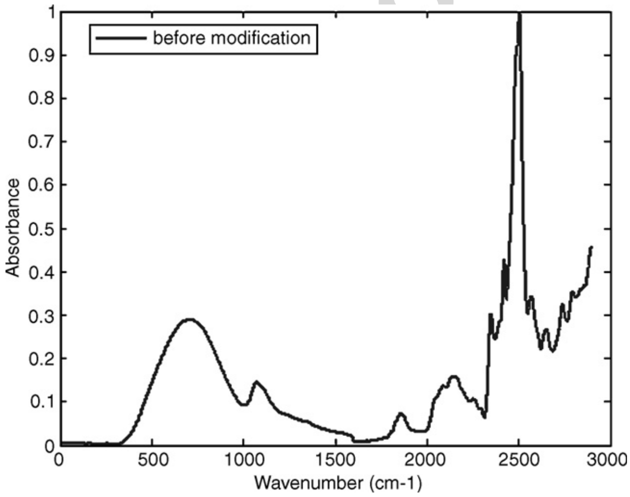

Fig. 2. Sample 4 before and after the second modification.

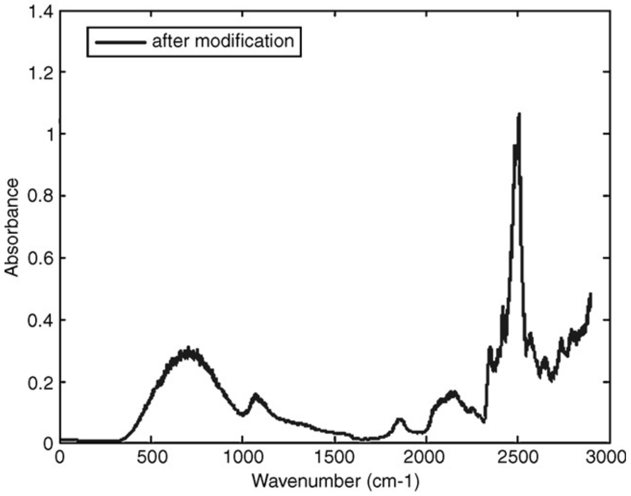

Fig. 3. Sample 4 before and after the second modification.

This outlier could also be detected as the same way as the previous outlier, i.e. by subtracting consecutive data points, after removing sample 4. The results are shown in Fig. 6.

2.4. Fourth modification

The last change was to add a slope to sample 17 and create sample 20. Fig. 7 shows the spectrum of the modified sample before and after the change.

By using the tools for outlier detection included in the PLS toolbox for Matlab, it is possible to calculate a lack of fit statistic for PCA models, Q. Q is simply the sum of squares of the residuals (E) for the PCA model. This Q statistic indicates how well each sample conforms to the PCA model. It is a measure of the difference, or *residual* between a sample and its projection into the k principal components retained in the model. As shown in Fig. 8 the Q residual test detects outliers 2, 4 and 20 but not 43 (spike). But sample 43

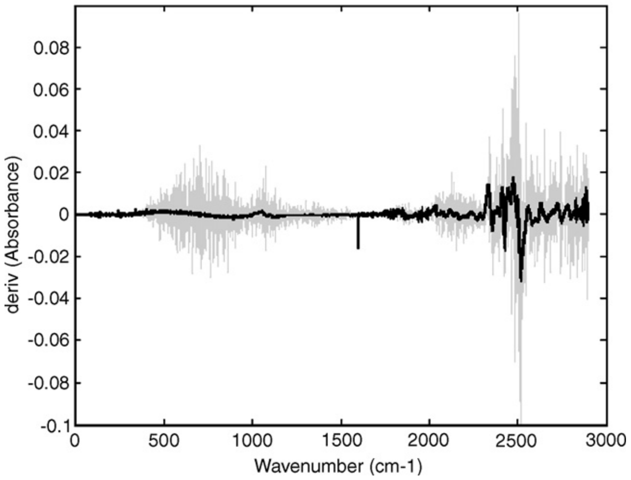

Fig. 4. Results of the derivative putting in evidence sample no. 4.

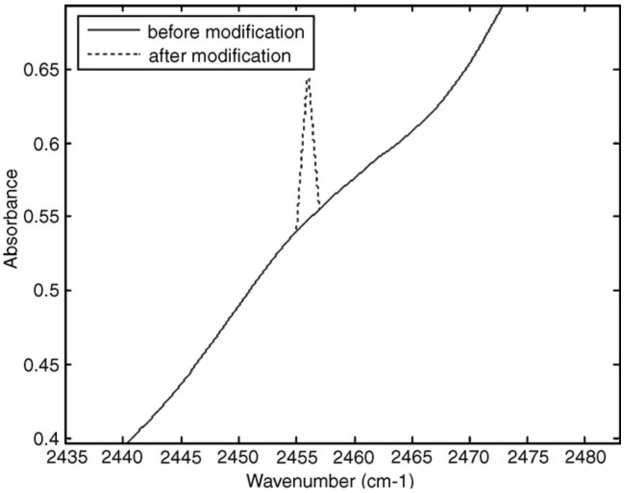

Fig. 5. Sample 43 before and after the third modification.

can be easily detected to the large spike that produces the data.

