The Representation of Chemical Spectral Data for Classification

Diana Porro[1,2], Robert W. Duin[2], Isneri Talavera[1], and Noslen Hdez[1]

[1] Advanced Technologies Application Centre, Cuba
[2] Pattern Recognition Group, TU Delft, The Netherlands
{dporro,italavera,nhernandez}@cenatav.co.cu, r.duin@ieee.org

Abstract. The classification of unknown samples is among the most common problems found in chemometrics. For this purpose, a proper representation of the data is very important. Nowadays, chemical spectral data are analyzed as vectors of discretized data where the variables have not connection, and other aspects of their functional nature e.g. shape differences (structural), are also ignored. In this paper, we study some advanced representations for chemical spectral datasets, and for that we make a comparison of the classification results of 4 datasets by using their traditional representation and two other: Functional Data Analysis and Dissimilarity Representation. These approaches allow taking into account the information that is missing in the traditional representation, thus better classification results can be achieved. Some suggestions are made about the more suitable dissimilarity measures to use for chemical spectral data.

Keywords: Pattern Recognition, Chemometrics, Classification, Spectral Data, Dissimilarity Representation, Functional Data Analysis.

1 Introduction

One of the main problems that can be found in any research area is related to the classification of unknown objects. A good representation of the data is one of the most important aspects to be considered in this process. The more information about the real data is described in its representation, the higher the probability of a good classification of the samples.

Although chemical spectral data are typically curves plotted as functions of wavelengths, product concentration, etc., they are traditionally represented as a sequence of individual observations (features) made on the objects, ignoring important aspects of their functional nature i.e. connectivity, shape changes, etc.

Functional Data Analysis (FDA) [1] and Dissimilarity Representation (DR) [2] are rather new approaches that, in their own way, can take the functional information into the data representation. FDA is an extension of the traditional multivariate analysis for data with a functional nature, and is based on considering the observed spectra as a continuous real-valued function instead of an array of individual observations. Several classical multivariate statistical methods been extended to work on it e.g. linear discriminant analysis (LDA) [3]. In the case of linear modeling, studies have also

E. Bayro-Corrochano and J.-O. Eklundh (Eds.): CIARP 2009, LNCS 5856, pp. 513–520, 2009.
© Springer-Verlag Berlin Heidelberg 2009

been made in regression [4]. A number of estimation methods for functional non-parametric classification and regression models have been introduced. Namely, k-Nearest Neighbor classifier (k-NN) [5], kernel classifiers e.g. Support Vector Machine (SVM) based on the Radial Basis Function (RBF) methods [6], [7], showing its application for chemical spectral data.

Although profound studies of the DR on chemical spectral data sets have not been done, there are already some results on spectral data in general [8], demonstrating its advantages for its classification. In this approach, based on the important role that proximities play in the classification process, the authors propose to work on a space defined by the dissimilarities between the objects [2]. This way, the geometry and the structure of a class are defined by the dissimilarity measure, by which we can take into account the information that can help to discriminate between objects of the different classes. So, the selection of a suitable measure for the particular problem is important. The DR has shown to be advantageous in problems where the number of objects is small, and also when they are represented in high dimensionality spaces, which are both common characteristics of chemical spectral data sets.

On the chemometrics side, some work has been done in the comparison of chemical spectral data. In [9], the authors are looking for similarity measures for infrared (IR) spectrometry. A more recent research [10] is about the comparison of drugs UltraViolet (UV) spectra by clustering, where they also try different dissimilarity measures.

The goal of this paper is to show, how the classification results can improve by using representations of the data that give more information about the real spectra than the feature representation. With this purpose, we make a comparison of the performance of 1-NN, Regularized LDA (RLDA), Soft Independent Modeling of Class Analogy (SIMCA) [11] and SVM classifiers on the three mentioned representations: feature, FDA and DR of four chemical spectral datasets. We also make a study of some dissimilarity measures that have already been used on these types of data, in order to propose which could be more suitable to take into account the main differences that can exist in spectral data sets: structure (shape) and/or concentration or intensity.

