Talanta 104 (2013) 128–134

Contents lists available at SciVerse ScienceDirect

Talanta

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

Short communication

Feature selection strategies for quality screening of diesel samples by infrared spectrometry and linear discriminant analysis

Mohammadreza Khanmohammadi [a,*], Amir Bagheri Garmarudi [a,b], Miguel de la Guardia [c]

[a] Chemistry Department, Faculty of Science, IKIU, Qazvin, Iran
[b] Department of Chemistry & Polymer Laboratories, Engineering Research Institute, Tehran, Iran
[c] Department of Analytical Chemistry, University of Valencia, 50 Dr. Moliner Street, E-46100 Burjassot, Valencia, Spain

ARTICLE INFO

Article history:
Received 14 August 2012
Received in revised form
11 November 2012
Accepted 12 November 2012
Available online 20 November 2012

Keywords:
Diesel
ATR-FTIR
Chemometrics
Wavelength selection
Classification

ABSTRACT

A rapid approach has been developed for the characterization of diesel quality, based on attenuated total reflectance – Fourier transform infrared (ATR-FTIR) spectrometry, which could be useful for diagnosing the sample quality condition. As a supervised technique, linear discriminant analysis (LDA) was employed to process the spectrometric data. The role of variable selection methods was also evaluated. Successive projection algorithm (SPA) and genetic algorithm (GA) feature selection techniques were applied prior to the discriminative procedure. It was aimed to compare the effect of feature selection procedures on classification capability of IR spectrometry for the diesel samples according to their quality passed or quality failed situation. Predictive capability of LDA was compared with that obtained by GA-LDA and SPA-LDA. Results showed 91.1%, 93.3% and 95.6% of accuracy for LDA, GA-LDA and SPA-LDA respectively. Thus SPA-LDA together with ATR-FTIR spectrometry was proposed as a fast screening analytical test for the evaluation of quality passed/failed situation in diesel samples.

1. Introduction

Diesel (petrodiesel) is a specific fraction of petroleum distillation process, which is consumed as fuel in diesel engines. It is a formulated mixture of different chemicals, such as linear chain hydrocarbons, naphthenic and aromatic compounds. Quality of diesel fuels is affected by several parameters while the most important factor is the relative proportion of each type of hydrocarbon in the formulation. Consistency in the quality of diesel is of high importance for commercial aims, customer rights, business competition and environmental risks. This concern would confirm the need of a careful quality control in the refining and fuel manufacturing industrial systems. Accordingly, each refinery industrial plant maintains a strategic department of quality control (QC), equipped with appropriate instruments and experimental expert human resources which would monitor the quality of fuel products in different stages of refining-manufacturing process. All the QC tests must be performed based on standard test procedures, which have been validated in the petroleum industry previously. These methods are utilized to evaluate the physico-chemical characteristics of fuel samples. Environmental, economic and industrial reasons encouraged the specialists for the development of rapid, precise and reliable experimental monitoring approaches to evaluate the fuel quality inside the refineries. There are several reports, dealing with analytical proposals for quality screening of diesel and biodiesel products [1]. Chromatographic methods, i.e. thin layer (TLC) [2], high performance liquid (HPLC) [3], gas (GC) [4] and gel permeation (GPC) [5] are the most common analysis approaches. However, most of these well-known methods are destructive, tedious and labor, consuming huge amount of chemical reagents to be performed. On the other hand, IR spectroscopy based methodologies are getting consolidated as rapid, reliable and non-destructive ones, which are free of sample preparation steps [6], thus offering green alternatives for the determination of sample parameters and characteristics without any previous sample pretreatment [7].

Research efforts concerning the application of IR spectrometry in diesel analysis involve the quantitative determination of some parameters in diesel samples [8,9] and pattern recognition strategies to assess them. In case of pattern recognition of diesel samples by IR spectrometry, reports are mostly related to recognizing the adulteration in diesel or its blend with biodiesel [10,11]. Most of these reports are based on chemometric processing of the spectrometric data and there are many efforts to improve the mathematical treatment of data in order to obtain as accurate as possible information while the analyst is enabled to

* Correspondence to: Chemistry Department, Faculty of Science, IKIU, P.O. Box: 34149-1-6818, Qazvin, Iran. Tel./Fax: +98 281 3780040.
E-mail address: mrkhanmohammadi@gmail.com (M. Khanmohammadi).

0039-9140/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.talanta.2012.11.032

increase the quality of results. An approach for this aim is to improve the previous methods by variable selection techniques and thus, in this work, we have evaluated the improvement in the capability of a supervised pattern recognition technique (LDA) in discrimination between quality passed and quality failed diesel samples from their ATR-FTIR spectra by incorporating two different variable selection methods, successive projection algorithm (SPA) and genetic algorithm (GA) in order to make a clear comparison between the outputs.

