Classification of continuous multi-way data via dissimilarity representation

Classification of continuous multi-way data via dissimilarity representation

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft;
op gezag van de Rector Magnificus prof.ir. K.C.A.M. Luyben;
voorzitter van het College voor Promoties
in het openbaar te verdedigen op 15 oktober 2013 om 12.30 uur

door

Diana PORRO MUÑOZ

Computer engineer van
Technische Universiteit "José Antonio Echeverría" (CUJAE)
geboren te Camaguey, Cuba

Dit proefschrift is goedgekeurd door de promotor:
Prof. dr. ir. M. J.T. Reinders

Copromotor: Dr. ir. R.P.W. Duin

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. ir. M. J.T. Reinders,	Technische Universiteit Delft, promotor
Dr.ir. R.P.W. Duin	Technische Universiteit Delft, copromotor
Prof.dra. I. Talavera Bustamante	Universiteit van Havana
Prof.dr. H.A.L. Kiers	RU Groningen
Prof.dr. A.K. Smilde	Universiteit van Amsterdam
Prof.dr. J.N. Kok	Universiteit Leiden
Prof.dr. L.F.A. Wessels	Technische Universiteit Delft
Prof.dr.ir. G. Jongbloed	Technische Universiteit Delft, reserve lid

This work was partly supported by the Foundation for Neural Networks SNN, as well as by the FET programme within EU FP7, under the SIMBAD project (contract 213250).

ISBN/EAN: 978-94-6186-220-4
© 2013, Diana Porro-Muñoz, all rights reserved.

Classification of continuous multi-way data via dissimilarity representation

Thesis

presented for the degree of doctor
at Delft University of Technology
under the authority of the Vice-Chancellor,
prof.ir. K.C.A.M. Luyben,
to be defended in public in the presence of a committee
appointed by the Board for Doctorates
on 15 October 2013 at 12.30

by

Diana PORRO MUÑOZ

Computer Science Engineer from
Technical University "José Antonio Echeverría" (CUJAE)
born in Camaguey, Cuba

This thesis is approved by the supervisor:
Prof. dr. ir. M. J.T. Reinders

Adjunct supervisor: Dr. ir. R.P.W. Duin

Composition of the Doctoral Examination Commitee:

Vice-Chancellor,,	chairman
Prof. dr. ir. M. J.T. Reinders,	Delft University of Technology, supervisor
Dr.ir. R.P.W. Duin	Delft University of Technology, adjunct supervisor
Prof.dra. I. Talavera Bustamante	University of Havana
Prof.dr. H.A.L. Kiers	RU Groningen
Prof.dr. A.K. Smilde	University of Amsterdam
Prof.dr. J.N. Kok	Leiden University
Prof.dr. L.F.A. Wessels	Delft University of Technology
Prof.dr.ir. G. Jongbloed	Delft University of Technology, reserve member

This work was partly supported by the Foundation for Neural Networks SNN, as well as by the FET programme within EU FP7, under the SIMBAD project (contract 213250).

ISBN/EAN: 978-94-6186-220-4
© 2013, Diana Porro-Muñoz, all rights reserved.

Contents

Chapter 1

Introduction

Within analytical chemistry, many new technologies are being generated over the last years, giving scientists huge opportunities to analyze and describe their problems. Such is the case for sensors, spectroscopic equipments like NIR Spectroscopy, NMR Spectroscopy, Mass Spectroscopy, Thermal Analysis and Atomic Spectroscopy. Such instrumental techniques allow analyzing a variety of case studies in many other branches of science, including biology, medicine, material characterization, food, environment, physics, geophysics, pharmaceutical and forensics.

Spectral information (depending on the applied instrumental technique) can be seen as a distinctive signature of objects. Therefore, spectra are often a source of challenging classification problems.

Classification is one of the most common tasks of pattern recognition. It is based on learning the structure of groups of objects with similar patterns. In this learning process, it is assumed that there are training examples that are representative and contain sufficient information to find a good (generalizable) model for predefined groups/classes of those examples. Therefore, new unseen objects can be assigned into these groups reliably. The formalized representation of objects is an important aspect in the determination of a good class description.

In the case of spectra and other processes/objects we are interested in, they have a shape as a function of e.g. time/position/frequency. They are often continuous functions (they don't jump) and it is in fact their shape that is significant for the classification. However, they are traditionally represented by sampling, i.e. sequence of individual observations (feature vector), such that an object is represented in a high-dimensional space. This representation of a spectrum in a feature [1] vector is, not optimal as it, generally, considers each feature independently. The continuous nature[2] of the data is ignored, therefore it is difficult to find discriminative spectral characteristics.

When it comes to analytical chemistry, the amount of objects available for training may range from billions in geophysics to only a few in chemical applications. However, research areas like these last two are a 'here and now' subject. Therefore, most of the time only a limited amount of objects is available, because experiments are very time consuming and costly, and often answers are required immediately [23]. When there are a limited amount of training objects, they are usually not sufficient to estimate the number of parameters required to find good models for the classes in these high-dimensional spaces. This phenomenon, known as the curse of dimensionality [45, 109], makes many statistical pattern recognition methods fail. Furthermore, if only a few objects are available, it becomes even more difficult to recover information about the type of data. A good separation/discrimination between classes is hard

[1] From now on, we will use the word feature to denote measurements or attributes of the objects. Thus, feature vector representation/ feature space will refer to objects represented by these types of features.

[2] By 'continuous nature data' it is meant: a sampling of a continuous function (which can be 1-d, 2-d or n-d function with well defined derivatives) of which the shape is informative for the class.

to obtain. These observations call for a new representation for spectral data that takes the information on their continuous nature into account.

Moreover, the raw spectra are usually contaminated with high-frequency noise. The topic of feature reduction (e.g. feature selection, extraction), aims at tackling this problem. Nevertheless, one could search for an alternative representation that in some way could also address this problem, if it encapsulates the information captured in a lower dimensionality representation of the original data.

The previously described problems are even more remarkable when it comes to multi-dimensional spectral data sets. Representation of objects by higher-order generalizations of vectors and matrices has become very common for many research areas e.g. image analysis, signal processing and chemometrics. The main issue with this type of data is how to perform a proper analysis. Multi-way data analysis is the field specially dedicated to it. In the next section, a brief review of this field will be given, focusing on the main contribution area of this thesis.

1.1 Multi-way data analysis

The early ideas on what would be later known as Multi-way data analysis, date back to the 50's in the psychometrics field, when Raymond Cattell [29] introduced the term of multi-way arrays. Usually, multivariate data has a two way structure (matrix), where there are a number of objects (rows) described by a set of measurements/properties (columns). For a wide variety of problems, the structure of the data can often, however, be more complex; one can have several sets of properties measured on different objects, e.g. data collected at different times or conditions. Such data would be more appropriately represented by higher-order generalization of vectors and matrices $\underline{Y} \in \mathbb{R}^{I_1 \times I_2 \times I_3 \times \ldots \times I_n}$, $n > 2$, which are the so-called multi-way arrays (See Figure 1.1).

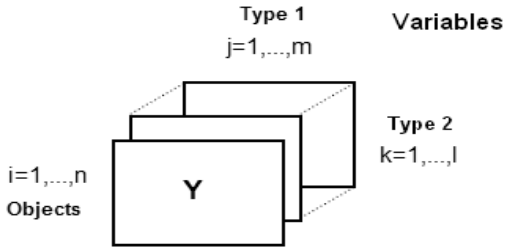


Figure 1.1: Design of a three-way array

The terminology for this type of data is different from that of two-way arrays. Each dimension/direction of a multi-way array is known as a mode or way and the number of attributes (according to the type of data these attributes can be objects or measurements) in each mode is used to indicate its dimensionality. Specifically, in three-way data the parts of the array are called rows, columns and tubes. In general, they are all called fibers [60] (See Figure 1.2).

There can be different designs of multi-way data, depending on what the features on the different modes are. One of the most common types of multi-way data is defined by Kroonenberg [64] as profile data. In this design, objects are always in one of the modes and the features used to describe these objects are found in the rest of the modes(See Figure 1.1). Note that these data sets are not the same as a stack of images (which can be invariant to shifting or rotation). In multi-way data, each position in the multi-dimensional structure of each object has a physical meaning, thus features should be aligned. In this thesis we restrict ourselves to the profile design of multi-way data. Lets take an example for chemistry, in which objects can

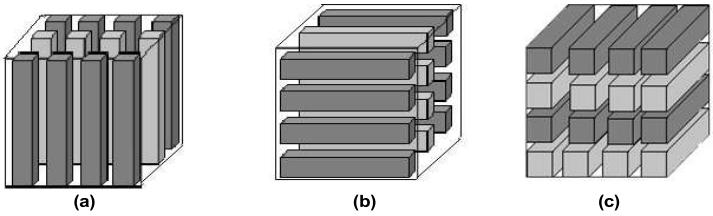


Figure 1.2: Definition of (a) columns, (b) rows and (c) tubes in three-way data

be measured by instrumental analytical tools like fluorescence emission-excitation spectroscopy and Gas Chromatography/Mass Spectrometry (GC/MS). A two-way array from a single object is generated, leading to a three-way array if all information is put together. This multi-way data can then be used to solve regression and/or classification tasks for the measured objects. With these purposes, it is important to employ proper tools in order to analyze the multi-way data.

An early and still common procedure to deal with multi-way data is called unfolding or matricization [60, 126, 64]. It is based on creating a two-way matrix out of the multi-way array, by laying out all matrices (slices) from one mode next to each other. For example, assume we have a three-way array $\underline{Y} \in \mathbb{R}^{I \times J \times K}$, which is rearranged into a matrix $Y \in \mathbb{R}^{I \times JK}$. Any two-way analysis tool can be used afterwards. However, multi-way data will not be analyzed optimally in this way, i.e. the information of the multi-way structure is lost and the dimensionality of the problem $(J * K)$ becomes huge.

Multi-way analysis is the extension of multivariate analysis (two-way data analysis), when data are arranged in multi-way arrays [126, 64, 31]. Development of specialized multi-way analysis techniques date back to psychometrics in 1964, when Ledyard Tucker [132] introduced the three mode factor analysis with the TUCKER models. These early methods have a strong exploratory character and are used for extracting hidden structures and explore the interrelations in the data. The multi-way component analysis can be considered as an extension of Principal Component Analysis (PCA) when the analyzed data is in the form of multi-dimensional arrays.

Let us introduce the multi-way component analysis by generalizing the singular value decomposition. A two-way component analysis of a matrix $\mathbf{X}(I \times J)$ with elements x_{ij}, based on a singular value decomposition of F components can be defined as:

$$x_{ij} = \sum_{f=1}^{F} a_{if} g_{ff} b_{jf} \tag{1.1}$$

where $\mathbf{A}(I \times F)$, with elements a_{if} is called the score matrix (scores on the components, projection of objects in the components space), and $\mathbf{B}(J \times F)$, with elements b_{jf} the loading matrix (weights used to form the linear combinations of the original features, transformation matrix), which are both orthogonal matrices. $\mathbf{G} = diag(g_{11}, \ldots, g_{FF})$ is a diagonal matrix called the singular value or core matrix containing the F largest singular values of $\mathbf{X}$. This indicates that only the fth column of $\mathbf{A}$ is interacting with the fth column of $\mathbf{B}$. [126]

Likewise, the TUCKER3 model of a three-way array $\underline{Y}(I \times J \times K)$ with elements y_{ijk} is given by one score matrix $\mathbf{A}$, and two loading matrices $\mathbf{B}$, and $\mathbf{C}$ (in multi-way analysis literature they are usually referred as three loading matrices) with typical elements a_{ie}, b_{jf}, and c_{kh}:

$$y_{ijk} = \sum_{e=1}^{E} \sum_{f=1}^{F} \sum_{h=1}^{H} a_{ie} b_{jf} c_{kh} g_{efh} \tag{1.2}$$

where the notation (E, F, H) is used to indicate that the model has E, F and H factors in the

three different modes. Thus, the column dimensions of the loading matrices can be accommodated individually in each mode. In this model, the core-matrix $\underline{G}$ is a three-way non-diagonal matrix, which explicitly means that unlike in traditional PCA, there are interactions between factors [24, 126, 64]. This core-matrix contains the strength (or weight) of the linkage between the components of the different modes [64]. The Tucker3 model is best seen as a way to generalize PCA to higher orders, i.e. its usefulness rests in its capability to compress variation, extract features, explore data, etc [24].

Besides giving the magnitudes of interactions, the core can be considered an approximation of $\underline{Y}$. It approximates the variation of the original array by expressing it in terms of the truncated basis matrices **A**, **B**, and **C**. An approximation of the original data can thus be obtained by transforming $\underline{G}$ back into the original space as in equation 1.2 [24].

One of the problems of this model is that the components are not unique because they can be rotated or non-singularly transformed in the same way that they can be transformed in two-mode PCA. This gives the model a higher complexity, affecting also the interpretability.

Another important method in the multi-way analysis field is known as parallel factor analysis, PARAFAC [53] (also proposed as CANDECOMP [27] in a different context). The idea behind PARAFAC is also based on component decompositions of n-way data. Unlike the TUCKER3 method, the same components are used to describe the variation in several matrices simultaneously but with different proportions or scales depending on the conditions. This entails the most attractive aspect of the PARAFAC method; a unique solution is found. However, a not very appealing characteristic of this method are the so-called degenerate solutions. Sometimes the algorithm converges very slowly (it is even not guaranteed to converge). In such cases, the TUCKER3 method can be a better option.