2 Functional Data Analysis

Functional Data Analysis (FDA) [1] was proposed as a way to retrieve the intrinsic characteristics of the underlying function from the discrete functional data. In this approach, the observations can be seen as continuous single entities, instead of sets of different variables. However, if the algorithms work on the functional spaces, their infinite dimensions can lead to theoretical and practical difficulties. To deal with the infinite dimensional problem, a filtering approach was constructed to reach a representation of a finite dimensionality.

For this approach, we have to select a proper family of basis functions to match the underlying function (s) to be estimated. In the case of spectral data, the basis of B-splines seems to be the most appropriate. A number of knots (points) between the start and end wavelengths has to be chosen, and a B-spline is run from one knot to another; the different splines overlap. The spectral function $x_i = x_i(\lambda)$ for sample i and

wavelengths λ , can be described by the linear combination of the basis functions $x_i = \sum_{k=1}^{K} c_{ik}\phi_k$, where $\{\phi_k\}_{k=1}^{K}$ is the basis of B-splines with K the number of basis functions, and c_{ik} the B-spline weights (coefficients). These are computed by minimizing the vertical distance between the observed spectral information and the fitted curve:

$$\min_{c_{ik}} \sum_{j=1}^{m} (x_{ij} - \sum_{k=1}^{K} c_{ik}\phi_k(\lambda_j))^2 ,$$

where x_{ij} is an element of the matrix conformed by a set of i spectra of j wavelengths. The function will be explained by the coefficients and the methods will take these as the new representation of the data instead of the original data points.

3 Dissimilarity Representation

The Dissimilarity Representation (DR) [2] proposes to work on the space of the proximities between the objects, instead of the space defined by their characteristics (features), as it is usually done.

In the new representation, instead of having a matrix $\mathrm{X}(m \times n)$, where m goes for the objects (spectrum) and n for the measured variables e.g. wavelengths, the set of objects will be represented by the matrix $\mathrm{D}(m \times q)$. This matrix contains the dissimilarity values between each object $x \in \mathrm{X}$ and the objects of the representation set $\mathrm{R}(p_1, p_2, ..., p_q)$, $d(x_m, p_q)$. The elements of R are called prototypes, and have preferably to be selected by some prototype selection method [12]. These prototypes are usually the most representative objects of each class $(\mathrm{R} \subseteq \mathrm{X})$, but the whole set of objects X can be used too, obtaining the square dissimilarity matrix, $\mathrm{D}(m \times m)$; R can also be a completely different set of objects.

For the DR three main approaches exist. In the first, the given dissimilarities are addressed directly e.g. k-NN. Another one is based on an approximate embedding of the dissimilarities into a pseudo-Euclidean space. The third and last one is defined as the dissimilarity space $\mathscr{D} \subseteq \mathbb{R}^n$, which is the one to be used here. This space is generated by the column vectors of the dissimilarity matrix, where each dimension corresponds to the dissimilarity value between the objects and a prototype $d(\cdot, p_q)$.

As the dissimilarities are computed to the representation set, already a dimensionality reduction is reached and therefore it can be less computationally expensive for the classification process. Furthermore, any traditional classifier that operates on feature spaces can also be used in the dissimilarity space.

3.1 Dissimilarity Measures

A general dissimilarity measure for all types of data does not exist. For each problem at hand, a dissimilarity measure adapted to the type of data should be selected. In the

case of spectral data, the connectivity i.e. continuity, ordering between the measured points, may be taken into account. In this work, we present some initial studies on dissimilarity measures for the dissimilarity representation of chemical spectral data, based on: their structures (shape changes) and/or concentration or intensity changes.

For this purpose, we studied dissimilarity measures that are more commonly used in the comparison of chemical spectral data (see Section 1). Such is the case of the very well known Manhattan (L1-norm) and Euclidean distances.