2. Experimental

2.1. Samples, apparatus and software

Commercial diesel samples (a total number of 90 samples) were obtained from the quality control laboratory of Shahid Tondguyan Oil Refining Company (Tehran, Iran) collected during 3 months of manufacturing. All the samples were kept refrigerated prior to IR spectrum recording, avoiding the probable loss of volatile compounds. Using a Tensor-27 Bruker FT-IR spectrometer, mid-IR spectra were obtained in triplicate for each sample and the average of spectra was used to be processed. In all cases, spectra were recorded at room temperature (22 ± 2 °C) using a horizontal, fixed path ATR cell (ZnSe, 45°, and single reflection by Pike®), a Ge–KBr beam splitter, a DTGS detector and Beer–Norton apodization. The spectral resolution was fixed at 8 cm^{-1} and 64 scans were accumulated over the range from 600 to 4000 cm^{-1} for each spectrum. Chemometric data processing was performed by MATLAB Ver. 8.0

2.2. Reference quality assessment of diesel by standard experimental procedures

The standard quality situation of diesel samples was established from the results obtained by ASTM reference test methods. The main monitored parameters were specific gravity at 60 F [12], cetan number [13], initial and final boiling temperatures [14], 10–90% distillation temperatures (within 10% intervals) [15], kinematic viscosity [16], cloud point [17], pour point [18], water and sediment content [19,20] and color [21]. All the aforementioned tests were made on each sample by triplicate and the average was considered as the analysis output. Considering the quality limits of the refinery, after performing the standard test experiments, the QC decision was made according to the defined values. Obtained samples were separated into two different groups: QC passed (60 samples) and QC failed (30 samples).

3. Chemometric methods employed

3.1. Linear discriminant analysis (LDA)

LDA is a powerful classification method, widely employed in analytical studies [22]. As a pattern recognition technique, LDA maximizes the ratio of between-class variance to the within-class variance in any particular data set in order to obtain a maximum discrimination. It does not make any change in the location of objects but only tries to obtain clear class discrimination, making a decision between the given classes. This would also help for understanding the distribution of the featured data. LDA is capable of handling the classification cases where the within-class frequencies are unequal and the performances have been examined on randomly generated test data. Investigating an unknown sample by LDA, a critical parameter is prior probability of class k (π_k), which is usually estimated simply by empirical frequencies of the training set, being evaluated through the use of Eq. (1)

$$\sum_{k=1}^{K} \pi_k = 1k \tag{1}$$

being π defined in (2)

$$\hat{\pi} = \frac{\text{number of samples in class } k}{\text{total number of samples}} \tag{2}$$

The class-conditional density of a sample of X in the class is defined in (3)

$$G = k \quad \text{is} \quad f_k(x) \tag{3}$$

and posterior probability of this classification can be evaluated as (4)

$$\Pr\left(G = k \mid X = x\right) = \frac{f_k(x)\,\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l} \tag{4}$$

3.2. Successive projection algorithm

SPA is a powerful algorithm for variable selection by minimizing the redundant information content of obtained signals and resolving the problems caused by collinearity [23–25]. The prediction capability of SPA coupled to a classification model is very critical.

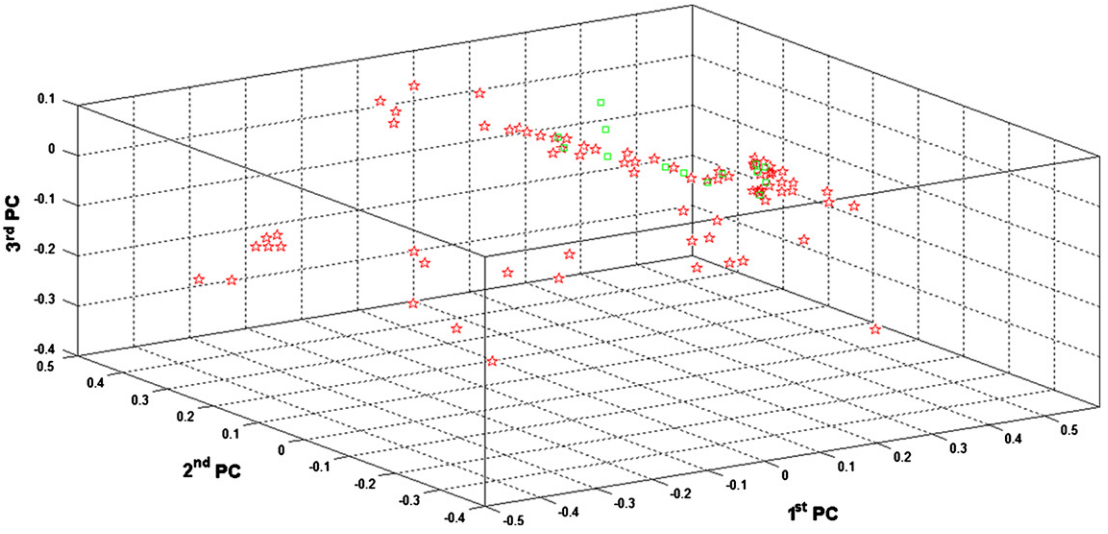

Fig. 1. PCA score pot obtained for quality based analysis of ATR-FTIR spectra of diesel samples.

