Analytica Chimica Acta 671 (2010) 27–35

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Gasoline classification using near infrared (NIR) spectroscopy data: Comparison of multivariate techniques

Roman M. Balabin[a,*], Ravilya Z. Safieva[b], Ekaterina I. Lomakina[c]

[a] Department of Chemistry and Applied Biosciences, ETH Zurich, 8093 Zurich, Switzerland
[b] Gubkin Russian State University of Oil and Gas, 119991 Moscow, Russia
[c] Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, 119992 Moscow, Russia

ARTICLE INFO

Article history:
Received 8 April 2010
Received in revised form 6 May 2010
Accepted 7 May 2010

Keywords:
Discriminant analysis (LDA, QDA, RDA)
Petroleum (crude oil)
Biofuel (biodiesel, bioethanol, ethanol–gasoline fuel)
Soft independent modeling of class analogy
K-Nearest neighbor method
Support vector machine
Probabilistic neural network
Near infrared spectroscopy

ABSTRACT

Near infrared (NIR) spectroscopy is a non-destructive (vibrational spectroscopy based) measurement technique for many multicomponent chemical systems, including products of petroleum (crude oil) refining and petrochemicals, food products (tea, fruits, e.g., apples, milk, wine, spirits, meat, bread, cheese, etc.), pharmaceuticals (drugs, tablets, bioreactor monitoring, etc.), and combustion products. In this paper we have compared the abilities of nine different multivariate classification methods: linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), regularized discriminant analysis (RDA), soft independent modeling of class analogy (SIMCA), partial least squares (PLS) classification, *K*-nearest neighbor (KNN), support vector machines (SVM), probabilistic neural network (PNN), and multilayer perceptron (ANN-MLP) – for gasoline classification. Three sets of near infrared (NIR) spectra (450, 415, and 345 spectra) were used for classification of gasolines into 3, 6, and 3 classes, respectively, according to their source (refinery or process) and type. The 14,000–8000 cm^{-1} NIR spectral region was chosen. In all cases NIR spectroscopy was found to be effective for gasoline classification purposes, when compared with nuclear magnetic resonance (NMR) spectroscopy or gas chromatography (GC). KNN, SVM, and PNN techniques for classification were found to be among the most effective ones. Artificial neural network (ANN-MLP) approach based on principal component analysis (PCA), which was believed to be efficient, has shown much worse results. We hope that the results obtained in this study will help both further chemometric (multivariate data analysis) investigations and investigations in the sphere of applied vibrational (infrared/IR, near-IR, and Raman) spectroscopy of sophisticated multicomponent systems.

1. Introduction

Near infrared (NIR) spectroscopy is a non-destructive measurement technique for many chemical compounds, including products of petroleum refining and petrochemicals, food products (tea, fruits, milk, meat, etc.), pharmaceuticals (drugs, tablets, bioreactor monitoring, etc.), combustion products, and many other [1–4]. It has proved its efficiency for laboratory and industrial applications [1,2]. Many current studies are focused on NIR sensor design for on-line [3] or portable operation [4].

One of the possible usages of near infrared spectroscopy is in a petroleum industry [5–12], because of the fact that many products of petroleum refining and petrochemicals consist of hydrocarbons, whose content can be estimated by NIR method. Near infrared spectroscopy (Tables 1 and 2) is useful to overcome many limitations, especially in a complicated real process, where on-line measuring is important to monitor the quality of products (e.g., gasoline production) [13].

Multivariate statistics techniques have boosted the use of NIR instruments [13,14]. Only methods of chemometrics are able to process enormous amounts of sophisticated experimental data that are provided by NIR technique. Many calibration methods were used for gasoline quality prediction [5–12,15]. Linear techniques, "quasi-nonlinear" techniques, and "truly" nonlinear techniques – all of them were used for gasoline properties and quality coefficients prediction [5–12].

Despite the fact that gasoline (or other products of petroleum refining) classification is also a practically important task, not many papers are devoted to this problem [11,16–18]. In this paper we have tried to evaluate the efficiency of different classification methods for gasoline classification by source and type.

Gasoline identification by source (refinery) is an important factor for both quality control and identification of gasoline adulteration. This method is based on the difference in crude oil that results in chemical difference of gasolines from variant refineries. Gasoline type is needed for classification of gasolines by quality

* Corresponding author. Tel.: +41 44 632 4783.
E-mail address: balabin@org.chem.ethz.ch (R.M. Balabin).

0003-2670/$ – see front matter © 2010 Published by Elsevier B.V.
doi:10.1016/j.aca.2010.05.013

Table 1
Gasoline sample sets.