In [13], the Spectral Angle Mapper (SAM) measure (Eq. 1) was proposed for spectral data. If we have samples (spectra) $x_1, x_2 \in \mathbb{R}^n$, the SAM dissimilarity is computed as follows:

$$d(x_1, x_2)_{sam} = \arccos\left(\sum_{j=1}^{n} x_{1j} x_{2j} \Big/ \sqrt{\sum_{j=1}^{n} x_{1j}^2 \sum_{j=1}^{n} x_{2j}^2} \right). \tag{1}$$

$$d(x_1, x_2)_p = 1 - \left(\sum_{j=1}^{n} \left(x_{1j} - \bar{x}_1\right)\left(x_{2j} - \bar{x}_2\right) \Big/ \sqrt{\sum_{j=1}^{n} \left(x_{1j} - \bar{x}_1\right)^2 \sum_{j=1}^{n} \left(x_{2j} - \bar{x}_2\right)^2} \right). \tag{2}$$

The dissimilarity measure in Eq. 2 is based on the Pearson Correlation Coefficient (PCC), and measures the angle between two vectors, like the SAM measure. The PCC can be also seen as the cosine of the angle between two mean-centered samples. Although the previous dissimilarities are of the most used measures in the comparisons of chemical spectral data, the connectivity between the n measured variables is not taken into account in neither of them. The variables could be easily reordered and the same dissimilarity value is obtained.

The Kolmogorov-Smirnov distance (KS) (Eq. 3) is a dissimilarity measure between two probability distributions:

$$d(x_1, x_2)_{ks} = \max_j \left(\left| \hat{x}_{1j} - \hat{x}_{2j} \right| \right). \tag{3}$$

$\hat{x}_{1j}$ and $\hat{x}_{2j}$ are the cumulative distribution functions of the object vectors. Spectra need to be normalized to unit area, thus the areas under the original distribution of the data can be compared and their shape reflected.

In [8], the authors propose to compute the Manhattan measure on the first Gaussian derivatives (Eq. 4) of the curves (Shape measure), to take into account the shape information that can be obtained from the derivatives:

$$d(x_1, x_2)_{shape} = \sum_{j=1}^{n} \left| x^{\sigma}{}_{1j} - x^{\sigma}{}_{2j} \right| \quad \text{with} \quad x^{\sigma} = \frac{d}{dj} G\left(j, \sigma\right) * x \,. \tag{4}$$

where $*$ denotes convolution and σ stands for a smoothing parameter.

4 Experimental Section and Discussion

To evaluate the performance of different classifiers, a comparative study will be made with the three different representations of the data and four classifiers: 1-NN, RLDA, SIMCA and SVM. All the experiments were performed in Matlab. For FDA the FDAFuns toolbox was used, and the PRTools toolbox for the DR and classification of the data. For FDA, each spectrum was represented by an l order B-spline approximation, with K basis functions. The optimal values for the number of B-spline coefficients and the degree of the spline was chosen using leave-one-out cross validation. For the DR, all the samples were used as representation set.

The comparison among the models was made by the averaged error of a 10 times 10-fold cross-validation (CV), on the three representations: feature, functional (FDA), and the DR for the different dissimilarity measures presented in Section 2. For the SVM classifier, after trying with different kernels, the best results were achieved with the Gaussian kernel for Tecator dataset and the linear kernel for the rest. The regularization parameter C was optimized, as well as the number of principal components in SIMCA. To find the regularization parameters of RLDA an automatic regularization process was done. The details of all datasets are related in Table 1.

The first data set (Fig. 1a) is composed by near infrared (NIR) transmittance spectra of pharmaceutical tablets [14] of four different (classes) dosages of nominal content of active substance. In this data, the spectra of the samples of the different classes