This is due to the effect of the variable selection strategy. In order to generate a robust model, truly "causal" variables must be used; otherwise they may produce good correlations through mere chance from fortuitous noise trends [26]. As a forward selection method to process spectrometric data, SPA starts with one wavenumber and then incorporates a new one during iterations, until a specified number (N) of wavenumbers is reached.

During orthogonalization, SPA may resemble the Gram–Schmidt algorithm to manipulate the studies data in order to generate a new set of orthogonal vectors, without any physical meaning. SPA would not modify the original data vectors, since projections are used only for selection purposes. So, the relation between spectral variables and data vectors is preserved. SPA procedure includes from the first wavenumber $k(0)$ to a final number of data used N and each $\boldsymbol{x}_j$ is considered as the jth column of signal data matrix $\mathbf{X}_{cal}$; $j=1,\ldots,J$ prior to the first iteration, while $S=\{j \text{ such that } 1\leq j\leq J \text{ and } j\notin\{k(0),\ldots,k(n-1)\}\}$ is the set of wavenumbers which have not been selected yet. The projection of $\boldsymbol{x}_j$ on the orthogonal subspace defined by $\boldsymbol{x}_{k(n-1)}$ is computed for all $j\in S$ as indicated in eq. (5)

$$\mathbf{P}\boldsymbol{x}_j = x_j - \left(x_j^T x_{k(n-1)}\right)x_{k(n-1)}\left(x_{k(n-1)}^T x_{k(n-1)}\right)^{-1} \tag{5}$$

being $\mathbf{P}$ the projection operator, $k(n)=\arg(_{\max}\ ||\mathbf{P}\boldsymbol{x}_j||,\ j\in S)$ and $\boldsymbol{x}_j=\mathbf{P}\boldsymbol{x}_j,\ j\in S,\ n=n+1$ and if $n<\mathrm{N}$ the procedure is repeated once again. If the resulting wavenumbers are $\{k(n);\ n=0,\ldots,N-1\}$, then the computation is stopped and the number of projection operations performed in the selection process is $(N-1)(J-N/2)$.