Several extensions of PARAFAC and TUCKER3 have been proposed [4, 64]. They aim at relaxing some of the restrictions of the former methods. Within chemometrics, some methods for second order calibration and/or regression that are generalizations of two-way calibration methods to multi-way data have been developed [126], e.g. Multi-way Partial Least Squares (NPLS) [126].

Surprisingly, multi-way classification tools have not been much studied, although classification problems appear frequently in many research areas. There are three main strategies to tackle these problems when the data has a multi-way structure:

Unfolding + Classification: The multi-way array is unfolded in the feature modes e.g. if we have a three-way array structured as $objects \times features1 \times features2$, the features modes are concatenated next to each other. Any traditional multivariate classification tool can be used afterwards [11, 12]. Disadvantages of this approach for multi-way data analysis were previously mentioned in this introduction.

Multi-way Decomposition method + Classification: Multi-way data structure is used and dimensionality of data reduced by decomposing the array with a multi-way factor decomposition tool such as PARAFAC or TUCKER3 [62]. The new representation of objects (score matrix from decomposition) is then used as input for traditional classifiers.

Multi-way classifier: To the extent of our knowledge, the few multi-way classifiers developed until today are based on an integration of the decomposition and the classes information, unlike the other two strategies, where the representation and classification can be seen as two different steps. Such is the case of Multi-way Partial Least Squares Discriminant Analysis (NPLS-DA) [11] and N-SIMCA [37] classifiers. They are both extensions of the former PLS-DA and SIMCA methods for two-way data to multi-way data. In the case of NPLS-DA, the N-PLS regression method aims at establishing a linear relationship between a set of explanatory variables,i.e. measurements taken on the object, and

response/dependent variables, e.g. some type of properties. Although the PLS method is primarily regarded as a calibration method, it can also be used for classification. In such case, the dependent variables are the classes, so the regression is performed to a binary label. For N-SIMCA classifier, like for two-way SIMCA [154], a separate classification model by means of a component analysis is built for each of the classes and classes boundaries are estimated to classify the new unseen objects. In this case, multi-way decomposition methods e.g. PARAFAC, are applied.

Except for the unfolding approach, the others succeed in exploiting the interrelations between the modes in the complex multi-way structure for the classification.

Despite the capabilities of multi-way analysis methods, there is an important aspect to be taken into account. They do not consider important context/background information from objects in their analysis. Such is the case of multi-way spectral data or any continuous nature data, which have a particular shape that characterizes them. However, as for simple spectra, in multi-way spectral data analysis, no assumption on the continuous nature of spectra is made. Methods are insensitive to any ordering in the data samples e.g. wavelengths or product concentration. Therefore, discriminative context information from their original nature is not taken into account in the representation, and thereby not in the classification. Thus, we wonder if an alternative representation for multi-way spectral data could tackle these problems.

1.2 Fundamentals of Dissimilarity Representation

The concept of proximity plays an important role in any pattern recognition problem. A class of objects is defined as a collection of objects with similar characteristics. Therefore, the process of classifying an unknown object is based on determining how similar it is to a set of objects from which we know their class label.

There are many ways on how to compare objects, and it eventually depends on what we consider that makes objects similar. This implies that the suitability of a proximity measure depends on the problem at hand; its definition should be based on the knowledge one has about the data.

A (dis) similarity can be seen as a function that assigns high (low) values to alike objects and low (high) values to objects that have distinct characteristics. Therefore, a large similarity and a small dissimilarity both mean the same thing with respect to objects comparison, i.e. similar objects. There are several manners to switch from one to the other, thereby studies can be based on one of them solely. From now on, we will refer to dissimilarities to be consequent with the approach to be described in this Chapter.

Given the intuitive notion of dissimilarity, there is a set of properties that a dissimilarity measure can fulfill. Let X be a set of objects and $d : X \times X \to \mathbb{R}$ a dissimilarity measure, for all $x, y, z \in X$, some of these properties can be defined as follows:

1. Positivity: $d(x, y) \geq 0$
2. Reflexivity: $d(x, x) = 0$
3. Definiteness: $(d(x, y) = 0) \Rightarrow (x = y)$
4. Symmetry: $d(x, y) = d(y, x)$
5. Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$

If a dissimilarity measure fulfills the five properties above, it is said to be a *metric*. But this is just a particular case of dissimilarity measures; depending on the fulfilled properties, a

more/less restrictive type of measure can be defined. The more properties a measure fulfills, a better description/information about the measure we have, allowing for a better understanding of its behavior. Moreover, a higher number of mathematical tools is available to analyze the data.

Nevertheless, when facing an applied problem, the expressivity of the measure can be more important than the mathematical properties it holds. The applied dissimilarity measure should be chosen or designed such that background knowledge on the data could be included, thus resembling the differences between the objects as good as possible. However, it is many times difficult to find a measure that accomplishes this and holds the desired constraints. Therefore, in practice the fulfillment of many of these (or other) properties might not be possible.

A new research field within pattern recognition that deals with different types of dissimilarity measures (does not demand metrics) is known as Dissimilarity Representation (DR) [93].

The field of DR deals with finding a representation of objects, such that prior knowledge on the classes as well as their structural information can be included into the learning process. The very basic assumption is that objects from the same class should be similar to each other, and that objects from different classes should be different from each other. This is called the compactness property. An example of including context knowledge is to take into account that spectra are ordered features, when comparing spectral data. Exploiting this context increases the informativeness of the representation. Consequently, DR links the statistical and structural approaches for pattern recognition. See [94] for more details on the fundamentals of the DR.

A dissimilarity representation basically consists in representing objects by its dissimilarities with respect to other objects. Assume we have an $N \times N$ dissimilarity matrix $\mathbf{D}$. This matrix contains the dissimilarity values $d(x_i, x_j)$ between all objects $x_i \in \mathbf{X}$. Once we have the dissimilarity-based representation there are three ways to build a classifier:

Nearest Neighbor Rule (k-NN): A very known and commonly used approach based on dissimilarities is the nearest neighbor rule [32, 45]. The k-NN rule assigns a new object to a class that most occurs among its k nearest neighbors, by interpreting the dissimilarities directly.

Dissimilarity Space: Dissimilarities are considered to be the *attributes/features* of the objects (not the distances between the objects). Consider a mapping $\phi(\cdot, R) : \mathbb{R}^{m \times l} \rightarrow \mathbb{R}^h$, such that for every object $x_i \in \mathbf{X}$, a representation based on its dissimilarities to the objects from a representation set $\mathbf{R} = \{r_1, r_2, ..., r_h\}$ is obtained: $\phi(x_i, R) = [d(x_i, r_1), d(x_i, r_2), ..., d(x_i, r_h)]$. $\mathbf{R}$ can be $\mathbf{X}$ itself or a set of representative objects (prototypes) $\mathbf{R} \subseteq \mathbf{X}$. Such objects can be found by a prototype selection method [94]. The features of an object are then its distances to each of the prototypes. Hence, now the complexity (dimensionality) is defined by the number of objects in the representation set. Classifiers are then built in this vector space in a similar way as in the traditional feature space. [91, 93].

Pseudo-euclidean Embedding: It is another alternative to tackle dissimilarity data. It consists in embedding the dissimilarities into some vector space, imposing that the distances in the new space reflect the original ones. Assuming any symmetric dissimilarity representation is available, it is possible to find a distance preserving mapping onto a pseudo-Euclidean space [94].

Although it seems to be a promising approach, embedding new unobserved objects into the pseudo-Euclidean space is still problematic. See [94] for further details. On the other hand, studies have shown that the dissimilarity space approach often outperforms the nearest neighbor rule [89, 91]. The k-NN classifier does not use the dissimilarities between objects of the training

set when classifying new objects, discarding the chance of gaining some knowledge from them. The dissimilarity space and embedding approaches however, aim at having the possibility of using this information, thus trained classifiers can be designed.

For spectra, the dissimilarity-space approach has the advantage that the number of features can be controlled by the number of prototypes that one uses. This is favorable because the number of objects in spectral data sets are generally much smaller than the number of samples in spectra. In this way, a clever dimensionality reduction is performed, while still using the complete spectra to calculate the dissimilarities. Therefore, in this thesis, we will use the dissimilarity space approach for the analysis of spectral data. As an additional advantage, this approach leaves complete freedom with respect to the choice of the classifier.

1.3 Motivations and Goals of the thesis

So far, important problems that have been detected in the representation and/or classification of simple (1D) and multi-way spectral data have been explained in Section 1. Moreover, the Dissimilarity Representation approach and its potential for data with similar characteristics to those of spectral data sets were mentioned. The advantages of this approach, including the fact that analysis of spectral data can be enriched by taking into account knowledge about the domain, have led us to devote this thesis to investigate the DR as a new tool for classifying multi-way spectral (or more in general, continuous) data.

Besides the main difficulties in the actual representation and for the classification of continuous multi-way data, there are other issues that can affect their analysis. In this kind of data, it is very common to have non-informative, noisy or redundant information. This is usually counter-productive for classifiers. In the best case, when it does not influence their performance, it causes the computational complexity to increase. Thereby, it is common to make a selection of the most contributive features to the modeling of the problem at hand.

Another problem is related to the frequent presence of the so-called missing values in this type of data. In contrast with the 1D case, there is limited work addressing the problem of missing values in multi-way data. Most of the related studies are dedicated to factorization methods [126, 64]. This issue has also been investigated for the DR approach [79]. However, it has been studied for 1D data only, so it does not fit to multi-way data.

From the observations reported thus far, we have defined the main goals of this thesis:

- Study and development of new alternative representations of spectral data, which contribute to a better discrimination of their classes by taking into account the information of its continuous nature.

- Development of a new tool for the classification of multi-way data, based on the Dissimilarity Representation approach. Development of new dissimilarity measures for multi-way continuous data, which make use of the information in the multi-dimensional structure and the nature of the data.

- Development of alternatives to tackle common problems in this type of data, which can affect the accuracy of the dissimilarity-based classification. Such is the case of the high-dimensionality with noisy, redundant features and the presence of missing values.

1.4 Outline

In Chapter 2, advanced representations for spectral data sets are studied. Section 2.2 aims at showing that classification of spectral data can be improved by including the information on the continuous nature of spectra in the process. As a first approach, this information is taken into account by approximating spectra with spline functions. In the second approach, we make use of the physical knowledge of the spectral background of the data by modeling their relations in a dissimilarity representation. Different dissimilarity measures for spectral data were analyzed. This study has been published in [97]. Finally, a new approach is introduced in Section 2.3, which is based on incorporating the continuity information into the DR by functional data descriptors. A comparison of the feature-based representation and various DR on different examples of spectral classification problems is carried out. It is shown in which situations each of the representations can be more suitable. This work has been published in [101].

In Chapter 3, a more general technique for classifying multi-way data is proposed. In the first part of the chapter (Section 3.2), the DR is introduced as a new tool for this purpose. Context information from the data can be included in its design. As a particular case, a 2D measure was developed for three-way spectral data sets, based on a combination of 1D dissimilarities. It allows taking into account the functional information of the data e.g. the shape of peaks. In cases where continuity is just present in one of the ways (directions), e.g. GC-MS data sets, the measure can be adapted. This has been published in [100].

In Section 3.3, a more thorough study on DR for spectral data (simple and multi-way) is presented. It is shown how the curse of dimensionality phenomenon, which is typical in this type of data sets, can be tackled by the DR. This work has been published in [102]. By using a suitable dissimilarity measure (shape-based for spectra), classes are more compact and only a few number of objects is enough to achieve good classification results.

A modification of the DR for three-way data is introduced in Section 3.4. Spectral data is often contaminated by a large amount of noise and redundant information. Improvements in the classification accuracy and a reduction in the computational cost of the dissimilarity measure could be achieved by neglecting this non-informative information. Section 3.4 studies the effects of selecting just a part of the spectral data (most discriminative peaks) for the computation of the DR. This approach is compared with that on the full data and the traditional multi-way classification approaches (See Section 1.1). This paper has been resubmitted for the second review after a major revision [104].

The last part of this chapter (Section 3.5) presents a new dissimilarity measure for 2D continuous data (CMS), which also takes the 2D shape structure of objects into account. This measure can be easily extended to multi-way objects. It is important for the generalization of the dissimilarity-based classification of multi-way data. This paper has been published in [103].

The presence of missing data is very common in many real world applications. This is a problem for classifiers, as it can lead them to poor performances. Chapter 4 is dedicated to the treatment of missing values for the classification of multi-way data based on dissimilarities. We study the performance of the dissimilarity-based classifiers after applying different imputation methods. A modification of the CMS measure is also proposed in order to deal with incomplete data, without the need of dependencies and extra computations of reconstructing the missing parts. Depending on the reason for these values to be missing, they can also appear with different patterns. We analyze how the different methods perform for the different patterns of missing data. A reduced version of this chapter has been accepted for publication in [105].

1.5 Main contributions

The principal aim of this work is to find a novel tool for the classification of multi-way continuous data, such that not only the information in its multi-dimensional structure is used, but the contextual information that can help for its better discrimination.

Although multi-way data analysis has been a hot topic for some years now, little attention (as far as we know) has been paid to the classification task and/or including background knowledge information in the analysis. We propose the use of the Dissimilarity Representation (DR) approach as a new tool for the supervised classification of multi-way data, more specifically multi-way continuous data. This approach, besides implying an alternative powerful representation for this type of data, opens a wide range of possibilities to classify multi-dimensional objects. Any traditional classification method can be applied on dissimilarity spaces.

As a preliminary work, in order to establish a solid foundation for this approach, alternative representations were studied for simple spectral data, such that their discriminative shape information could be used in the analyses. A new representation based on the combination of the dissimilarity representation and functional data analysis is introduced for spectral data.