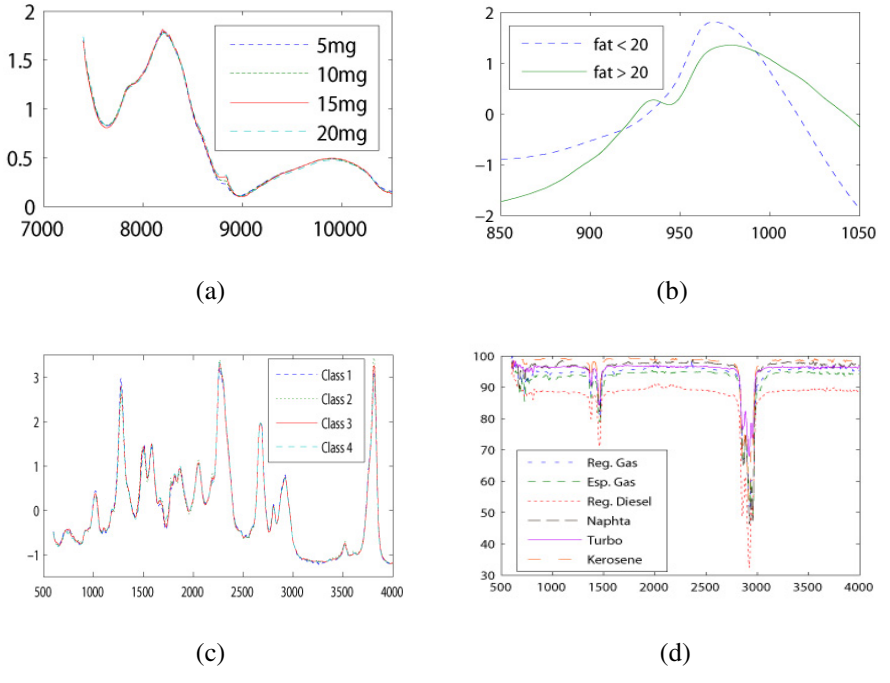

Fig. 1. Spectrum of one sample from each of the classes of each datasets: a) Tablet, b) Tecator, c) Oil and d) Fuel

are very similar, they variate in the intensity of only one peak at 8830 cm^{-1}. This peak corresponds to the only visually characteristic band of the active substance. Multiplicative scatter correction (MSC) was used as preprocessing method.

The second, named Tecator [15] (Fig. 1b), consists of NIR absorbance spectra of meat samples. In this data, the samples of the two classes differ in their fat content which is reflected in changes in the shape of the spectra (structure). Standard Normal Variate (SNV) was used as preprocessing method. The second derivative of the spectra is computed on the functional representation.

The third dataset consists of oil samples of different origins, analyzed by Mid-Infrared (MIR) technique [16] and was transformed to have zero mean and unit variance. The variations in the spectra of the classes are based in the difference in concentration of some substances and some shape changes also exist.

And the last dataset consists of fuel samples of Fourier Transform Infrared (FT-IR) transmittance spectra; base line correction and smoothing were performed on the data. The samples of these classes differ in the substances by which they are composed (structure), and therefore they differ in shape.

Table 1. Details about the # samples, features and samples per class of each dataset. The last column is related to the # basis functions used for the FDA of each dataset.

DataSet	#Samples	#Features	# Samples per Class	#Basis Functions
Tablets	310	404 (7400 to 10500 cm^{-1})	Types: A (5mg), B (10mg), C(15mg) and D(20mg)	100
Tecator	215	100 (850-1050 nm)	Fat content: Low (77) , High (138)	48
Oils	80	571 (600-4000 cm^{-1})	Origin: A (18), BB (8), BC (29) and D (25)	100
Fuels	80	3528 (600-4000 cm^{-1})	Type: Regular Gasoline (16), Especial Gasoline (15), Regular Diesel (16), Naphtha(16), Turbo Diesel(8) and Kerosene(9)	300

As can be seen in Table 2, in general for the four datasets, the SVM shows good results on all the representations, outperforming the rest of the classifiers. These could be due to these datasets are mostly non-linear. The exception is Tablets, where RLDA seems to outperform the other classifiers for its feature and functional representation, but in the DR, SVM again shows superiority. The experiments show that, most of the time, for most classifiers, their accuracy improves when using the DR and functional representation of these datasets. This demonstrates the importance of a good and descriptive representation of the data. In the case of DR, the results depend on whether a suitable dissimilarity measure is used to explain the discriminative characteristics of the curve, in order to obtain a better and more reliable classification of the data. It is worth to notice that, for both representations, the dimensionality of the datasets are reduced to half (or more) of the dimensionality of the feature representation. From the comparison of the different dissimilarity measures used, we can observe that very good results are achieved with the Shape dissimilarity, in which connectivity and shape information are considered. This proves the fact outlined in the previous paragraph, and suggests that this dissimilarity measure could be a good option for our purpose.