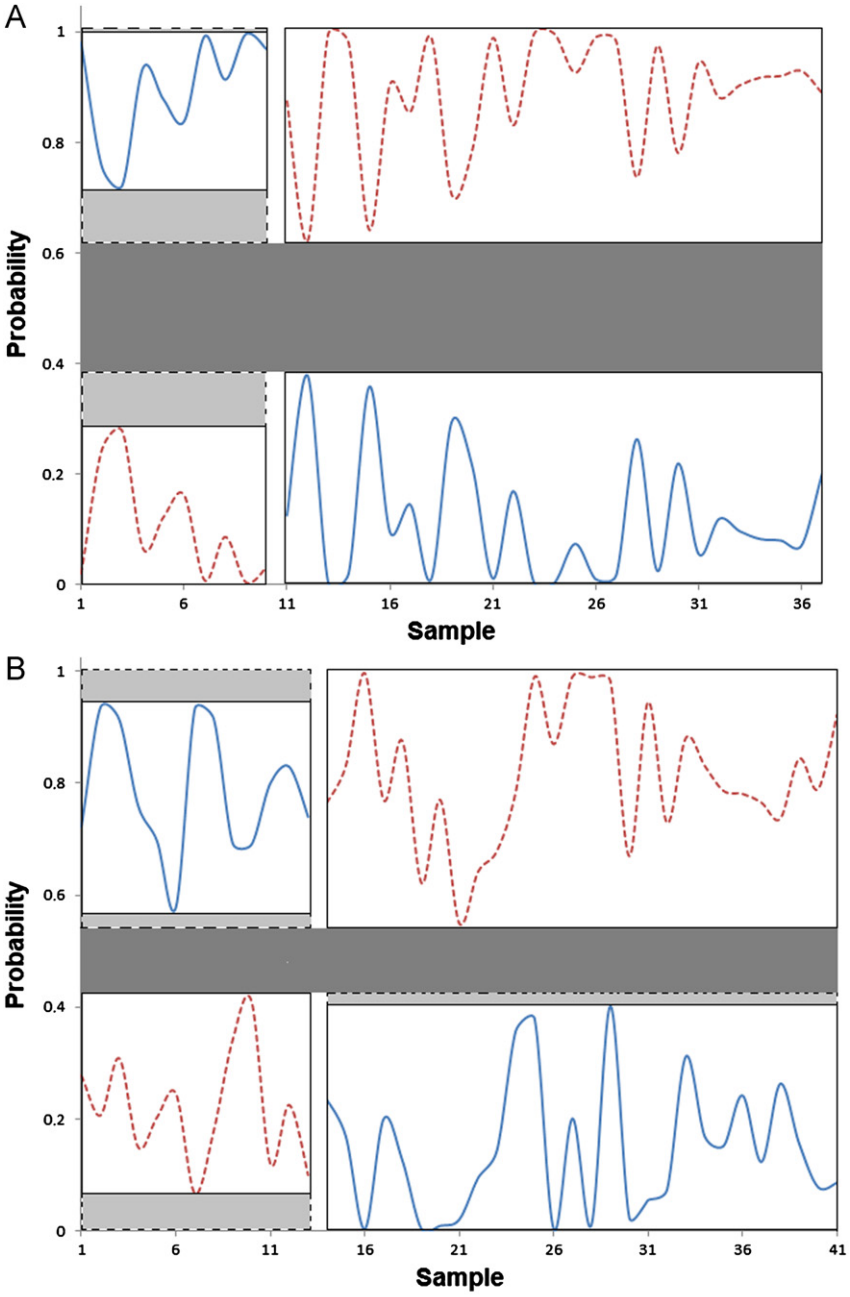

Fig. 2. Graphical explanation of the discrimination success of LDA for both quality passed (dash line) and quality failed (solid line) samples in training (A) and validation (B) sets.

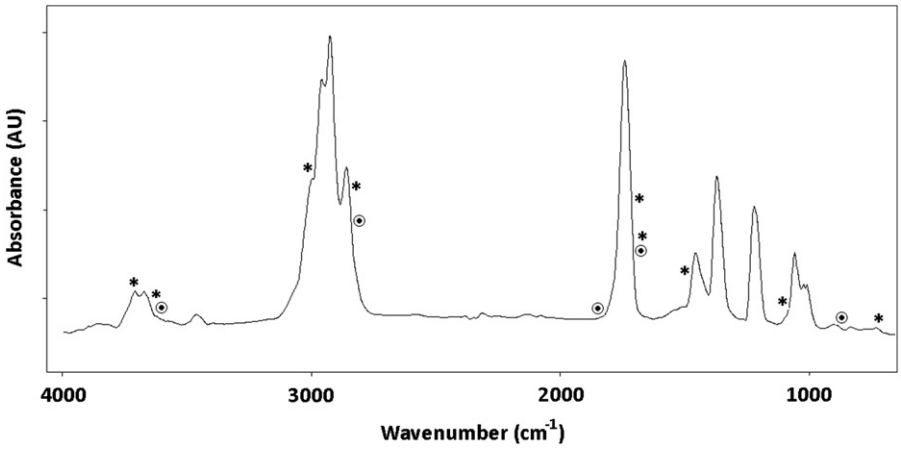

Fig. 3. Spectral features selected by SPA (*) and GA (*).

3.3. Genetic algorithm

As a well-known feature selection method, GA is aimed to optimize a given response function. The main explanation of this method is inspired by the evolution theory which claims that in specific environment best individuals have more chance to spread their genomes by reproduction. In case of spectral feature selection, GA is applied to select the most representatives of the huge spectral data set. In case of infrared spectrometry these selected wavenumbers would reflect the information content of the whole spectral region. By mixing the selected informative variables (spectral features) the best one is resulted. Selected variables correspond to genomes and different values of the different variables in a spectral data set correspond to a chromosome. The main steps of a GA procedure are:

i). Spectral variable coding e.g. direct conversion of the spectral value to a mathematical code such as binary coding system.
ii). Initiation of general population, composed of some chromosomes (usually 20–100) while the structure of each chromosome is determined in a totally random way.
iii). Response evaluation for each chromosome, associated with the corresponding spectral data to be evaluated. (If the output is out of the range, a null response is selected).
iv). Reproduction by creating a new population of N chromosomes, considered as the next generation. At this stage a new population is obtained in which best chromosomes are present and a better average response could be achieved.
v). Mutation for random combination at minimum data level and obtaining the probable useful data via forming the new population by random pairs of previous ones.

The last three steps must be repeated until reaching the aimed criterion which usually is the loss of noticeable improvement in the response.

4. Results and discussions

4.1. Data preprocessing

Taking into account the role of experimental conditions and sampling situation, the data preprocessing strategy was defined prior to performing any classification approach. Row mean centering (MC) for removal of the constant error was the starting point of data preprocessing. MC was accomplished by subtracting the data set mean from each data entity. As the data mining method by which the diesel samples are going to be qualitatively classified is affected by the magnitude of the spectral expressions, MC is a helpful practice to consider the relative changes, instead of the absolute magnitudes. It would remove the dependence on magnitude and mean centered data possess a mean expression of zero. The second step was to conduct standard normal variate transformation on individual spectra to reduce the baseline shift and collinearity. By making this transformation the spectra were centered and scaled by their own standard deviation.