In order to apply the dissimilarity-based approach for multi-way data classification, we proposed two novel dissimilarity measures. One of them, the 2Dshape, can only be used for 2D objects, but it allows comparing 2D objects that have measurements of different nature in the different directions. The other, the continuous multi-way shape measure, can be used for multi-way continuous objects. A modification of this measure for missing values is proposed, as this is a very common problem in many applications. On the basis of optimizing the computation of the dissimilarity-based representation for multi-way data, a feature selection approach for spectral (continuous) 2D objects is introduced. However, this is independent of the DR approach, therefore it can be used as a pre-processing step in order to reduce noisy/redundant information that can affect further data analyses.

Chapter 2

Dissimilarity Representation in Spectral data

2.1 Overview

Spectral data is commonly encountered in many pattern recognition applications, e.g. based on time signals, hyper-spectral images and chemometrics. Each class of objects can have a particular spectral profile (group of objects with similar spectra), a particular shape that differs from the other classes. Therefore, an unknown object could be classified by finding its resemblance with the spectral profile of any of the previously defined classes. However, this discriminative information is not really used for the automatic analysis of spectra.

A very important aspect in pattern recognition is to find a representation of objects that is optimal for the following generalization step. It should guarantee that objects of the same class are closer than those of different classes (compactness property), such that good class models can be found [36, 44]. In the case of spectral data, they are captured and represented with high-dimensional vectors, as sequences of individual numerical values resulting from a sampling procedure. Traditional multivariate analysis tools just make a purely statistical analysis of this set of numerical values, without taking into account the structure of spectra [23]. Thus, the real important discriminative characteristic of this type of data i.e. their functional nature, the ordering of the measurements, their shape, is not considered in the classification process.

In this chapter, there are two main contributions. In section 2.2, we investigate the benefits of using advanced representations for spectral data sets. It aims at showing that accuracy in classification of spectral data can be improved, by including the information on their continuous nature in the analysis. The first studied approach is known as Functional Data Analysis (FDA) [108]. FDA is an extension of the traditional multivariate analysis for data with a functional nature, and it is based on considering the observed spectra as a continuous real-valued function instead of an array of individual observations. In the second approach, Dissimilarity Representation (DR) [94], spectra are represented by its dissimilarities to other spectra (dissimilarity space). 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 spectra i.e. shape. An analysis of dissimilarity measures that have been used for spectral comparison is performed.

The second contribution of this chapter (section 2.3) consists in incorporating the functional approximation descriptors into the design of a dissimilarity measure for spectral data sets. This way, the advantages of both approaches, namely FDA and DR, can be used for a better classification of spectral data sets. A comparison between the proposed approach and the formers is performed, in order to study the suitability of each of them for different patterns of spectra.

2.2 The Representation of Chemical Spectral Data for Classification

This section has been published as 'The Representation of Chemical Spectral Data for Classification', by D. Porro-Muñoz, I. Talavera, R.P.W Duin, and N. Hernández, in *Lecture notes in computer science vol. 5856, Springer Verlag, Berlin, 14th Iberoamerican Congress on Pattern Recognition, Proc. CIARP 2009 (Guadalajara, Mexico)*, 513–520, November 2009. In order to make the text more readable, some corrections in language were made.

Abstract

The classification of unknown objects 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 (wavelengths, m/z) have no connection, and other aspects of their functional nature e.g. shape differences, are also ignored. In this paper, we study some advanced representations for chemical spectral data sets. We make a comparison of the classification results of 4 data sets by using their traditional representation and two other: Functional Data Analysis and Dissimilarity Representation. These approaches allow taking into account the information about spectra, which is missing in the traditional representation, thus better classification results can be achieved. Some suggestions are made about the more suitable dissimilarity measures to be used for chemical spectral data.

2.2.1 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 a significant aspect 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 objects.

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

Functional Data Analysis (FDA) [108] and Dissimilarity Representation (DR) [94] 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 it 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 have been extended to work on it e.g. linear discriminant analysis (LDA) [28]. In the case of linear modeling, studies have also been made in regression [106]. A number of estimation methods for functional nonparametric classification and regression models have been introduced. Namely, k-Nearest Neighbor classifier (k-NN) [30], kernel classifiers e.g. Support Vector Machine (SVM) based on the Radial Basis Function (RBF) methods [139, 54], showing its application for chemical spectral data.

Profound studies of the DR on chemical spectral data sets have not been carried out. However, there are already some results on spectral data in general [87], demonstrating its advantages for classification. In this approach, based on the meaningful role that proximities play in the classification process, the authors propose to work on a space defined by the dissimilarities between the objects [94]. 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. Thus, the selection of a suitable measure for the particular problem is substantial. 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 papers have addressed the comparison of chemical spectral data. In [134], the authors are looking for similarity measures for infrared (IR) spectrometry. A more recent research [63] 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 classification results can be improved 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) [154] and SVM classifiers on the three mentioned representations: feature, FDA and DR of four chemical spectral data sets. We also make a study of some dissimilarity measures that have already been used on these types of data. It will help us propose those measures that 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.

2.2.2 Functional Data Analysis

Functional Data Analysis (FDA) [108] 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 features. 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 have 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 example 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:

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

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.

2.2.3 Dissimilarity Representation

The Dissimilarity Representation (DR) [94] proposes to work on the space of the proximities between the objects, instead of the space defined by their attributes (features), as it is usually done.

In the new representation, instead of having a matrix $\mathbf{X}(m \times n)$, 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 $x \in \mathbf{X}$ and the objects of the representation set $\mathbf{R}(p_1, p_2, ..., p_q)$, $d(x_m, p_q)$. The elements of R are called prototypes, and have preferably to be selected by some prototype selection method [92]. These prototypes are usually the most representative objects of each class ($\mathbf{R} \subseteq \mathbf{X}$), but the whole set of objects $\mathbf{X}$ can be used too, obtaining the square dissimilarity matrix, $\mathbf{D}\,(m \times m)$; $\mathbf{R}$ can also be a completely different set of objects.

For the DR, three main approaches exist [94]. In the first one, 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 $\mathbf{D} \subseteq \mathbb{R}^n$, 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 dissimilarities are computed to the representation set, a

dimensionality reduction is already reached. 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.

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 shape of peaks 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 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 [158], the Spectral Angle Mapper (SAM) measure (Eq. 2.1) was proposed for spectral data. If we have two examples (spectra) $x_1, x_2 \in \mathbb{R}^n$, the SAM dissimilarity is computed as follows:

$$d(x_1, x_2)_{sam} = ar\cos\left(\sum_{j=1}^{n} x_{1j}x_{2j} \Big/ \sqrt{\sum_{j=1}^{n} x_{1j}^2 \sum_{j=1}^{n} x_{2j}^2}\right) \tag{2.1}$$

$$d(x_1, x_2)_p = 1 - \left(\sum_{j=1}^{n} (x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2) \Big/ \sqrt{\sum_{j=1}^{n} (x_{1j} - \bar{x}_1)^2 \sum_{j=1}^{n} (x_{2j} - \bar{x}_2)^2}\right) \tag{2.2}$$

The dissimilarity measure in Eq. 2.2 is based on the Pearson Correlation Coefficient (PCC), and measures the angle between two vectors, like the SAM measure. The PCC can also be seen as the cosine of the angle between two mean-centered spectra. Although the previous dissimilarities are of the most used measures in the comparison of chemical spectral data, the connectivity/ordering 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. 2.3) is a dissimilarity measure between two probability distributions:

$$d(x_1, x_2)_{ks} = \max_j \left(|\hat{x}_{1j} - \hat{x}_{2j}|\right) \tag{2.3}$$

$\hat{x}_{1j}$ 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 [87], the authors propose to compute the Manhattan measure on the first Gaussian derivatives (Eq. 2.4) of the curves (Shape measure), to take into account the shape information that can be obtained from the derivatives. The operator $*$ denotes convolution and σ stands for a smoothing parameter.

$$d(x_1, x_2) = \sum_{j=1}^{m} |x_{1j}^{\sigma} - x_{2j}^{\sigma}|, \quad x^{\sigma} = \frac{\mathrm{d}}{\mathrm{d}_j} G(j, \sigma) * x \tag{2.4}$$

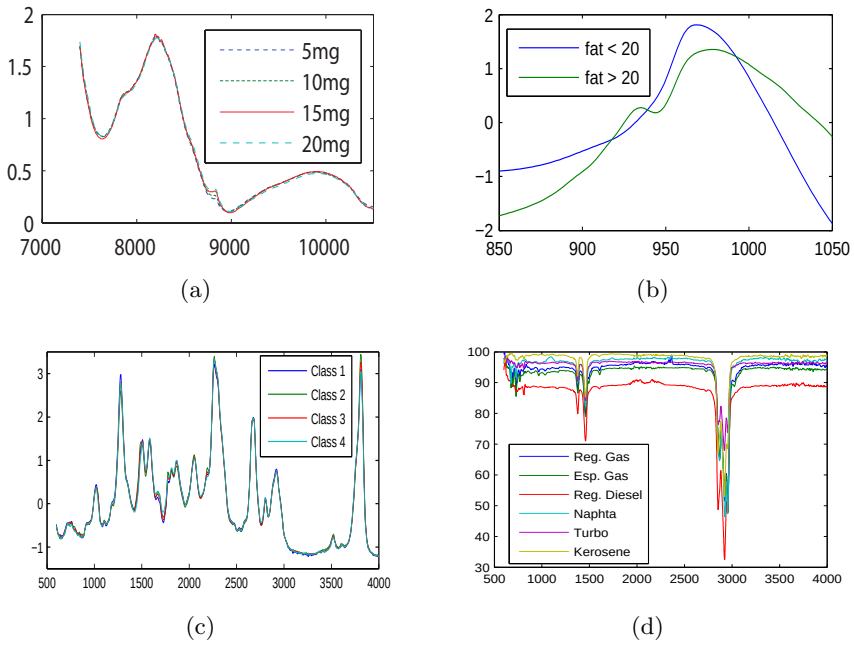
2.2.4 Experimental Section and Discussion

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 were chosen using leave-one-out cross validation. For the DR, all objects 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 data set 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 data sets are related in Table 2.1.

The first data set (Fig. 2.1(a)) is composed by near infrared (NIR) transmittance spectra of pharmaceutical tablets [150] of four different (classes) dosages of nominal content of active substance. In this data, the spectra of the examples of the different classes 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.



(a) (b) (c) (d)

Figure 2.1: Spectrum of one example from each of the classes of each data sets: a) Tablet, b) Tecator, c) Oil and d) Fuel.

The second, named Tecator [151] (Fig. 2.1(b)), consists of NIR absorbance spectra of meat examples. In this data, the examples 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 data set consists of oil examples of different origins, analyzed by Mid-Infrared (MIR) technique [146] and was transformed to have zero mean and unit variance. The variations in the spectra of classes are based in the difference in concentration of some substances and some shape changes also exist. The last data set consists of fuel examples of Fourier Transform Infrared (FT-IR) transmittance spectra; base line correction and smoothing were performed on

Table 2.1: Details about the # examples, features and examples per class of each data set. The last column is related to the # basis functions used for the FDA of each dataset.

Data Sets	#Examples	#Features	# Examples 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, High	48
Oils	80	571 (600-4000 cm^{-1})	Origin: A, BB, BC and D	100
Fuels	80	3528 (600-4000 cm^{-1})	Type: Regular Gasoline, Especial Gasoline, Regular Diesel, Naphtha, Turbo Diesel and Kerosene	300

Table 2.2: Averaged CV error with its standard deviation (%). Results are shown for the four classifiers on the feature, functional, and DR of each data set for the six dissimilarity measures presented. The numbers highlighted in bold and underlined, stand for the lowest error among all representations for each classifier. In the case of the dissimilarities, the one that performs best in general for each data set is also highlighted 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	10-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)*

the data. The objects of these classes differ in the substances by which they are composed (structure), and therefore they differ in shape.

As it can be seen in Table 2.2, in general for the four data sets, the SVM shows good results for all representations, outperforming the rest of the classifiers. These could be due to these data sets 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 data sets. This demonstrates the importance of a good and descriptive representation of the data.

In the case of the DR, results depend on whether a suitable dissimilarity measure is used to explain the discriminative characteristics of the curve, such that a better and more reliable classification of the data is obtained. It is worth to notice that, for both representations, the dimensionality of the data sets are reduced to half (or more) of the dimensionality of the feature representation. From the comparison of the dissimilarity measures, we can observe that very good results are achieved with the Shape dissimilarity, in which shape information is considered. This proves the fact outlined before, and suggests that this dissimilarity measure can be a suitable option for our purpose.

If we compare the results with the functional representation (FDA) and the DR of the data, it is shown that both approaches are competent when shape variations between objects of different classes are appreciable. But it can be observed that, the DR gives the best results for most data sets (with the Shape measure). It shows the capability of the Shape measure, which performs well not only in data sets 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 data sets 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.

In the case of Tecator data set, 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 data set, a similar result could be expected if the same procedure is carried.

However, in spite of the satisfactory performance of the DR for most cases, this is not the case for Oil data set. This suggests that, although the dissimilarity measures have shown their ability to discriminate between spectra that are very similar (see Tablet data set in Fig. 2.1(a)); 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 research most be done on this aspect.

2.2.5 Conclusions

We presented two alternative ways to improve the representation of chemical spectral data. The first makes use of the spectral continuous nature 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 data sets, 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 their shape changes into account.

2.3 Dissimilarity Representation on Functional Spectral Data for Classification

This section has been published as 'Dissimilarity Representation on Functional Spectral Data for Classification', by D. Porro-Muñoz, I. Talavera, R.P.W Duin, N. Hernández and M. Orozco-Alzate, in *Journal of Chemometrics*, **25**: 476–486 (2011). The sections of the paper were structured differently to make them fit in the thesis style.