3. Participant's approaches

3.1. Participant no. 1

In order to detect outliers PCA was performed after centering the data. Then, the scores have been used to calculate the Mahalanobis distances between the samples and the center of each class. Two samples were removed after this test. Three outliers (2, 4 and 20) were easily found. This study has also permitted the participant to show that classes 3 and 4 seem to be easily discriminated from the other two classes and that classes 2 and 3 present similar structure and therefore are difficult to discriminate.

Once the outlier detection has been performed, the discrimination method used follows a hierarchical approach: a first model is constructed in order to discriminate between three classes: (1+2), (3) and (4). Then, a second model is used to discriminate between class 1 and class 2. For the construction of these 2 models, PLS-DA by using leave-one-out has been chosen.

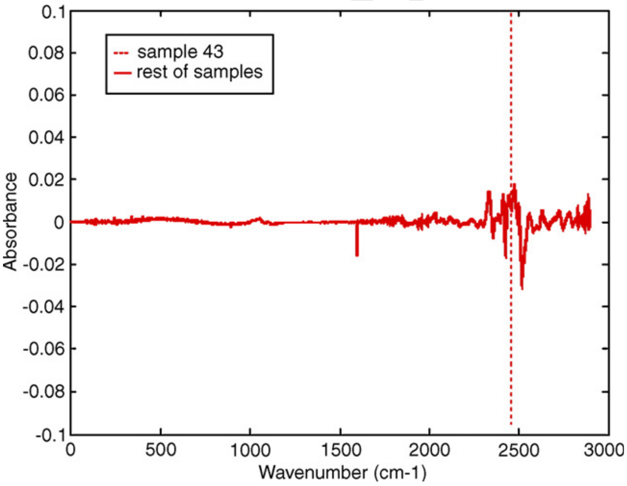

Fig. 6. Results of the derivative putting in evidence sample no. 43.

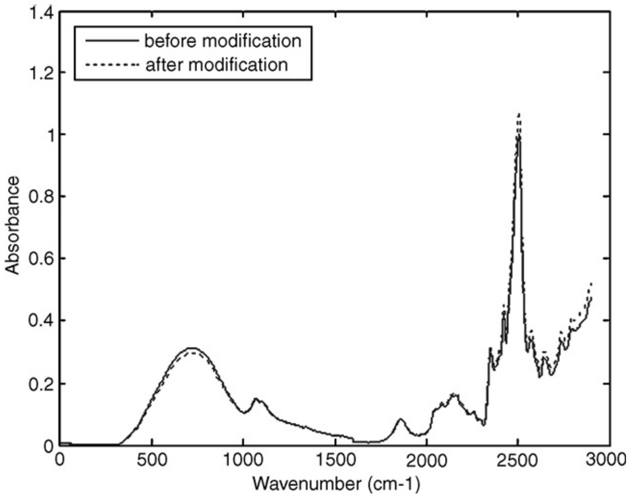

Fig. 7. Sample 20 before and after the fourth modification.

3.2. Participant no. 2

In a first step variable selection has been applied based on the importance of these variables in a PLS model. 72 variables have been selected as representative. Then, 3 different models were constructed: ROC, PLS and SVM. The final conclusion is done by taking a combination of the predictions performed by the three models by vote by taking into account the error for each class of each model. No outlier detection has been performed.

4. The author's approaches

The authors tested two approaches: PLS-DA and SVM [3]. Because the procedure has been already explained in [1], only the main characteristics of the models and the results are shown

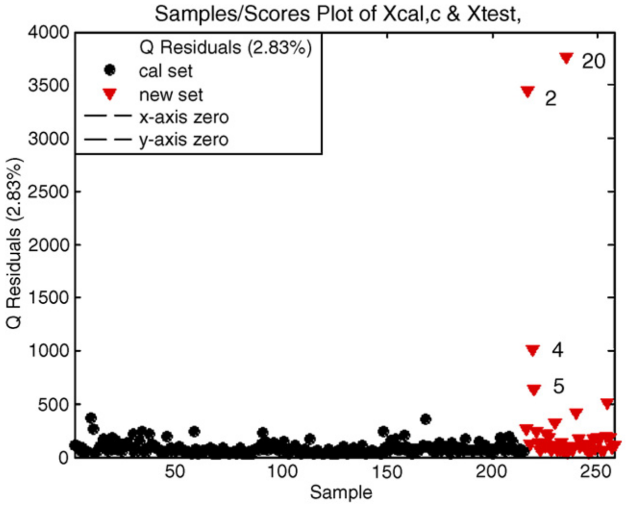

Fig. 8. Q residual test to detect outliers.