4.2. Principal component analysis (PCA) and outlier detection

As a size reduction technique for dealing with big data sets, PCA was performed to obtain most of the original variability in a smaller space. The output of PCA consists of a loading matrix which would represent the principal components (PCs) that are spectrum like patterns. The other part of PCA output is the score matrix which contains the coordinates of the original spectra on the new axes determined by the corresponding PC's. Considering leverage criterion to flag atypical samples, PCA based outlier detection was utilized to avoid the disadvantage of taking into account the probable outliers in the data set [27]. The leverage of a sample is a measure of its spatial distance to the main colony of samples in the PCA provided data space. In this work, the scores of the three first PCs were utilized to find out the outliers. Leverage threshold selected in this work was 0.09 and considering this threshold, the number of spectra in the data set was reduced from 90 to 85. PCA demonstrated that 82.7%, 10.8% and 3.2% of data variance are covered by 1st, 2nd and 3rd PC respectively. However, graphical output of PCA in the score plot of the abovementioned PCs does not show any evidence on differentiation of quality passed/failed samples (see Fig. 1).

4.3. Linear discriminant analysis

In this study LDA was applied to the spectral data set to investigate two different objectives simultaneously. In order to achieve a predictive method, with the goal of formulating a discrimination rule used to predict or allocate the quality situation of unknown diesel samples in "quality passed" and "quality failed" predefined classes and also to evaluate it as an exploratory tool to increase the understanding about the differences between classes. Mahalanobis distance (also called "statistical" distance) was the measure to explore a decision border between considered classes. The boundary plane (hyper plane) was calculated in order that the variance between the classes could be maximized while the variance within the individual classes was minimized. LDA