Abstract

In chemometrics, spectral data are typically represented by vectors of features in spite of the fact that they are usually plotted as functions of e.g. wavelengths and concentrations. In the representation, this functional information is thereby not reflected. Consequently, some characteristics of the data that can be essential for discrimination between objects of different classes or any other analysis are ignored. Examples are the continuity between measured points and the shape of curves. In the Functional Data Analysis (FDA) approach, the functional characteristics of spectra are taken into account by approximating the data by real valued functions, e.g. splines. Another solution is the Dissimilarity Representation (DR), in which classifiers are trained in a space built by dissimilarities with training examples or prototypes of each class. Functional information may be incorporated in the definition of the dissimilarity measure. In this paper, we compare the feature-based representation of chemical spectral data, with three representations that take the functional characteristics into account: FDA, DR defined on raw data and DR defined on FDA descriptions. We analyze the classification results of these four representations for five data sets of different types, by using different classifiers. We demonstrate the importance of reflecting the functional characteristics of chemical spectral data in their representation, and show when the presented approaches can be more suitable.

2.3.1 Introduction

The increasing possibilities of chemometrics, raises a growing interest in advanced approaches to an automatic analysis of the collected data. If data sets are small, the accuracy of the results is of concern. One way to improve it is by considering advanced data representations. This paper focusses on finding better representations for the classification of spectral data.

The traditional way of representing spectra is by sampling. The higher the sampling resolution, the more accurate the spectrum is described. However, in case of the design of a classification system for spectra, this implies a representation in a high dimensional space. For small training sets of spectra, the resulting classifier will thereby be inaccurate due to the curse of dimensionality or overtraining. Dimension reduction by Principal Component Analysis (PCA) or Partial Least Squares (PLS) is needed, but may not solve the problem fully as they are still based on a statistical analysis of high-dimensional data.

Another way to tackle the problem of small training sets is to improve the original representation at the start. In particular, it may be advantageous to directly include the knowledge that spectra are one-dimensional signals and that neighboring points are connected, i.e. that their difference in amplitude is limited. The so-called Functional Data Analysis (FDA) [123, 108] uses this structural property of spectra by a functional approximation e.g. by B-spline basis functions. The dimensionality of the description of a spectrum is thereby reduced from the number of samples to the number of functional parameters.

A recently developed alternative in pattern recognition is the Dissimilarity Representation (DR) [89, 91, 93, 94]. This representation was mainly designed for discriminating between different classes of objects (classification), based on the important role that dissimilarities play for this purpose. The fact (or property) that dissimilarities should be smaller for similar objects (the same class) and larger for different objects, suggests that they could be used as more discriminative features due to their crucial performance in the class constitution. Therefore, in this approach objects are represented by distances as new features, determined by some "appropriate" dissimilarity measure, to a set of prototype objects usually named the representation set. Classifiers may be then built in the dissimilarity space, where each dimension corresponds to the dissimilarity to an object of the representation set (most representative objects for each class), and then applied to a new object represented the same way. Consequently, the geometry and the structure of a class are determined by the user defined dissimilarity measure, in which

application background information may be expressed. Like the FDA, the DR may make use of the structural data characteristics e.g. shape of the spectra. Some studies have already been reported on the DR for spectral data [88, 87, 85]. It is important to remark that, any traditional classifier that operates in feature spaces can also be used in the dissimilarity space.

FDA and DR are rather young techniques that have received a good acceptance in chemometrics and pattern recognition, respectively. Both aim to solve the problem of a statistical analysis for high-dimensional data generated by sampling spectra by introducing the possibility of integrating structural knowledge in the representation.Some classical multivariate techniques have been extended for FDA e.g. Functional Principal Component Analysis PCA [123], Canonical Correlation Analysis [70], Partial Least Squares [6, 28, 106] and Linear Discriminant Analysis [28]. If the used models are correct they are expected to perform better than the traditional techniques, as these have to learn (linear) relations from the data. More recently, a number of estimation methods for functional nonparametric classification and regression models have also been introduced. Namely, k-Nearest Neighbor classifier [30], kernel methods [1, 18, 42] such as Support Vector Machine [139, 54, 107, 55].

DR can be considered a generalization of the kernel approach studied in machine learning [94, 89] as it accepts almost any (dis)similarity measure between objects (spectra), including indefinite ones. Several applications have been studied [91], including spectra [88, 85]. Also in chemometrics studies appeared in which similarities between spectra are applied [134, 22, 63], but these mainly aim at studying correlation, cluster analysis or visualization. The use of representation has almost not been studied at all.

In this paper, we will compare the DR directly defined on distances between spectra with the FDA approach as well as with their combination, the DR based on the functional description of spectra. By this proposed combination, advantages of both approaches may be combined. By approximating the spectral data with the B-spline basis functions, the structural information in the spectra e.g. shape of peaks, can be incorporated in the dissimilarity measure. According to the basis of the DR, when adding this information about the structure of the objects, the distances between them should be more discriminative features. Therefore, one of the main issues in chemometrics, the small number of objects in high-dimensional spaces can be tackled, as with less but more discriminative data should be enough for the classification task. Hence, non-linearly separable problems in the feature space, can be converted to linear problems in the dissimilarity space.

As a baseline procedure for the comparison, it is used the traditional feature representation in which spectra are represented by their samples. The following classifiers are used: the k-Nearest Neighbor (k-NN) rule [44, 36], Regularized Linear Discriminant Analysis (RLDA) [44], Soft Independent Modelling of Class Analogy (SIMCA) [154] and the Support Vector Machine (SVM) [25, 133]. Section 2 summarizes and defines the foundations of FDA, DR and their combination. Data sets and experimental procedures are presented in Section 3 and results are discussed in Section 4. Finally, our conclusions are presented.

2.3.2 Dissimilarity Representation

The Dissimilarity Representation (DR) [91, 93, 94] was originally proposed as a more flexible representation of the objects than the traditional feature-based one. In this approach, which was mainly thought for classification purposes, new features are defined for the objects, such that they are represented by their dissimilarities to a set of representative objects of each class. The fact (or property) that dissimilarities should be smaller for similar objects (the same class) and larger for different objects, suggests that they could be used as more discriminative features due to their crucial role in the class constitution.

It aims at including more information about the characteristics and structure of the objects through the dissimilarity measure e.g. shape of spectra. There is not a general dissimilarity

measure for all problems. Hence, the first task in the DR is to select a suitable dissimilarity measure for the problem at hand. The fact that it has to be user-specified, is a way for the expert to integrate his knowledge and application [93].

Thus, in this approach, given a set of training objects $\mathbf{X} = x_1, x_2, ...x_n$ e.g. spectra, a representation set (a set of prototypes or representative objects for each class) $\mathbf{R}(r_1, ..., r_p)$ e.g. the reference spectrum for each substance that constitutes a class, and a dissimilarity measure; the distance between each object $x_i \in \mathbf{X}$ to each object $r_h \in \mathbf{R}$ will be defined as $d(x_i, r_h)$. The representation set $\mathbf{R}$ can be a subset of $\mathbf{X}$, $\mathbf{R} \subseteq \mathbf{X}$ or $\mathbf{X}$ itself, being then $\mathbf{D}(\mathbf{X}, \mathbf{X})$ a square dissimilarity matrix, or $\mathbf{R}$ and $\mathbf{X}$ can be completely different sets. There are some approaches to select prototypes of the representation set [92]. See reference for further details.

An object from the training set is then represented by a vector of dissimilarities $\mathbf{D}(x_i, \mathbf{R}) = [d(x_i, r_1), ..., d(x_i, r_p)]$, which relates him to the prototypes in the representation set. Therefore, in place of the traditional feature matrix $\mathbf{X} \in \mathbb{R}^{n \times q}$, where n runs over the objects and q over the features, the training set is now represented by the dissimilarity matrix $\mathbf{D}(\mathbf{X}, \mathbf{R})$ of size $n \times p$, which associates all objects from the training set to all objects from the representation set:

$$\mathbf{D} = \begin{pmatrix} d(x_1, r_1) & d(x_1, r_2) & \dots & d(x_1, r_p) \\ d(x_2, r_1) & d(x_2, r_2) & \dots & d(x_2, r_p) \\ \vdots & \vdots & \vdots & \vdots \\ d(x_n, r_1) & d(x_n, r_2) & \dots & d(x_n, r_p) \end{pmatrix}$$

We build from this matrix a dissimilarity space $\mathbb{D} \subseteq \mathbb{R}^p$. Objects are represented in this space by the row vectors of the dissimilarity matrix, such that each dimension corresponds to the dissimilarities with one of the representation objects. Using the DR, classifiers are trained in the space of the dissimilarities between objects, instead of the traditional feature space. Consequently, the relationship between all objects in the training and representation sets is used for the classification. If a suitable measure is chosen, the compactness property (objects from the same class should be similar and objects from different classes should be different) of the classes should be more pronounced. Therefore, it should be easier for the classifiers to discriminate between them, such that linear classifiers in dissimilarity space may correspond to non-linear classifier in feature space. In general, any arbitrary classifier operating on features can be used [89].

Given a test set $\mathbf{Y} = y_1, y_2, ..., y_g$, these objects are classified in the dissimilarity space, using their distances to the prototypes in $\mathbf{R}$, $\mathbf{D}(\mathbf{Y}, \mathbf{R})$, which is a $g \times p$ matrix.

Some dissimilarity measures have been already proposed for spectral data [87, 88]. In recent studies, the shape (ShD) measure demonstrated to have good capabilities for capturing the functional information. It consists of a sum of the absolute differences between the first Gaussian derivatives of the curves:

$$d(x_1, x_2) = \sum_{j=1}^{m} |x_{1j}^{\sigma} - x_{2j}^{\sigma}|, \quad x^{\sigma} = \frac{\mathrm{d}}{\mathrm{d}_j} G(j, \sigma) * x \tag{2.5}$$

The expression of x^{σ} corresponds to the computation of the first Gaussian (that is what G stands for) derivatives of spectra. Thus, a smoothing (blurring) is done by a convolution process $(*)$ with a gaussian filter and σ stands for the smoothing parameter [88]. Good performances have been obtained for chemical spectral data with this measure [88, 97].

2.3.3 Functional Data Analysis

In chemical spectral data as Near-Infrared, Ultra-Violet, each spectrum is a function of e.g. wavelengths, concentrations. However, they are usually observed and recorded discretely and so ana-

lyzed with multivariate data analysis techniques that consider the spectrum as high-dimensional vectors of different but high-correlated features. Therefore, when working with this type of representation, many practical problems can be encountered as the characteristics of the functional nature of the data are not taken into account.

FDA is based on retrieving the intrinsic characteristics of the underlying function from the discrete functional data. Thus, the observations (spectra) can be seen as continuous single entities, instead of sets of different features. Nevertheless, if the algorithms work on the functional spaces, they can also lead to theoretical and practical difficulties as these have infinite dimensions.

For dealing with the infinite dimensional problem, FDA methods have been constructed on two general principles: regularization and filtering. The filtering approach is based on using representation methods that allow working in finite dimension. This way of approximation is used here. The first step in FDA is to choose a proper family of basis functions matching best the function(s) to be approximated. Of a variety of bases that exist (Fourier series, polynomial, wavelet and splines), as spectra are generally smooth, it seems that B-splines [140] are more appropriate to approximate them. To make this basis of B-splines $\{\phi_k\}_{k=1}^{K}$ with K the number of basis functions, a number of knots (points) between the start and end wavelengths are defined. A B-spline is run from one knot to another; the different splines can overlap.

Hence, the spectral function $x_i = x_i(\lambda)$ for example i and wavelengths λ can be described by the linear combination of the basis functions:

$$x_i(\lambda) = \sum_{k=1}^{K} c_{ik}\phi_{ik}(\lambda) \ ,$$

where $c_i = [c_{i1}, c_{i2}, ..., c_{ik}]$, is the vector of B-spline weights (coefficients) correspondent to each spectrum (object) x_i. These coefficients are computed by minimizing the distance between the observed discrete spectrum x_i at wavelengths λ_j; $\forall j = 1, 2, ..., m$ and the fitted curve $x_i\lambda$:

$$c_i = arg_c \in \mathbb{R}^k min \sum_{j=1}^{m} (x_{ij} - \sum_{k=1}^{K} c_{ik}\phi_{ik}(\lambda_j))^2$$

Filtering can therefore be considered a preprocessing step in which functional data are consistently transformed into vector data [114]. As we are operating now in a finite-dimensional space, it is possible to work with the coefficients instead of working on the approximating functions. It has been demonstrated that working with these coefficients vectors c_i is strictly equivalent to working directly on the phi_i functions [113].

The function for each spectrum is thus explained by k coefficients, which are represented in a vector c_i, obtaining for the entire data set a matrix $\mathbf{C}(\mathbf{n} \times \mathbf{k})$:

$$\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1k} \\ c_{21} & c_{22} & \dots & c_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ c_{n1} & c_{n2} & \dots & c_{nk} \end{pmatrix}$$

Matrix $\mathbf{C}$, will be taken as the new representation of the data set, and the classifiers or any other data analysis method can use it as input, instead of the original data points. A dimensionality reduction is achieved in this process, which is an important task when working with spectral data. Besides the functional representation by B-splines is already a way of smoothing the curve, other functional processing techniques such as derivation can be done on it. This processing could be beneficial when the analysis of the shape of the function (curvature) is essential for the solution of the problem at hand.