	Set A	Set B	Set C
Classification by	Source (refinery[a])	Source (process[b])	Type
Number of samples	150	117	115
Number of classes	3	6	3
Classes	Refinery 1; Refinery 2; Refinery 3	Straight-run; reformate; catalysate; isomerizate; hydrocracking gasoline; mixture[c]	Normal[d]; Regular; Premium
Distribution of samples[e]	50; 50; 50	30; 12; 15; 13; 12; 35	55; 45; 15

[a] Refinery 1, Orsknefteorgsintez (Russia); Refinery 2, Kirishinefteorgsintez (Russia); Refinery 3, Astrakhan'gazprom (Russia).
[b] For details one can consult Refs. [46,47].
[c] All gasoline fractions (straight-run, reformate, catalysate, isomerizate, hydrocracking) are presented in mixtures.
[d] Normal (Russia) – regular gasoline with octane number of 80.
[e] Distribution of samples is given according to the classes above.

(and price). The process of gasoline production defines the chemical composition of gasoline fraction. This difference is an important factor for gasoline mixing.

In this paper three gasoline sample sets were used to compare prediction capabilities of linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), regularized discriminant analysis (RDA), soft independent modeling of class analogy (SIMCA), partial least squares (PLS) classification, *K*-nearest neighbor (KNN), support vector machines (SVM), probabilistic neural network (PNN), and multilayer perceptron (MLP). All classifications are based on gasoline near infrared (NIR) spectroscopy data.

2. Brief descriptions of the classifiers

In this section we have tried to give some basic remarks about the methods used. The references on detailed papers were also provided.

2.1. Linear discriminant analysis (LDA)

Among many possible techniques for data classification, linear discriminant analysis (LDA) is a commonly used one. Linear discriminant analysis (LDA) is used to find the linear combination of features which best separate two or more classes of object or event. The resulting combinations may be used as a linear classifier, or more commonly in dimensionality reduction before later classification. This method maximizes the ratio of between-class variance to the within-class variance in any particular data set thereby guaranteeing maximal separability. Details about LDA can be found in Refs. [19–21].

2.2. Quadratic discriminant analysis (QDA)

Quadratic discriminant analysis (QDA) is closely related to LDA (see above). Unlike LDA however, in QDA there is no assumption that the covariance of each of the classes is identical. When the assumption is true, the best possible test for the hypothesis that a given measurement is from a given class is the likelihood ratio test.

Using QDA each of the covariance matrices is estimated separately, which requires a larger sample than that used in LDA to reach the same level of reliability in the estimations and hence in the predictions. QDA details can be found in Refs. [19–21].

Table 2
NIR experimental parameters.

	Set A	Set B	Set C
Spectral range (cm^{-1})	14,000–8000	14,000–8000	13,500–8500
Sample volume (cm^3)	100	100	100
Optical path (cm)	10	10	10
Spectra resolution (cm^{-1})	16	8	8
Number of scans	64	64	72
Cell material	Quartz	Quartz	Quartz
Time of one measurement (min)	2–3	3–4	2–3
Number of measurements (per sample)	3	3–5	3

No spectra preprocessing was used.

2.3. Regularized discriminant analysis (RDA)

Regularized discriminant analysis (RDA) can be called a "compromise" between LDA and QDA.

RDA introduces a double type of bias in the estimation of the covariance matrices of each class with the aim of stabilizing the prediction in the case of a deficient estimation of their elements. This is a very frequent problem encountered with real data because the estimations of the matrices are very sensitive to the presence of different data. For a detailed analysis one can consult Refs. [21,22].

2.4. Soft independent modeling of class analogy (SIMCA)

Soft independent modeling of class analogy (SIMCA) is a method for supervised data classification that requires a training data set consisting of samples (or objects) with a set of attributes and their class membership. The term "soft" refers to the fact the classifier can identify samples as belonging to multiple classes and not necessarily producing a classification of samples into non-overlapping classes.

In order to build the classification models, the samples belonging to each class need to be analyzed using principal components analysis (PCA). For a given class, the resulting model then describes a line, plane or hyperplane. For each modeled class, the mean orthogonal distance of training data samples from the line, plane or hyperplane (calculated as the residual standard deviation) is used to determine a critical distance for classification.

New observations are projected into each PC model and the residual distances calculated. An observation is assigned to the model class when its residual distance from the model is below the statistical limit for the class. The observation may be found to belong to multiple classes and a measure of goodness of the model can be found from the number of cases where the observations are classified into multiple classes. Details about SIMCA model building and usage can be found in Ref. [23].

2.5. Partial least squares (PLS) regression and classification

Partial least squares (PLS) regression is an extension of the multiple linear regression model.

PLS method has found widespread use for multivariate analysis of spectral data. The main goal of PLS usage is calibration model building, but this technique can also be applied for classification purposes. Some information about the idea of PLS classification can be found in Section 3.6.3 and Refs. [24–27].