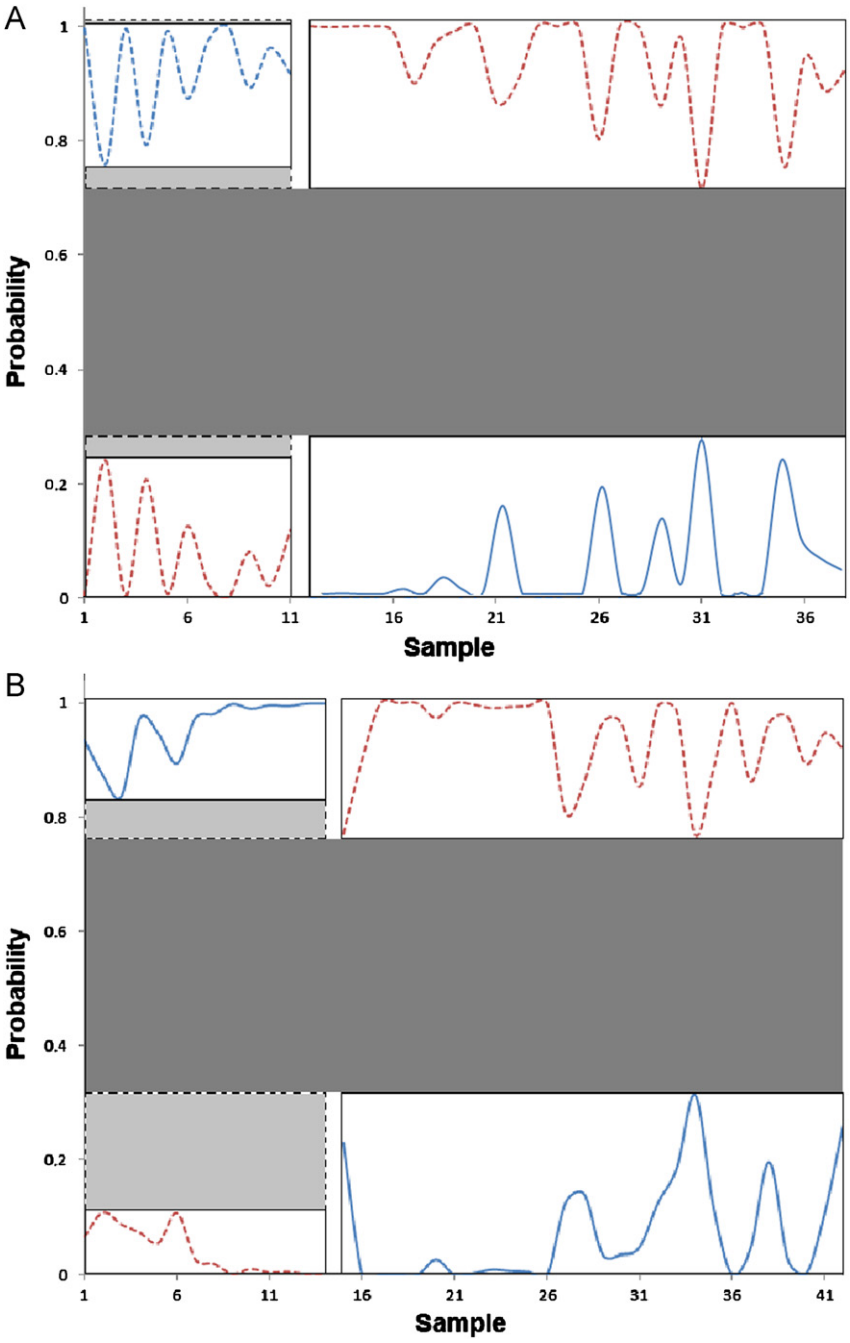

Fig. 4. Graphical explanation of the discrimination success of SPA-LDA for both quality passed (dash line) and quality failed (solid line) samples in training (A) and validation (B) sets.

training model was built using 40 samples (28 passed and 12 failed) and remaining 45 samples (30 passed and 15 failed) were considered as unknown test set to be used for model evaluation. Kennard–Stone algorithm was applied for data partitioning. This algorithm is based on finding two points with the most separated position in the data set and finding the smallest distance to any object already selected. It also maximizes the minimal distances between already selected objects and the remaining objects. During the validation of the training model one QC passed and two QC failed samples were misclassified while in case of the test set two samples of each class were classified wrongly. Evaluating the capability of LDA in discrimination of analyzed samples, the accuracy of the built model was 92.5% and 91.1% for the training and prediction steps respectively. Leaving the misclassified samples out and making a two dimensional plot presentation of the probability for the remaining samples of test set, a nonlinear behavior (see Fig. 2) was observed in the probability to find samples in both classes while the variation in probability of quality passed class was wider. The free space between two classes is considered as a data space in which no probability is observed (dark gray colored part of the graph) and the probability curves do not occupy it. This part of the graph could be desired as the "safety region" and used as a measure of discriminative capability of the LDA model. It is obvious that the safety region is decreased while comparing the