2.3.4 Dissimilarity representation and Functional Data Analysis

Based on the advantages that both of the previous approaches show for the representation of spectral data, we propose to compute the dissimilarity representation from the functional representation of the data (DR-FDA), instead of computing it from the feature-based one. By using the functional approximation, we are highlighting the structural (in terms of how the composition of the substances is reflected in the shape of the spectra) information of the spectra. Hence, we have a more faithful representation of the real spectrum than by using the feature-based one. Moreover, by using B-spline basis, the interpretability of the data is still maintained. Because the new features (coefficients) depend only on some spectral regions, a range of these original features can be associated to each new feature. Therefore, the spectral regions responsible for the discrimination can still be depicted from the functional representation. A method to achieve this association was recently introduced [114].

Moreover, if spectra are represented by dissimilarities, the use of a suitable dissimilarity measure allows emphasizing important details of the particular problem, which cannot be easily treated by the simple feature representation e.g. shape and continuity of the measured points of the spectrum. Any knowledge on the problem e.g. discriminative spectral regions, or on the spectral background can also be included in the measure. Besides, by using the dissimilarities as features, we are considering the relationship between all objects (structure of the classes) as information. This is very important for the discrimination between the classes, and even more when the number of objects is very small (a typical problem in chemometrics), as with less but more discriminative data should be enough for the classification task. Thus, non-linearly separable problems in the feature space, can be converted to linear problems in the dissimilarity space. Furthermore, by using the DR for chemical spectral data, the problem of high-dimensionality spaces is also eradicated. The highest dimension that could have the data now, is the number of objects in the training set, which are usually much less than features (spectral bands) in these types of data sets. Therefore, it can be less computationally expensive.

In consequence, it is to be expected that, if the dissimilarities between a set of spectra $\mathbf{D}(\mathbf{X}, \mathbf{R})$ are derived from their functional representation i.e. the vector of coefficients c obtained from the approximation of each spectrum by B-spline basis functions; the classification results may improve. A better description of the objects than feature-based will be used, where patterns in the structure of the spectra can be more exploited. Example of a simple dissimilarity measure that can be computed on the functional data is the Manhattan (L1-norm) distance:

$$d(x_1, x_2) = \sum_{k=1}^{K} |c_{1k} - c_{2k}| \tag{2.6}$$

The Manhattan distance is one of the most commonly used in many research areas, and particulary for the comparison of spectral data [87, 63, 85, 97]. This measure views the spectrum as a high-dimensional feature vector, making just a band-to-band comparison, therefore neglecting the connectivity/ordering between the measured points of the spectra. However, the functional information from the spectra can be taken into account if the distance is computed from their functional representation by B-splines approximation.

In the following section, the feature, functional and DR approaches will be compared to the proposed one on different chemical spectral data sets. Four classifiers are used to show the efficacy of this approach compared to the others.

2.3.5 Materials and Methods

A comparative study is carried out between the two representations presented above (DR and FDA) and the proposed approach (DR-FDA), using the feature representation as a baseline

comparison. The performance of four classifiers will be evaluated on these four representations of five chemical spectral data sets.

Data sets

The first data set, named Tecator (Figure 2.2 (a)), originates from the food industry [151]. It consists of 215 near infrared absorbance spectra of meat examples, recorded on a Tecator Infratec Food and Feed Analyzer. Each observation consists in a 100 channel absorbance spectrum in the 850-1050 nm wavelength range. It is associated to a content description of meat example, obtained by analytic chemistry.

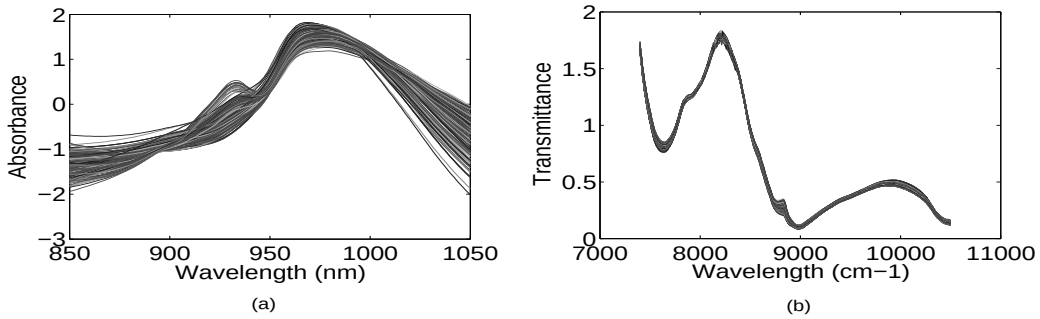


Figure 2.2: Data sets (a) Tecator and (b) Tablets

The classification problem consists in separating 77 meat examples with a high fat content (more than 20%), from 138 examples with a low fat content (less than 20%). Original spectra are preprocessed, each spectrum is reduced to zero mean and unit variance.

The second data set (Figure 2.2 (b)) is composed of near infrared (NIR) transmittance spectra of pharmaceutical tablets [39]. It consists of 310 spectra and 404 features in a range of wavelengths from 7400 to 10500 cm^{-1}. Four different (classes) dosages of nominal content of active substance are analyzed: class A (5 mg), B (lOmg), C (15mg), and D (20 mg) per tablet. There are 70 objects in class A and 80 in each of the other classes. As reported in the reference [39], a Multiplicative Scatter Correction (MSC) [46] was used as preprocessing method. The MSC transforms the spectrum x to z, such that $z(j) = (x(j) - a)/b$, with a the intercept and b the slope of a "least-squares" regression of the values $x(j)$, on the corresponding values $r(j)$ for a reference spectrum. Usually this reference spectrum is the mean of all the available spectra [41].

The third data set is a real-world data set, which was obtained from a cooperation with the Oil Industry in Cuba. It consists of 80 fuel examples of Fourier Transform Infrared (FT-IR) transmittance spectra (Figure 2.3 (a)) in a wavelength range of 600-4000 cm^{-1}. A base line correction and smoothing were performed on the data. The classification problem consists in determining the fuel type of the examples: regular gasoline (16 objects), especial gasoline (15 objects), regular diesel (16 objects), naphtha (16 objects), turbo diesel (9 objects) and kerosene (8 objects).

The fourth data set is another fuel real-world data set of 101 examples measured at 127 wavelengths in a range of 275-220 nm, but this time measures have been taken by a Ultra-Violet Visible (UV) spectrophotometer (see Figure 2.3 (b)). The classification problem consists also in determining the fuel type of the examples: regular gasoline (23 objects), especial gasoline (21 objects), regular diesel (22 objects), naphtha (18 objects) and turbo diesel (17 objects).

The last data set consists of 101 NIR spectra of four different common pharmaceutical excipients (classes), with $27, 14, 17$ and 13 objects in each class respectively, measured at 700 wavelengths (see Figure 2.4). The goal is to develop a classification model to identify to which

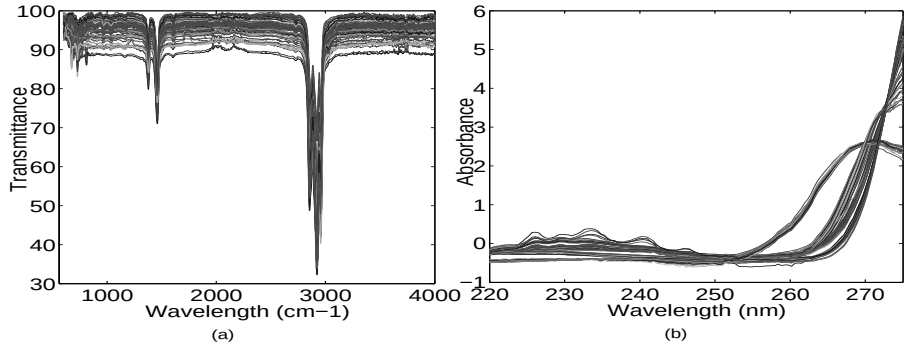


Figure 2.3: Fuel data sets from (a)FT-IR and (b)UV-VIS

of the pharmaceutical excipients belongs a new object. This is an example data set from the chemometrics software Pirouette [148]. For both of the previous data sets, original spectra are preprocessed such that each spectrum is reduced to zero mean and unit variance.

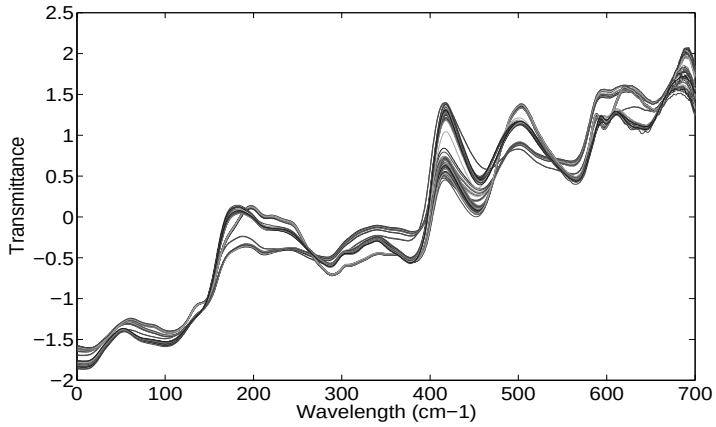


Figure 2.4: Pharmaceutical excipients data set

Software and optimization

The experiments were all performed in Matlab. For the case of FDA the FDAFuns [144] toolbox was used, and PRTools toolbox [149] for the DR and classification of the data. The experiments have been designed in the following way:

1. A comparison is made between the performance of classifiers on the four representations of each dataset: feature-based (spectral), functional (FDA), dissimilarity representation (DR) and the proposed combination (DR-FDA).

2. For the classification of the data we used four classifiers: k-Nearest Neighbor (k-NN), Support Vector Machine (SVM), Soft Independent Modeling of Class Analogy (SIMCA) and Regularized Linear Discriminant Classifier (RLDC). To solve the multiclass problems of some of the data sets, the one-versus-all classification scheme is applied. For the k-NN classifier, a leave-one-out optimization for k is computed. An optimal number of k= 1 was obtained for all data sets, thus a 1-NN classifier is applied. For the SVM classifier, the Gaussian and Linear kernel were applied in the five data sets, to show their performances

on their different representations. The optimal regularization parameter C and Gaussian kernel width parameter σ, were tuned in a grid search based on a k-fold cross-validation procedure. In the case of the Linear kernel, the regularization parameter was optimized in a k-fold cross-validation. The Linear Discriminant Classifier (LDC) assumes that the classes are described by multi-normal distributions with the same covariance matrices. Since for $n \times n$ dissimilarity representations the estimated covariance matrix S is singular, its inverse cannot be determined. Therefore, its regularized version is used instead (RLDC). Regularization takes care that the inverse operation is possible by emphasizing the diagonal values (variances) of the matrix S with reference to the off-diagonal elements (covariances) [89, 94]. To find the regularization parameters of RLDC, an automatic regularization (optimization over training set by cross-validation) process was done.

3. For the functional representation, each spectra was represented by a l-th order B-spline approximation with K basis functions. The optimal values for the number of B-spline bases and the order of the splines were chosen by using a leave-one-out cross-validation, using the error in the approximation of the curve as evaluation criteria [113]. In the comparison with all representations, the results for the FDA are reported for the performance of classifiers on the functional representation, and when the second derivative is applied on it.

4. For the DR, the shape distance was applied (Equation 2.5). For DR-FDA, we used the Manhattan distance on the functional representation as defined in Equation 2.6. The results shown in this case, are the ones computed on the functional representation version (see above) for which the classifiers performed better. The entire set of objects was used as representation set for all data.

5. For all data sets, a k-folds cross-validation procedure was repeated 10 times, such that all objects are used for training and test at some moment. Consequently, the information of all objects is taken into account for the modeling of the problem. Classifiers performances are evaluated in terms of the Average Classification Error (ACE), and the standard deviation from the different repetitions is taken into account.

2.3.6 Results and discussion

Tecator data set

The classification results (averaged classification error) for the different representations of Tecator data set are shown in Table 2.3. The data set was split in different training and test sets in a 10-fold cross-validation repeated 10 times. For the functional approach, the leave-one-out error calculation leads to the selection of an optimal basis of 48 B-splines of order 5.

As stated above, in this data, the objects of the two classes differ in their fat content, which is reflected in changes in the shape of the spectra. We can observe remarked differences in curvature of the spectra between the examples of the two classes (fat<20 and fat>20). High fat content spectra (fat>20) have sometimes two local maxima instead of one (fat<20) (Figure 2.5). As it is reported in the literature [139], we computed the second derivative of the functional data, to highlight these differences.

In the case of the derivative based distance, shape (ShD), the second derivative was also applied. The smoothing parameter σ was optimized in a 10-fold cross-validation procedure repeated 10 times. The best results were achieved with $\sigma = 2$.