If we compare the results with the functional representation (FDA) and the DR of the data, they show that both approaches are good when the shape variations between the samples of different classes are appreciable. But it can be observed that, the DR gives the best results for most datasets (with the Shape measure). It shows the capability of the Shape measure, which performs well not only in datasets where the differences are based in changes in the curvature of the spectra, but also when concentration or intensity changes are present. On the other hand, in datasets like Tablet, where the functional information to be extracted is very poor, the FDA does not work very well. This lack of information in the functional data, can also be due to some of the information could have been lost by using only the coefficients resulting from the projection of the function in the B-spline basis.

Table 2. Averaged CV error with its standard deviation (%). The results are shown for the four classifiers on the feature, functional, and DR of each dataset for the six dissimilarity measures presented. The numbers highlited in bold and underlined, stand for the lowest error among all the representations for each classifier. In the case of the dissimilarities, the one that performs best in general for each dataset is also highlited in italic.

Data Sets		Feature	FDA	Dm	De	Dsam	Dpcc	Dks	Dshape
Tablets	1-NN	12,9(0,18)	**9(0,15)**	48,2(0,03)	13(0,02)	25,1(0,01)	13(0,02)	14,5(0,01)	***15,7(0,06)***
	RLDA	9,9(0,06)	10,6(0,09)	6,8(0,02)	11(0,1 e^{-17})	15,8(0,01)	8,4(0,03)	30,3(0)	***5,1(0,1 e^{-17})***
	SIMCA	25,7(0,16)	23,3(0,27)	17,2(0,02)	16(0,03)	20,2(0,06)	35,4(0,03)	26,5(0,02)	***10,7(0,03)***
	SVM	13,6(0,03)	16(0,09)	5,1(0,01)	5,3(0,03)	6,8(0,1 e^{-17})	14,8(0,02)	14,1(0,02)	***5,1(0,01)***
Tecator	1-NN	3(0,17)	2,2(0,17)	5,3(0,14)	5,3(0,19)	***1,9(***0,04)	11,2(0,04)	11,1(0,04)	3,3(0,04)
	RLDA	4,7(0,02)	3,5(0,2 e^{-17})	4,7(0,09)	4,7(0,09)	1,4(0,19)	3,8(0)	15,6(0,19)	***1,4*****(0,04)**
	SIMCA	2,5(0,12)	**2**(0,2)	9,4(0,09)	9,8(0,4 e^{-17})	***2,4***(0,9 e^{-17})	16,8(0,9)	15,3(0,9)	3,2(0,04)
	SVM	1,9(0)	**1**(0)	1(0,04)	2,8(0,2 e^{-17})	***1***(0,2 e^{-17})	1,9(0,04)	4,7(0,2)	1,4(0,1 e^{-17})
Oils	1-NN	13,8(0,32)	**7,5(0,19)**	11,1(0,51)	13,1(0,47)	7,4(0,44)	13,1(0,47)	17,4(0,29)	***9,4(0,47)***
	RLDA	22,4(0,13)	20(0,4 e^{-17})	22,8(0,25)	21,4(0,12)	22,6(0,13)	23,6(0,12)	19(0,25)	***18,6(0)***
	SIMCA	7,9(0,56)	**6,6(0,62)**	16,3(0,81)	15,6(0,43)	17,9(0,42)	17(0,46)	19,2(0,62)	***14(0,36)***
	SVM	6,3(0)	**2,5(0)**	13,8(0,2)	15,9(0,37)	8,9(0,13)	8,8(0,4)	19,8(0,12)	***6,3(0)***
Fuel	1-NN	35,1(2,08)	17,7(1,71)	9,5(0,62)	33,3(0,75)	20,1(0,54)	14(0,52)	30,2(0,58)	***8,6(0,42)***
	RLDA	22,5(0)	21(0,79)	15,1(0,54)	39,8(1,16)	15,5(0,86)	19,6(0,42)	43,1(1,02)	***16,9(0,75)***
	SIMCA	30,4(3,73)	12,4(1,61)	12(0,38)	40,5(0,82)	20,4(0,65)	20(0,91)	57,5(0,49)	***11,9(0,43)***
	SVM	10(0,04)	7,5(0,4 e^{-17})	8,6(0,12)	25,3(0,25)	13(0,50)	16(0,25)	35,1(0,13)	***5,5(0,50)***

In the case of Tecator dataset, good results are achieved either with the FDA representation or the DR (for the different classifiers); there is barely a difference between the errors committed for some classifiers when operating on them (looking also at the standard deviation error). Nevertheless, FDA performed better in general. It can be explained by the fact that, from the functional point of view, a lot of information can be obtained when shape changes are present in the curve. So the FDA by B-splines is capable of using this information and the use of the second derivatives afterwards emphasizes the peaks in the curve, making easier to see the differences. In the Fuel dataset, a similar result could be expected if the same procedure is carried.