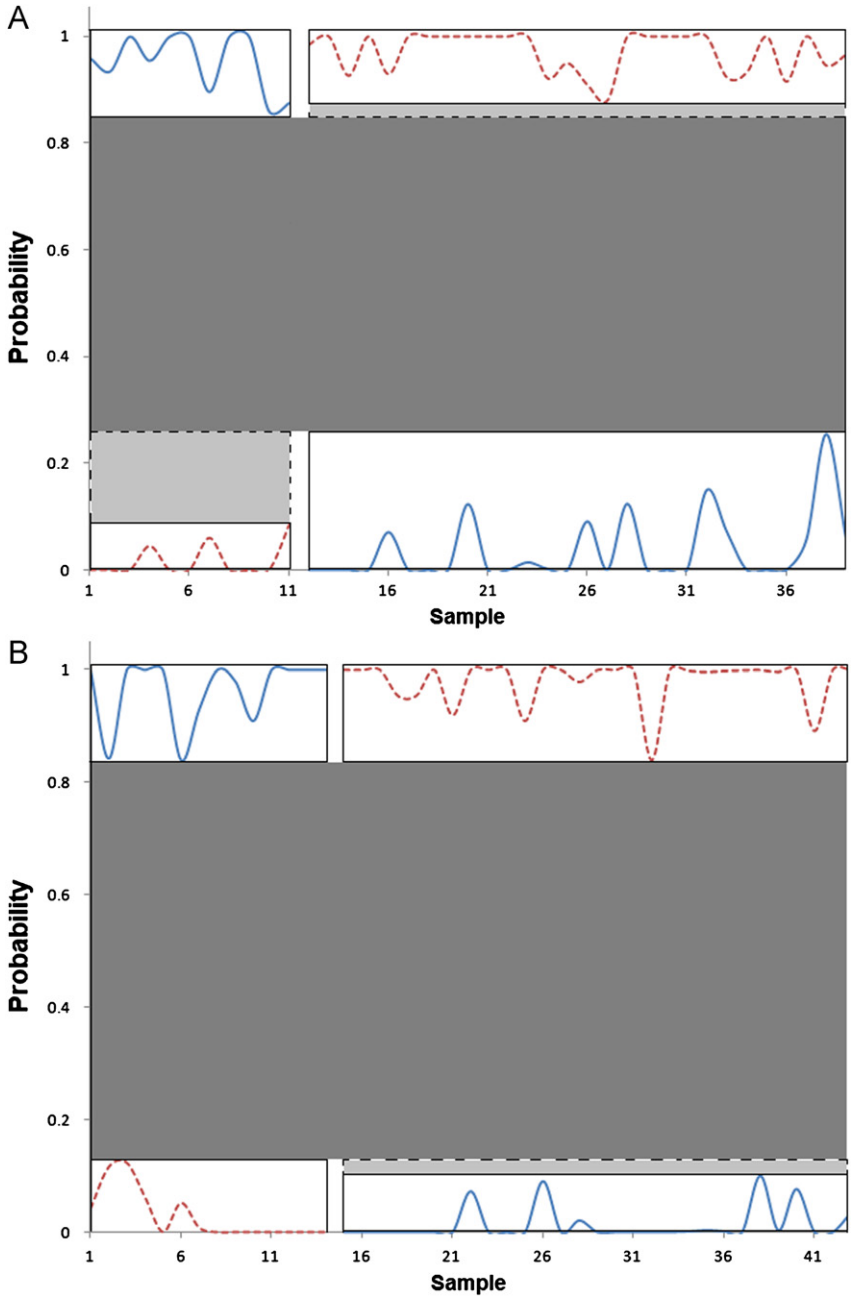

Fig. 5. Graphical explanation of the discrimination success of GA-LDA for both quality passed (dash line) and quality failed (solid line) samples in training (A) and validation (B) sets.

test set with the training set. So, it could be a caution for the generalization of the model, analyzing larger set of unknown samples. Lack of discrimination power (related to size of safety region) could be an effective parameter, reducing the accuracy of model in future applications for industrial aims.

The visual comparison of the boxes in which the probability curves are surrounded, evidences that in case of samples from quality failed class, the probability of a correct classification is never 1 (100% allocation) either in training or in validation test subset (see Fig. 2). On the other hand, the probability variation of this class is less than that of quality passed class. This could be due to the number of samples in quality passed class which is higher than that in the quality failed class and thus they would fall within a wider range of probability. Comparing the graph area in which one of the classes does not demonstrate any probability, called "dead region" (light gray colored area), it could be realized that in case of quality failed class the dead region is larger than in the quality passed class. Heterogeneity of the QC results in the failed class could also influence this wider distribution. Above mentioned tolerated observations in the classification of diesel samples by LDA encouraged us for improving the chemometric strategy. In this way, SPA and GA feature selection techniques were used in combination with LDA.

4.4. Successive projection algorithm

The same as other variable selection procedures in spectroscopy, SPA consists of selecting a particular region of the spectrum supposed to contain useful spectral variables. Thus it is better to start from a few variables. In this work, SPA was started with one variable and by an iterative process new variables were incorporated, until the specified number of variables was reached (five wavenumbers as shown in Fig. 3). Internal validation of the training set was utilized for selection of the best variables in the applied algorithm. On the other

hand, total or partial section of training set is also applicable as dataset for choice of variables. Gram–Schmidt algorithm was applied to construct the orthonormal basis for an inner product space. Applying SPA prior to LDA, there was one misclassified sample in each class for the training model while there were two and one misclassified samples in quality passed and quality failed sets respectively for the unknown test set. Accuracy of SPA-LDA was 95.0% and 93.3% for training and validation set respectively. Comparing the graphical presentation of probabilities in SPA-LDA with those of LDA, it was observed that safety region is extended by using SPA. On the other hand, dead region of quality failed class was increased from training to test set. An advantage of SPA-LDA toward LDA is that by using SPA there were some samples with probability of 1 which confirms the ensured belonging of sample to predicted class (see Fig. 4).

4.5. Genetic algorithm

GA is a variable selection method known to be useful for solving the optimization problems e.g. minimizing an objective function over a given space of arbitrary dimension. The diesel classification is an example of those problems which deal with associating a given input spectral pattern with one of the considered classes (QC passed or failed). A decision rule determines a decision border which partitions the feature space into regions associated with each class. It represents the best solution to the classification problem. Spectral patterns were specified by a number of nine features (wavenumbers) representing the spectral measurements made on the diesel samples to be classified (see Fig. 3). Patterns were some points in a multi dimensional space consisting of vectors made by selected variables and QC passed/failed classes were the sub-spaces. Constructing the genetic model according to the criteria as population size of 100, generation of 30 and mutation rate of 0.01 by LDA fitness function, nine wavenumbers were selected for LDA.

Performing LDA on GA selected variable data set only one sample of quality failed class was wrongly allocated in the quality passed class in the training subset while there was one misclassified sample in each considered class for the validation test set. Thus the accuracy of model was 97.5% and 95.6% for training and validation set respectively, evidencing the improvement of the previous treatment.

Investigating the graphical modeling of probabilities, the safety region was extended in comparison with LDA and SPA-LDA treatments (see Fig. 5, and compare with Figs. 2 and 4). On the other hand, all the probability sections were narrow and free of dead region, except the probability of quality failed samples being considered as QC passed ones. In this regard, it can be concluded that the general situation of the chemometric LDA output can be improved by using SPA and GA while the robustness of GA is higher than that obtained by SPA. Several reasons could be supposed for this improvement e.g. increment in the number of variables selected by GA as compared to SPA while the 1st derivative of spectral regions in which the GA selected variables do exist are higher than those employed by SPA.