In Table 2.3, it is shown that classifiers perform better on the functional data than on their original (feature) representation. The good performance of most classifiers on the functional space is due to the fact that, from the functional point of view, there is a great quantity of

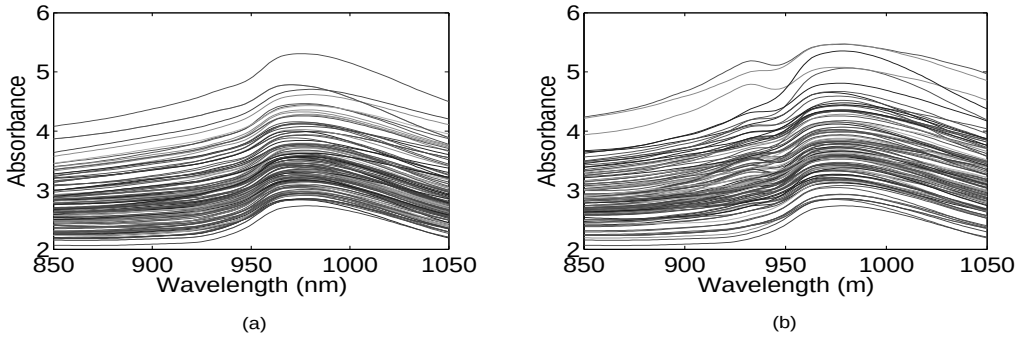


Figure 2.5: Data sets (a) Tecator (fat<20) and (b) Tecator (fat>20)

Table 2.3: Averaged cross-validation error in % (with standard deviation) for Tecator data set for different classifiers

	Representations				
Classifiers	Feature	DR	FDA	FDA (+ 2nd der)	DR-FDA
1-NN	4.11(0.2)	3.72(0.2)	2.97(0.2)	2.15(0.1)	1.41(0.1)
RLDA	6.82(0.5)	1.72(0.01)	3.02(0.3)	2.74(0.3)	0.98(0.1)
SVM (Gaussian kernel)	2.56(0.2)	2.65(0.2)	1.21(0.3)	0.93(0.2)	0.6(0.04)
SVM (Linear kernel)	3.51(0.2)	2.88(0.1)	1.8(0.2)	1.21(0.05)	0.47(0)
SIMCA	5.77(0.01)	3.3(0.2)	2.6(0.07)	2.31(0.3)	1.21(0.04)

information to obtain when shape changes are present in the curve. Hence, the FDA by B-splines is capable of using the information embedded in curvature of the spectrum. The use of the second derivatives emphasizes the peaks in the curve, therefore it makes easier to see the differences [139]. Nevertheless, it seems that it is not enough to discriminate between both classes.

For the DR, the results with the shape dissimilarity measure are very good compared to the results on the original feature-based data. This dissimilarity measure takes into account the shape information (functional) that can be obtained from the derivatives. Compared to FDA, the results are usually worst (taking the standard deviation into account). This can be due to in cases like this, the use of the B-splines and second derivatives afterwards are more capable of extracting the functional information than the shape dissimilarity measure.

In this data set, we can see that linear classifiers perform a bit better on the functional space, but not good enough. Results with non-linear classifiers are still better. This might be because classes are non-linearly separable. In Figure 2.6, the scores of all objects from a PCA of 2 principal components from each representation are plotted. In all cases, more than 95% is retained in this 2 principal components. These are not demonstrative plots, but they show in someway how the structure of the classes can be according to each representation. It can also be seen that all classifiers, even the linear classifiers, perform better when computing the dissimilarities on the functional representation of the data (see Equation 2.6), corroborating our hypothesis. These results are also better than the obtained in the literature [139]. The functional information that is not captured by the measure itself is obtained by the FDA, and thus included in the DR. The relationship between all objects is also considered when analyzing them all-against-all in the dissimilarity space, which is very important for the discrimination between the classes. These results could be even improved by some other expert knowledge introduced into the dissimilarity measure.

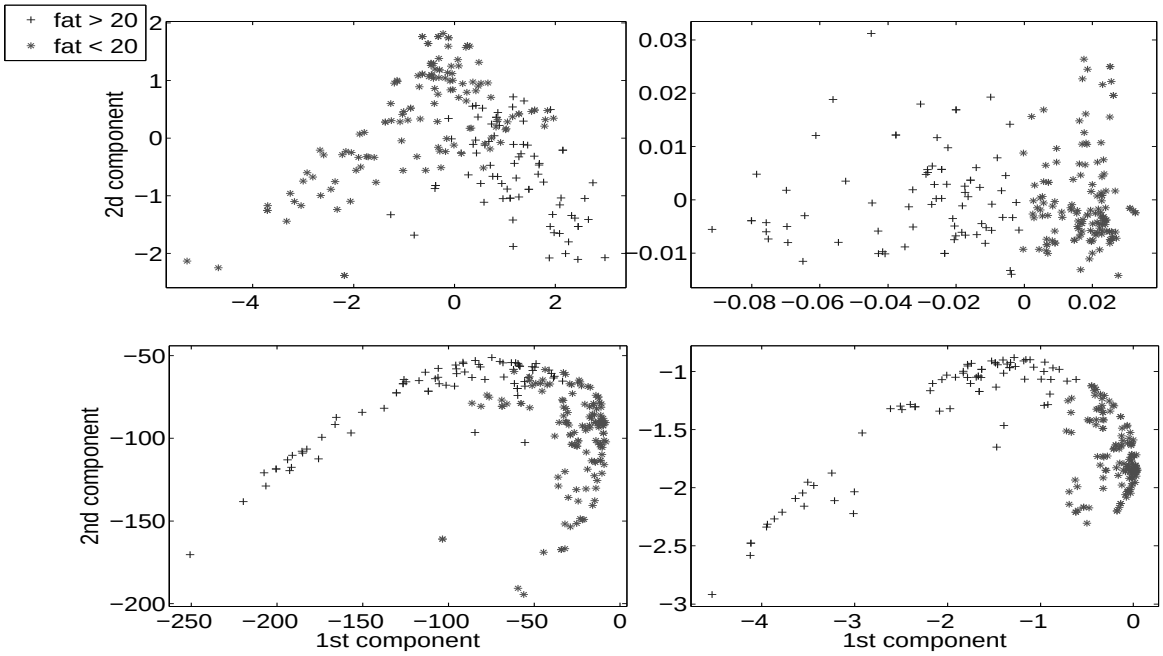


Figure 2.6: PCA of 2 components for the four representations of Tecator data set: Feature (top-left), FDA (top-right), DR (bottom-left) and DR-FDA (bottom-right)

Fuel Data set by FT-IR and UV-VIS

With these two data sets, we are tackling the same problem of discriminating between types of fuel, but using two different instrumental techniques. Besides, they are not based on the same objects, and in the FT-IR data set, one class more is analyzed. In this case, the objects of the classes differ in the substances by which they are composed (see Figure 2.7), and therefore they differ in shape (although sometimes it is difficult to determine for all of them) in some parts of their spectrum.

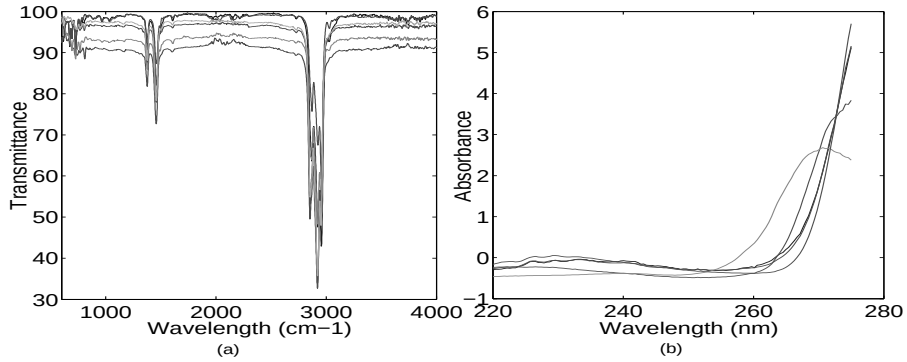


Figure 2.7: One example of each of the classes of Fuel (a)FT-IR and (b)UV-VIS data sets

The data sets were split in different training and test sets in a 8-fold and 10-fold cross-validation for the FT-IR and UV-VIS respectively; these splits were repeated 10 times. For the functional approach, the leave-one-out error calculation leads to the selection of an optimal basis of 850 B-splines of order 4 for the FT-IR data set and 30 B-splines of order 4 for the UV-VIS. We computed also for both of them the second derivative of the functional data, to highlight the curvature differences. In the case of the derivative based distance, shape, the

second derivative was applied. The smoothing parameter σ was optimized in a 5-times 10-fold cross-validation procedure and the best results were achieved with $\sigma = 3$ also for both cases.

Table 2.4: Averaged cross-validation error in % (with standard deviation) for Fuel (FT-IR) data set for different classifiers

	Representations				
Classifiers	Feature	DR	FDA	FDA (+ 2nd der)	DR-FDA
1-NN	32.4(0.8)	14.3(0.7)	18.1(0.8)	10.4(0.7)	7.5(0.8)
RLDA	16.1(0.5)	13.6(0.2)	14(0.5)	11.8(0.6)	9.5(0.4)
SVM (Gaussian kernel)	11(0.5)	8.8(0.3)	12.8(0.8)	9.6(0.5)	6.3(0.3)
SVM (Linear kernel)	17.8(0.8)	7.4(0.4)	14.9(0.5)	12(0.8)	4.5(0.2)
SIMCA	22.5(1.1)	15.6(0.6)	14.9(1.1)	10.7(1)	8.7(1)

Table 2.5: Averaged cross-validation error in % (with standard deviation) for Fuel (UV-VIS) data set for different classifiers

	Representations				
Classifiers	Feature	DR	FDA	FDA (+ 2nd der)	DR-FDA
1-NN	13.1(0.3)	19.8(0.4)	14.6(0.4)	11.4(0.2)	10.7(0.3)
RLDA	21.4(1)	13.3(0.8)	14.9(0.7)	10.8(0.4)	7(0.6)
SVM (Gaussian kernel)	13.1(0.2)	11.4(0.3)	16.8(0.1)	9.4(0.2)	8.1 (0.2)
SVM (Linear kernel)	15(0.5)	12.2(0.5)	14.1(0.6)	12.5(0.1)	8.7(0.2)
SIMCA	21.2(0.6)	19.5(0.4)	20.6(0.4)	17.5(0.7)	14.3(0.3)

From Tables 2.4 and 2.5, it can be observed that the pattern in the behavior of the results is very similar to the obtained for Tecator data set. For both of them, we are in presence of the same situation. The difference between the examples of the classes are mainly in the curvature of the spectra.

It can be noticed that in both of these data sets the results with the functional data and the DR outperform those obtained for the feature-based representation of the data. It is also remarkable, how the classifiers accuracy improves even more when the DR is computed on the second derivative of the functional representation of the data. However, the results are a bit worst for the UV-VIS spectra due to the characteristics of the instrumental techniques. It seems that the information obtained in the FT-IR spectra is more discriminative specially for Regular and Especial gasoline. It is worth to notice the advantage of the DR in this case, where we have six (FT-IR) and five (UV-VIS) classes and a few objects for each of them. If the dissimilarities have capture more structure from the data, it should be sufficient to discriminate with few data. The FT-IR data set is also the case, where in the functional representation a dimensionality reduction was achieved, but still it was high for the number of objects available. From this approximation more reduction could not be possible, otherwise important information could be lost in the smoothing process; besides it was the result from the optimization process. Thus, it is highlighted the advantage of the DR in these cases, as classifiers are built in a more balanced space and do not have to deal with the high-dimensionality problems.

Pharmaceutical excipient Data set

In Figure 2.8, an example of each of the four pharmaceutical excipient classes are shown. This is another example were the objects from different classes have differences in shape in some parts of their spectra. Thus, it would be important to take into account this information in their representation, such that it would be possible to discriminate better between them.

The data set was split in different training and test sets in a 10-fold cross-validation procedure. For the functional approach, the leave-one-out error calculation leads to the selection of an optimal basis of 150 B-splines of order 4. We computed also for both of them the second derivative of the functional data, to highlight the curvature. In the case of the derivative

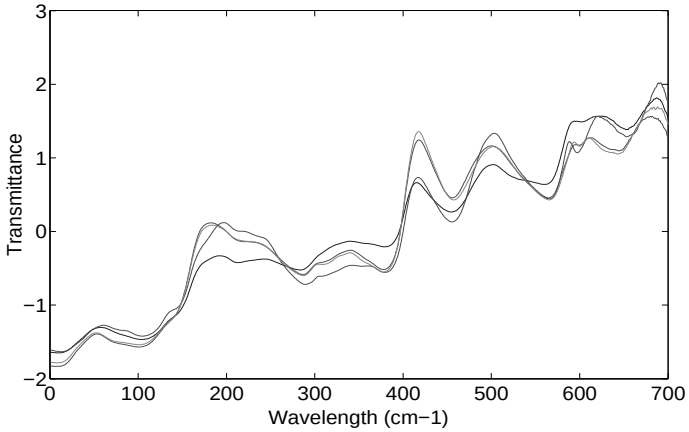


Figure 2.8: One example of each of the classes of Pharmaceutical excipient data set

based distance, shape, the second derivative was applied. The optimal value for the smoothing parameter was $\sigma = 1$.

As it is shown in Table 2.6 for all classifiers, the combination of DR with FDA outperforms the other representations in general. In this case, although there is an improvement, it is not as remarkable as in the previous data sets. The classification problem seems to be not so difficult as it can be observed in the results. Thus, it depends on the application, how much the users are willing to give up in some efficiency for a bit more of efficacy.

Table 2.6: Averaged cross-validation error in % (with standard deviation) for Pharmaceutical excipient data set for different classifiers

	Representations				
Classifiers	Feature	DR	FDA	FDA (+ 2nd der)	DR-FDA
1-NN	3.9(0.05)	2.7(0.3)	3.2(0.3)	2.1(0.5)	1.4(0.1)
RLDA	2.9(0.6)	1.7(0.4)	2.5(0.3)	1.4(0.1)	1(0.2)
SVM (Gaussian kernel)	2.8(0.05)	1.8(0.5)	2.1(0.3)	1.5(0.1)	1.21(0.2)
SVM (Linear kernel)	3.1 (0.1)	2.1(0.8)	2.8(0.2)	1.9(0.2)	1(0.3)
SIMCA	4.2(0.6)	3.7(0.1)	4.7(0.04)	3.8(0.4)	2.7(0.7)

From the studies of all these data sets, it can be seen that when the spectra of different classes are characterized by having differences in their curvature, this is discriminative information that should be taken into account. These differences between the classes are not always so visible or clear, such that the pattern for each of them could be extracted easily. Thus, the use of mathematical operators like the second derivative can emphasize the curvature of the spectrum; therefore the shape difference is more highlighted to be further used by the classifiers.