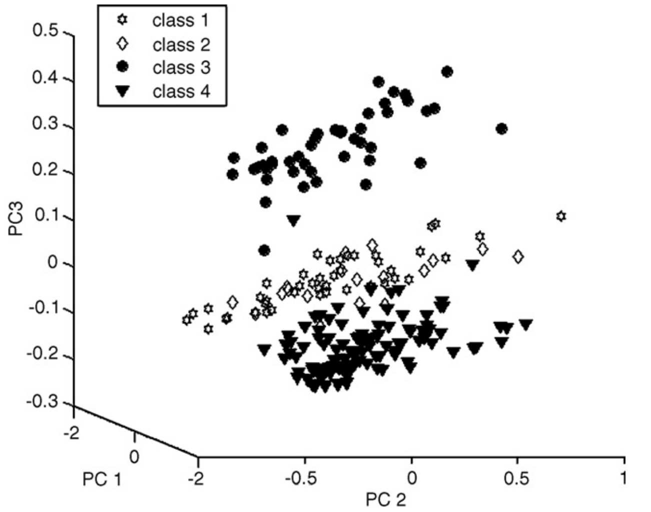

Fig. 9. Three-dimensional plot showing the results of the scores obtained using PCA (PC1 vs. PC2 vs. PC3).

here. For PLS-DA, in the calibration set, leave-one-out cross-validation was carried out to optimise the model, i.e., to find the number of variables, with which one obtains the best classification rate. To do that, PCA combined with a Fisher criterion (FC) was performed. The Fisher criterion describes the ratio of 'between-class' variance to 'within-class' variance that is helpful to decide which variables have an important discriminating power. In this study the PCs were ranked according to the FC in order to decide which PCs have an important discriminating power. Then, the number of latent variables was selected as the one that drove to the minimum RMSE (root-mean-square error). For SVM, the parameters were optimised using a separate training and validation subsets split from the calibration data using the Kennard-Stone method. This optimisation was performed using the grid-search technique [3]. The optimal parameter settings were selected as the values that give the minimum RMSE and the maximum classification rate. The classification or success rate is defined as:

$$\text{Success rate} = \frac{\left(\sum_{i=1}^{K} \frac{\text{Correctly classified samples in class } i}{\text{Total number of samples in class } i}\right)}{K}$$

with K being the number of classes. A success rate of, for instance, 0.9012 indicates that 90.12% of the objects are correctly classified.

5. Results

Fig. 9 shows a three-dimensional plot showing the results of the scores obtained using PCA. It becomes evident the difficulties of separating classes 1 and 2. Table 1 shows a summary of all the results obtained during the challenge. The first two columns represent the results obtained by the two participants, and the four last columns to the results for the solutions proposed by the authors.

Participant 1 found 3 of the 4 outliers proposed (2, 4 and 20) using PCA followed by a distance test. Then he built a model between three classes: (1+2), (3) and (4) obtaining a classification rate of 99.5%. When a second model is constructed, i.e. a model between class 1 and class 2, the classification rate decreases to 90.2%. But when a conclusion has to be taken, only 80% of the samples are correctly classified.

Participant 2 did not performed any outlier detection method and all the samples were predicted. His procedure of combining three different models produces a correct classification rate of 90.7%.

The procedures proposed by the authors drive to the conclusion that the best results are those obtained with the SVM model where, even if no outliers detection techniques were applied, the classification rate is larger than 93%.

6. Conclusion

Even if outlier detection is one of the most important task in chemometrics, it can be observed, as last year, that this step was not very deep. For this data set the outliers created have no large influence in the classification rate, but the aim of this challenge was to test the different techniques of the participants not only for discrimination but also for data treatment in general. Unfortunately only one participant applied (satisfactorily) some techniques for outlier detection and mentioned the modifications made on the spectra.

The session with the "contest" presentations of these results interested most of the participants and the final conclusion was that it will be repeated during the next conference.

Acknowledgments

We would like to thank all the participants who spent time to treat the data and present their results: Dr. S. Skander, Pertinence and Dr. J.M. Roger, Cemagref.

Table 1
Summary of all the results obtained during the challenge

	Participant number 1	Participant number 2	Author approaches			
Method used	PLS-DA	ROC+PLS+SVM	SVM	SVM	PLS-DA	PLS-DA
Model (s)	(1+2)-3-4 1-2	1-2-3-4	1-2-3-4	1-2-3-4	1-2-3-4	(1+2)-3-4-1-2
Outliers removed	3	0	0	4	0	0
Test samples to be predicted	40	43	43	39	43	43
Test samples correctly predicted	32	39	40	37	37	33
% test samples correctly predicted	80	90.7	93	94.9	86	76.7

References

[1] J.A. Fernández Pierna, P. Volery, R. Besson, V. Baeten, P. Dardenne, Classification of modified starches by FTIR spectroscopy using support vector machines, Journal of Agricultural and Food Chemistry 53 (17) (2005) 6581–6585.

[2] 'A NIR data set is the object of a Chemometric contest at 'Chimiométrie 2004''. P. Dardenne, J.A. Fernández Pierna, Accepted for publication in to Chemometrics and Intelligent Laboratory Systems, special issue Chimiométrie (2004).

[3] J.A. Fernández Pierna, V. Baeten, A. Michotte Renier, R.P. Cogdill, P. Dardenne, Combination of Support Vector Machines (SVM) and Near Infrared (NIR) imaging spectroscopy for the detection of meat and bone meat (MBM) in compound feeds, Journal of Chemometrics 18 (2004) 341–349.