5. Conclusions

The use of a feature selection procedure together with LDA classification of diesel ATR-FTIR spectra, clearly improves the prediction capability and robustness of LDA pattern recognition method for assessment of passed/failed quality of samples and the most important conclusion is that it could be considered as a spectral application thus opening exciting possibilities in both industrial and diagnostic fields. This could be introduced for industrial analysis of diesel samples, directly, with no sample preparation and in a fast way, leading to a clear decision about its quality failed/passed condition

References

[1] G. Knothe, J. Am. Oil Chem. Soc. 10 (2006) 823–833.
[2] R. Bacaud, V.L. Cebolla, L. Membrado, M. Matt, S. Pessayre, E.M. Gálvez, Ind. Eng. Chem. Res. 41 (2002) 6005–6014.
[3] L.F.P. Brandão, J.W.B. Braga, P.A.Z. Suarez, J. Chromat. A 1225 (2012) 150–157.
[4] K.M. Pierce, S.P. Schale, Talanta 83 (2011) 1254–1259.
[5] S. Florio, L. Pellegrini, S. Florio, L. Pellegrini, SAE Technical Paper, (2011), http://dx.doi.org/10.4271/2011-01-2098.
[6] A. Borin, R.J. Poppi, Vib. Spec. 37 (2005) 27–32.
[7] J. Moros, S. Garrigues, M. de la Guardia, Trends Anal. Chem. 29 (2010) 578–591.
[8] R.M. Balabin, E.I. Lomakin, R.Z. Safieva, Fuel 90 (2011) 2007–2015.
[9] M. Coronado, W. Yuan, D. Wang, F.E. Dowell, Appl. Eng. Agr. 25 (2009) 217–221.
[10] R.M. Balabin, R.Z. Safiev, E.I. Lomakin, Anal. Chim. Acta 671 (2010) 27–35.
[11] M.J.C. Pontes, C.F. Pereira, M.F. Pimentel, F.V.C. Vasconcelosa, A.G.B. Silva, Talanta 85 (2011) 2159–2165.
[12] ASTM D1298, Standard Test Method for Density, Relative Density (Specific Gravity), or API Gravity of Crude Petroleum and Liquid Petroleum Products by Hydrometer Method. ASTM International, Conshohocken, PA, USA.
[13] ASTM D976, Standard Test Method for Calculated Cetane Index of Distillate Fuels. ASTM International, Conshohocken, PA, USA.
[14] ASTM D2887, Standard Test Method for Boiling Range Distribution of Petroleum Fractions by Gas Chromatography. ASTM International, Conshohocken, PA, USA.
[15] ASTM D6352, Standard Test Method for Boiling Range Distribution of Petroleum Distillates in Boiling Range from 174 to 700 °C by Gas Chromatography. ASTM International, Conshohocken, PA, USA.
[16] ASTM D4486, Standard Test Method for Kinematic Viscosity of Volatile and Reactive Liquids. ASTM International, Conshohocken, PA, USA, 2010.
[17] ASTM D2500, Standard Test Method for Cloud Point of Petroleum Products. ASTM International, Conshohocken, PA, USA.
[18] ASTM D97, Standard Test Method for Pour Point of Petroleum Products. ASTM International, Conshohocken, PA, USA.
[19] ASTM D95, Standard Test Method for Water in Petroleum Products and Bituminous Materials by Distillation. ASTM International, Conshohocken, PA, USA.
[20] ASTM D1796, Standard Test Method for Water and Sediment in Fuel Oils by the Centrifuge Method (Laboratory Procedure). ASTM International, Conshohocken, PA, USA.
[21] ASTM D6045, Standard Test Method for Color of Petroleum Products by the Automatic Tristimulus Method. ASTM International, Conshohocken, PA, USA.
[22] M. Otto, Chemometrics, Statistics and Computer Application in Analytical Chemistry, Wiley VCH, Weinheim, Germany, 1998.
[23] M.J.C. Pontes, R.K.H. Galvão, M.C.U. Araújo, P.N.T. Moreira, O.D.P. Neto, G.E. José, T.C.B. SaldanhaChemom, Intell. Lab. Syst. 78 (2005) 11–18.
[24] A.C. Silva, L.F.B.L. Pontes, M.F. Pimentel, M.J.C. Pontes, Talanta 93 (2012) 129–134.
[25] U.T.C.P. Souto, M.J.C. Pontes, E.C. Silva, R.K.H. Galvão, M.C.U. Araújo, F.A.C. Sanches, F.A.S. Cunha, M.S.R. Oliveira, Food Chem. 119 (2010) 368–371.
[26] M.C.U. Araujo, T.C.B. Saldanha, R.K.H. Galvao, T. Yoneyama, H.C. Chame, V. Visani, Chemom. Intell. Lab. Syst. 57 (2001) 65–73.
[27] W.F. de Carvalho Rocha, R. Nogueira, B.G. Vaz, J. Chemom. 26 (2012) 456–461.