In the results shown for the previous data sets, the SVM shows the better results on all the representations. This could be due to these data sets are mostly non-linear. This classifier has shown a high flexibility to confront complex data sets. Such is the case of spectral chemical data sets where the number of objects is small and have a high dimensionality, and the classes are completely unbalanced with respect to the number of objects that belong to each of them. Nevertheless, it should be noticed that in most cases, when classifiers are built on the DR from FDA, the linear classifiers i.e. RLDA and SVM (linear kernel) have better or the same performance than the non-linear ones. Showing once more the feasibility or one of the advantages that can be taken from this approach. Non-linearly separable problems in the feature space, can be converted to linear problems in the dissimilarity space. All the previously discussed, corroborates the idea that the proposed combination can be optimal in cases like this, where the DR is generated from the functional data extracted by the approximation with B-splines.

Tablet Data set

In this data, the spectra of the examples of the different classes are very similar, they barely variate in the intensity of one peak at 8830 cm^{-1} (corresponding to 1132 nm) (Figure 2.9). This peak corresponds to the only visually characteristic band of the active substance, which is identified as the second overtone of the aromatic C-H stretch. It is partially overlapping with the peak at 8200 cm^{-1}(1220 nm), originating from the primary excipient, microcrysta11ine-cellulose [39]. It seems that we are in presence of another non-linearly separable classes problem.

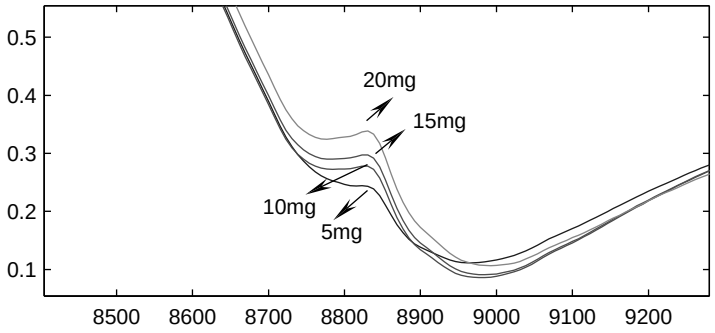


Figure 2.9: Nominal content of active substance(mg) for the classes

The data set was split in different training and test sets in a 10-fold cross-validation procedure. For the functional approach, the leave-one-out error calculation leads to the selection of an optimal basis of 150 B-splines of order 4. In the case of the derivative based distance, shape, the first derivative was applied. The smoothing parameter σ was optimized in a 5-times 10-fold cross-validation and the best results were achieved with $\sigma = 2$.

In Table 2.7, the results are shown. In this data set, it is to be expected that the results with the functional representation from the B-splines approximation, do not make much of a difference. This could be explained by the fact that only a few information can be extracted from this data, from the functional point of view. The second derivative does not give any information of curvature either.

Table 2.7: Averaged cross-validation error in % (with standard deviation) for Tablet data set for different classifiers

	Representations				
Classifiers	Feature	DR	FDA	FDA (+ 2nd der)	DR-FDA
1-NN	16.7(0.1)	8.7(0.2)	15.1(0,1)	22.9(0.5)	14.9(0.5)
RLDA	25.6(0.6)	7.7(0.4)	20.8(0.3)	24.8(0.4)	10.87(0.2)
SVM (Gaussian kernel)	14.1(0.3)	6.7(0.3)	10.7(0.2)	14.5(0.3)	9.6(0.3)
SVM (Linear kernel)	20.1(0.5)	10.5(0.2)	15.6(0.3)	17.8(0.4)	11.2(0.2)
SIMCA	23.8(0)	14.8(0.4)	24.2(0.1)	28.6(0.05)	16.2(0.1)

Indeed, it can be observed in the table above that the results with the functional representation are very similar to the obtained on the original spectra (considering the standard deviation). With the dissimilarity representation, good results are obtained in general, even more than for FDA based on B-splines. This shows that this measure is also capable of detecting the intensity changes between the different curves, even when they are so slight as in this case. In Figure 2.10 we try to show again with 2-components PCA (more than 95% of variance is retained for all representations) that, the classes could be slightly more separable in this space that in the others.

However, if we analyze the performance of classifiers on the dissimilarity space obtained from the functional representation, the behavior is reasonable. Most classifiers perform better on the dissimilarity space generated from the original feature-based data than that of the functional

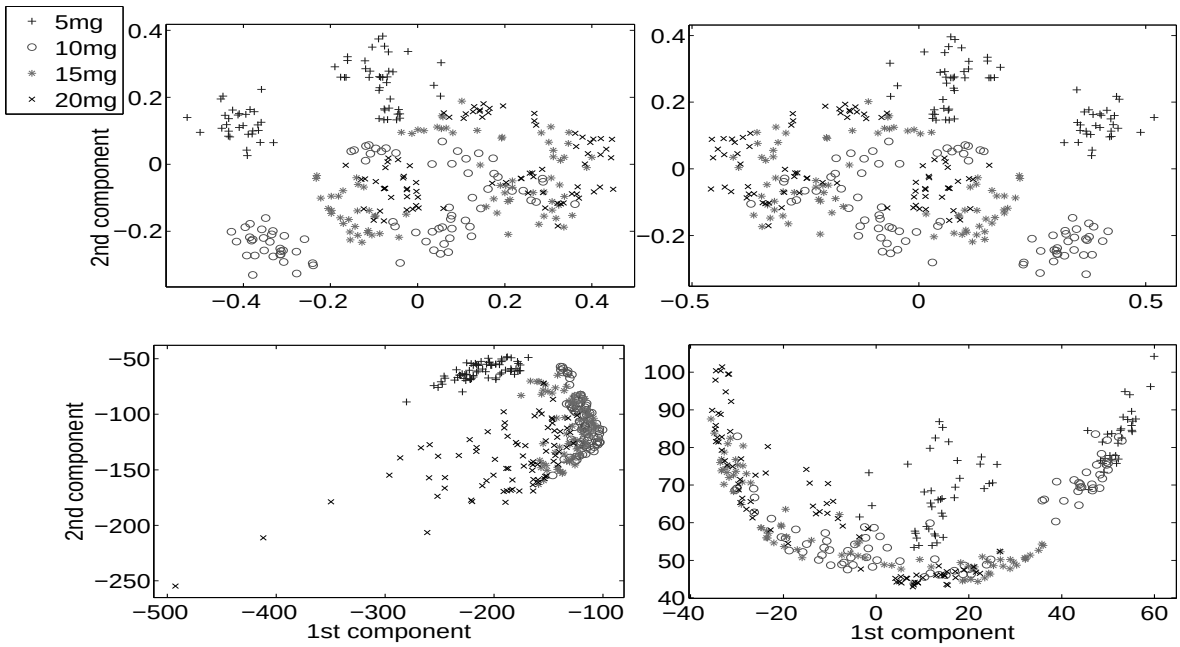


Figure 2.10: PCA of 2 components for the four representations of Tablet data set: Feature (top-left), FDA (top-right), ShD (bottom-left) and DR-FDA (bottom-right)

representation. It is understandable that in this case, the scarce functional information extracted with by the B-splines does not benefit the DR. This lack of information in the functional representation, can be caused by the loss of some information when using only the coefficients resulting from the smoothing process with projection of the function in the B-spline basis.

2.3.7 Conclusions

We presented three alternative ways to improve the representation of chemical spectral data. The first makes use of the physical knowledge of the spectral background of the data, by modeling their relations in a dissimilarity representation (DR). The second makes use of the spectral continuous nature by approximating the spectra by spline functions (FDA). In the third, we propose to compute the dissimilarity representation on the functional representation of the data (DR-FDA). Therefore, the functional information of spectra is taken into account and we can make use of the advantages of both approaches. Comparisons were made by classifying five chemical spectral data sets, expressed by their feature and the three other representations. We can conclude that: in chemical spectral data sets where changes in the shape of the spectra of different classes are present e.g. Tecator and Fuel data sets, both FDA and the DR outperform the results on the feature space. In the comparison of this type of data by their dissimilarities, the better results are obtained with measures that take the functional information into account. Such is the case of the shape dissimilarity measure and the proposed combination of the DR on the functional data. Nevertheless, the later has shown to be the best option in this case. In data sets where the differences between objects are referred only to intensity changes e.g. Tablet, the shape dissimilarity is capable of improving the results obtained on the feature space. However, FDA (with B-splines) is not able to extract the functional information from this type of data. Therefore, the computation of the DR on the functional data does not improve, since it is influenced by the errors of the functional approach.

2.3.8 Acknowledgements

We acknowledge financial support from the FET programme within the EU FP7, under the SIMBAD project (contract 213250). We would also like to thank to the project Cálculo científico para catacterización e identificación en problemas dinámicos (code Hermes 10722) granted by Universidad Nacional de Colombia.

Chapter 3

Dissimilarity Representation in Multi-way Spectral Data

3.1 Overview

In many application areas, the straightforward representation of objects is in a multi-dimensional array i.e. $features1 \times features2 \times ... \times featuresn$. Chemistry for example is very rich in this type of data sets, due to the development of many advanced instrumental analysis techniques such as gas chromatography- mass spectrometry, high performance liquid chromatography with photo-diodo array detection, excitation-emission fluorescence. Such structures, containing information about the relation between the different types of measurements can be useful for a better comprehension of the problem at hand. The field dedicated to study and develop proper tools to analyze this kind of data sets is known as multi-way data analysis [126, 64].

A number of methods for multi-way data analysis has been proposed [4, 98, 64]. However, most of these methods are for exploratory and regression purposes. Classification has been much less studied. Some classification approaches have been recently considered with the aim of making use of the multi-dimensional structure for the learning process e.g. Multi-way Soft Independent Modeling of Class Analogy (NSIMCA) [37]. Nonetheless, they are general classification approaches that do not take into account characteristics about the type of data that could help for a better discrimination. Such is the case of multi-way spectral data that, as simple (1D) spectral data are just analyzed as a set of individual measured values. Hence, after the results of the investigation in Chapter 2, we decided to introduce more advanced representations for multi-way spectral data.

This chapter contains three main contributions related to the classification of multi-way data based on dissimilarity representation. First, we introduce the dissimilarity representation approach as a new tool for the classification of multi-way data. The key issue is then to propose a suitable dissimilarity measure for this type of data. In section 3.2, we propose a 2D measure (2Dshape) for three-way spectral data. It is based on the combination of 1D measures (taking into account the information of both measurement directions). The functional (context) information on one or the two directions of the array is taken into account by comparing the first Gaussian derivatives. This measure has been mainly considered for data sets with features of different nature in the different directions e.g. data with a continuous and a non-continuous direction, like in social network analysis: $users \times keywords \times time\ samples$.

An unattractive characteristic of spectral data sets in general (simple and multi-way) is that they often tend to result in high-dimensional representations, causing classifiers to suffer from the curse of dimensionality. A small set of objects will cause traditional statistical methods to fail in finding a good class model [109]. As the dimensionality of the dissimilarity representation approach depends on (at most) the size of the training test, this problem could be tackled by the DR approach. In section 3.3, we study the importance of using a suitable dissimilarity measure for the problem at hand. We show that by taking into account discriminative information about spectra when designing the dissimilarity measure, not only leads to better classification results, but that a certain number of objects is enough to achieve it.

Another issue to be tackled in multi-way spectral data is related to the large amount of noise and/or redundant information it can contain. For the computation of the DR, the redundant information should not really affect the classification accuracy. However, due to the high dimensionality of these data sets, there is the problem of the computational cost for obtaining the dissimilarity representation of new elements. It can be considerably reduced if this redundant information is removed. The noisy information, however, is more worrying. If the amount of noise is considerable, it can really influence any interpretation of the data, i.e. the dissimilarity representation can no longer represent the desired separation between classes as to obtain a good classification.

Both problems (noisy and redundant features) are usually overcome by applying feature selection techniques. There are several approaches for feature selection in traditional multivariate

data analysis. However, this is not the case for multi-dimensional data arrays.

In the second part of this chapter, we investigate how the DR of three-way data could benefit from performing a data reduction prior to its computation. With this purpose, we introduce in section 3.4 a procedure based on selecting the most discriminative parts (peaks) in both directions of the three-way spectral data. The mutual information [122] between each peak and classes is used as selection criterion [95]. It is shown that the introduced procedure can be very useful for high-dimensional data sets.

The measure introduced in section 3.2 does not take into account how surfaces of 2D continuous data change simultaneously in both directions. This information on the relationship between measurements of the different directions can be very important for discrimination. In section 3.5, we introduce the Continuous Multi-way Shape (CMS) measure, which is based on the differences between the gradients of objects, thus considering simultaneous shape changes in the surfaces. This measure can also be easily extended to multi-way objects (not only 2D). Thereby, we are not limited to use the benefits of the DR approach for three-way data only, but for multi-way data in general.

An also remarkable aspect of this measure is that it is not restricted to measure shape in the principal directions of the multi-way array only. Although CMS is based on the idea of the gradient of an image, which is mathematically defined by a vector of 2 components i.e. horizontal and vertical, this measure allows analyzing other directions e.g. diagonals. Hence, more accurate approximations of the information on the shape of objects and the dependencies of the different directions, can be used for the comparison of objects. In this section, we also introduce a gradient kernel operator. It is based on computing a partial derivative as the derivative of a higher-degree polynomial fitted to the analyzed points, instead of a simple line. This way, data should be better fitted, thus leading to a better approximation of derivatives.