However, in spite of the good performance of the DR for most cases, this is not the case for Oil dataset. This suggests that, although the dissimilarity measures have shown their ability to discriminate between spectra that are very similar (see Tablet dataset in Fig. 1a); they might not be robust enough for cases like this, where the shape varies so abruptly and so frequently in the spectrum. Still, the results could be improved if the DR is computed on the FDA representation. Further researches most be done on this aspect.

5 Conclusions

We presented two alternative ways to improve the representation of chemical spectral data. The first makes use of the spectral connectivity by approximating the spectra by spline functions (FDA). The second makes use of the physical knowledge of the spectral background of the data by modeling their relations in a dissimilarity representation. Comparisons were made by classifying four chemical spectral datasets, expressed by their feature and the two other representations. It was shown that, with the studied representations, improved classification results can be obtained. But it shows that the use of either of them will depend on the characteristics of the data. We can also conclude that, for the comparison of spectral chemical data by their dissimilarities, the better results are obtained with measures that take the connectivity between the points, and shape information into account.

References

1. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis, New York (1997)
2. Pekalska, E., Duin, R.P.W.: The Dissimilarity Representation For Pattern Recognition. Foundations and Applications 64 (2005)
3. Cardot, H., Ferraty, F., Sarda, P.: Functional linear model. Statist. Probab. Lett. 45, 11–22 (1999)
4. Preda, C., Saporta, G.: PLS regression on stochastic processes. Comput. Statist. Data Anal. 48, 149–158 (2005)
5. Cérou, F., Guyader, A.: Nearest neighbor classification in infinite dimension. ESAIM: Probability and Statistics 10, 340–350 (2006)
6. Villa, N., Rossi, F.: Support Vector Machine For Functional Data Classification. In: ESANN 2005 (2005)
7. Hernández, N., Biscay, R.J., Talavera, I.: Support Vector Regression Methods for Functional Data. In: Rueda, L., Mery, D., Kittler, J. (eds.) CIARP 2007. LNCS, vol. 4756, pp. 564–573. Springer, Heidelberg (2007)
8. Paclik, P., Duin, R.P.W.: Classifying spectral data using relational representation. In: Spectral Imaging Workshop, Graz, Austria (2003)
9. Varmuza, K., Karlovits, M., Demuth, W.: Spectral similarity versus structural similarity: infrared spectroscopy. Anal. Chimica Acta 490, 313–324 (2003)
10. Komsta, L., Skibinski, R., Grech-Baran, M., Galaszkiewicz, A.: Multivariate comparison of drugs UV spectra by hierarchical cluster analysis-comparison of different dissimilarity functions. In: Annales Universitaits Marie Curie-Sklodowska, Lublin, Polonia, vol. 20, pp. 2–13 (2007)
11. Wold, S.: Chemometrics: Theory and Application. In: Kowalski, B.R. (ed.) ACS Symposium, vol. 52, pp. 243–282 (1977)
12. Yuhas, R.H., Goetz, A.F.H., Boardman, J.W.: Discrimination among semiarid landscape end members using the spectral angle mapper (SAM) algorithm. In: Third Annual JPL Airborne Geoscience Workshop, Pasadena, CA, pp. 147–149 (1992)
13. Pekalska, E., Duin, R.P.W.: Prototype selection for finding efficient representations of dissimilarity data. In: Kasturi, R., Laurendeau, D., Suen, C. (eds.) International Conference on Pattern Recognition, Quebec, Canada, vol. 3, pp. 37–40 (2002)
14. Tablets dataset, `http://www.models.kvl.dk/research/data`
15. Tecator dataset, `http://lib.stat.cmu.edu/datasets/tecator`
16. Oil dataset, `http://cac2008.teledetection.fr/shootout`