3.2 Classification of three-way data by the dissimilarity representation

This section has been published as 'Classification of three-way data by the dissimilarity representation', by D. Porro-Muñoz, R.P.W Duin, I. Talavera and M. Orozco-Alzate, in *Signal Processing*, **91 (11)**: 2520–2529 (2011). The structure of the paper was modified to fit in the thesis style. Notation was also changed in order to have a unified notation in the whole thesis.

Abstract

Representation of objects by multi-dimensional data arrays has become very common for many research areas e.g. image analysis, signal processing and chemometrics. In most cases, it is the straightforward representation obtained from sophisticated measurement equipments e.g. radar signal processing. Although the use of this complex data structure could be advantageous for a better discrimination between different classes of objects, it is usually ignored. Classification tools that take this structure into account have hardly been developed yet. Meanwhile, the dissimilarity representation has demonstrated advantages in the solution of such classification problems. Dissimilarities also allow the representation of multi-dimensional objects in a way that the data structure can be used. This paper introduces the use of dissimilarities as a tool for classifying objects originally represented by two-dimensional (2D) arrays. A 2D measure to compute the dissimilarity representation from spectral data with this kind of structure is proposed. It is compared to existent 2D measures, in terms of the information that is taken into account and computational complexity.

3.2.1 Introduction

The standard way of representing objects for classification is in a two-way structure (matrix), where a number of objects (rows) are simply characterized by feature vectors (1D representation). Nevertheless, in many research areas such as signal analysis, chemometrics and image analysis, objects observed by sensors are represented by higher-order generalizations of vectors and matrices i.e. several sets of features measured on objects, as for example, data collected at different times or conditions. The structure in which a set of objects with this representation is organized is called multi-way data.

Multi-way data analysis [98] is the extension of multivariate analysis when data is arranged in this multi-way structure. The most common is the three-way (3D) array, where objects are represented by matrices (2D representation) e.g. signals (objects, in the first direction of the array) represented by time points in the second direction and frequency in the third direction of the three-way data. The information obtained from such structures, e.g. interrelations between the different sets of features, can be advantageous for many purposes as regression and/or classification. Data will not be analyzed optimally by the traditional multivariate methods, which do not take into account the multi-way structure. If objects are analyzed in a 1D representation derived (unfolding for example) from the original higher-order one, the information of objects properties in one of the directions will be ignored. This may deteriorate the results. Moreover, fictitious relationships between the features of the different directions may be created in this process. A number of methods for multi-way analysis have been proposed [98, 64]. Most of these methods are for exploratory and regression purposes. Classification has been studied much less. This might be caused by the lack of classification tools able to operate on objects represented by multi-way arrays, that use all the available information.

Traditionally, pattern recognition systems are based on feature representations. Every object is represented by a feature vector that is constituted of object attributes that are characteristic for the differences between the classes. The representation itself neglects possible dependencies between these attributes. They have to be found from the statistics in a training set. This holds in particular when as features samples of an object image a time signal or a spectrum is taken. The connectivity between pixels, time samples or frequencies is lost in the representation. By operations like a PCA they may be re-found, but at the cost of a sufficient size of the training set.

To overcome the above problem of the feature representation, the Dissimilarity Representation (DR) has been developed [94, 85, 88]. It is based on a direct comparison of the total objects based on a dissimilarity measure. In this way, the geometry and the structure of a class

are determined by a user defined dissimilarity measure in which application background information may be expressed. As this measure may respect the fact that the object has some shape in an image, as a function of time or in its spectrum, the above problem can be avoided. The dissimilarity representation may generate a dissimilarity space, which is a general vector space. It can be used to train any of the traditional classifiers by a proper training set, represented by its dissimilarities to a selected set of prototype objects. This representation set may also be randomly selected or even be the entire training set.

In this paper, we introduce the use of the DR as a tool for classifying three-way data such that objects are analyzed in their 2D structure. Consequently, the relationship between the object properties in the different feature directions of the three-way array can be included if a suitable dissimilarity measure is selected. Moreover, the relations between objects are analyzed in the dissimilarity space. Thus, the key issue in this process is to find a dissimilarity measure that takes into account all relevant object differences for their classification. Information about the data that is not considered in the traditional representation or approaches can be included in the dissimilarity measure.

Although the introduced approach can be theoretically applied to any type of three-way data, we will focus in this paper on three-way spectral data e.g. signals represented by time points in one direction and frequency components in the other direction. With this purpose, we also try to construct a 2D measure that makes use of the 2D nature of the objects, and extracts the functional information e.g. shape from this type of data. We will show how, by making use of the 2D structure, the discrimination of objects improves with respect to that obtained just using the traditional vectorial (1D) representation.

The paper is organized as follows. In Section 3.2.2, related works about DR e.g. on spectral data and other 2D measures proposed in the literature are analyzed. In Section 3.2.3, a brief introduction to the DR theory is given. In addition, the proposed generalization of the DR for three-way data is presented, together with the new 2D dissimilarity measure. This measure is compared to the other related measures in Section 3.2.2. The comparison done is in terms of the analyzed data and the computational complexity. In Section 3.2.5, the materials and methods applied in the experimental section are detailed. Following, the experimental results are presented and discussed in Section 3.2.6. In this section, in order to show the advantage of using the 2D representation of objects, the proposed approach is compared to the 1D representation of the analyzed data sets. Moreover, the new measure is compared to other ones in the literature for the two three-way spectral data sets analyzed in the paper. Some characteristics of the proposed measure are also analyzed. Finally, the drawn conclusions are presented in Section 3.2.7.

3.2.2 Related studies

Several studies to classify spectral data starting from their vectorial feature e.g. wavelengths representation have been done [50, 135, 139, 116, 16, 17, 33]. Nevertheless, for a wide variety of problems, the structure of the data can often be more complex than this; one can have several sets of variables measured on different objects, as for example, data collected at different times or conditions. There have not been many attempts to use this multi-dimensional representation as it is, and take advantage of it for a better discrimination of objects.

In signal processing, multi-dimensional representation of objects (mainly 2D), has been used e.g. video signal, space-time-wave analysis for source localization and detection [15], blind multiuser detection-estimation in direct-sequence code-division multiple-access (DS-CDMA) communication [117], 3-D radar clutter modeling and mitigation [71, 153, 77], classification of sonar contacts [84], blind spatial signature estimation [112], detection of epilepsy [2] and alzheimer [67] from electroencephalography signals, analysis of seismic signals [14, 69, 76], and sound and speech recognition [111, 35], among others. However, what it is mainly done in these researches is to reduce the multi-way data to a vector for each object, by unfolding it in one of the features

direction e.g. from a spectrogram or scalogram, averaging in time to get the frequency spectrum, or vice versa. Afterwards, the authors do the processing of the data. For the classification task, a traditional classifier is applied after the data is transformed e.g. linear discriminant classifier and neural networks mainly [153, 69, 76, 111, 35].

In other research fields like chemometrics [126], the multi-way data analysis [64, 4, 98] has been one of the main topics for the last few years, but most efforts are directed to solve regression and exploratory analysis problems [126, 4]. These methods for multi-way data analysis have also been used in signal processing, but mainly for exploratory analysis (looking for the interrelation between the features in the different directions of the multi-dimensional array) or for dimensionality reduction purposes [71, 84, 117]; applying a traditional classifier afterwards. So, classification has not been that explored in multi-way analysis [98, 115, 40, 52]. However, this approach is just based on a numerical analysis, where other aspects like the shape of the signal, which can also have valuable information for the discrimination between different patterns of signals (classification), is not taken into account.

On the other hand, although it is a rather new technique, the DR has been applied in many fields e.g. image analysis [90, 61] and spectral data analysis [88, 85]. In the latter, the DR has shown to be advantageous as it addresses the dimensionality ill-posed problem of most spectral data sets. Normally, this kind of data, are characterized by having a few objects but represented with very high-dimensional vectors. Consequently, by using the DR, a dimensionality reduction is achieved, even if the entire training set is used as representation set. Moreover, background knowledge on the data can be expressed in the dissimilarity measure. Some studies have been done to find suitable dissimilarity measures for spectral data [88] to apply the DR approach, and more specifically for seismic volcanic signals [85, 86] and chemical data [97].

The DR can be generated from any other representation of the objects e.g. vectors of numbers, graphs, or multi-way data, as long as the suitable dissimilarity measure is found. However, it has not been done yet for the last one. Nonetheless, some 2D measures (mainly for image processing) have been used and developed for the dissimilarity-based k-Nearest Neighbor (k-NN) classifier, with the purpose of making use of the data two-dimensional array (2D) representation e.g. Frobenius [156], Yang [157] and Volume [78]distances, and the Assembled Matrix Distance (AMD) [161]. But none of these measures takes the spectral information, e.g. continuity, shape, into account. Some measures are analyzed in the subsequent sections together with the one proposed in this paper.

3.2.3 Dissimilarity Representation from Three-way data

The Dissimilarity Representation (DR) [94], was proposed as a more flexible representation than the feature representation, with the purpose of having more information about the structure of the objects. It is seen as a link between the statistical and structural approaches, as both types of patterns can be described by the (dis)similarity measure. The DR is also based on the role that (dis)similarities play in a class composition, where objects from the same class should be similar and objects from different classes should be different (compactness property). Hence, it should be easier for the classifiers to discriminate between them.

Using the DR, classifiers are trained in the space of the proximities between objects, instead of the traditional feature space. Thus, in place of the feature matrix $\mathbf{X} \in \mathbb{R}^{n \times q}$, where n runs over the objects and q over the features, the set of objects is represented by the matrix $\mathbf{D}(\mathbf{X}, \mathbf{R})$. This matrix contains the dissimilarity values $d(x_i, r_j)$ between each object x_i of $\mathbf{X}$ and the objects r_j of the representation set $\mathbf{R}(r_1, ..., r_h)$, where h is the number of prototypes. We build from this matrix a dissimilarity space. Objects are represented in this space by the row vectors of the dissimilarity matrix. Each dimension corresponds to the dissimilarities with one of the representation objects.

The elements of $\mathbf{R}$ are called prototypes, and have preferably to be selected by a prototype

selection method [94]. These prototypes are usually the most representative objects of each class, $\mathbf{R} \subseteq \mathbf{X}$ or $\mathbf{X}$ itself, resulting in a square dissimilarity matrix $\mathbf{D}(\mathbf{X}, \mathbf{X})$. $\mathbf{R}$ and $\mathbf{X}$ can also be chosen as different sets. As dissimilarities are computed to $\mathbf{R}$, a dimensionality reduction is reached if a good, small set can be found, resulting in less computationally expensive representation and classifiers.

For a t-dimensional array $\underline{Y} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_t}$ (See Fig. 3.1), the theory of the DR is the same. In fact, one of the advantages of the DR is that it can be generated from any representation of the objects e.g. vectors of numbers, graphs, as long as we have a proper dissimilarity measure. This applies also to the multi-way data. Originally, each object is represented by a (t-1)-dimensional array of numerical values and all the objects together compose the t-dimensional array. Hence, to obtain the dissimilarity space, a mapping $\phi(\cdot, \underline{R}) : \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_{t-1}} \rightarrow \mathbb{R}^h$ is defined, such that for every object $\underline{Y_i}$, $\phi(\underline{Y_i}, \underline{R}) = [d(\underline{Y_i}, \underline{R_1}), d(\underline{Y_i}, \underline{R_2}), ..., d(\underline{Y_i}, \underline{R_h})]$. , where h is the number of prototypes. Classifiers are then built in this space, as in any feature space.

Thus, to apply this approach to any classification problem the following steps should be

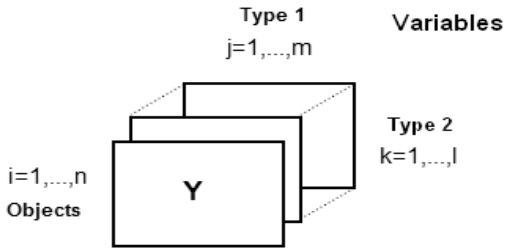


Figure 3.1: Design of a three-way array

followed:

1. Design the three-way (multi-way) data from the problem at hand, like $objects \times features1 \times features2$.

2. Define the dissimilarity measure according to the characteristics of the data. This measure should include all possible relevant knowledge, e.g. on the (in) significance of the tails of a spectrum.

3. Compute the dissimilarity matrix between the new objects and the representative objects (prototypes).

4. Prototype selection (if needed, depending on the classifier and the computational demands), with one of the methods reported in the references.

5. Build a classifier, using the dissimilarity matrix as input data, as the new description of the objects will be based on their dissimilarities to the prototypes from each class.

The issue to be addressed here is how to obtain the dissimilarities from the multi-way representation. Many ideas can arise to do this transformation. We propose as a first approach, focusing in three-way data, to take each object (matrix) (matrix) Y of $\underline{Y}$ and compute the dissimilarities between them by a 2D dissimilarity measure. Some 2D measures have been proposed in [161] for face and palm-print recognition. However, the selection of the suitable measure for the problem at hand is a very important aspect in the DR approach. To deepen in this task, we will focus in our case of study on three-way spectral data i.e. objects represented by spectra and/or time signals in the two feature directions. Thus, each object is represented by a matrix (2D). A comparative study is made about the characteristics of each data set and the dissimilarity measure to be used. A 2D dissimilarity measure is proposed.

3.2.4 2Dshape measure

In many types of data e.g. spectral data, it is necessary to take the shape information into account. In this way, the observations in the spectrum can be seen as continuous single entities, instead of sets of different features.

As mentioned in Section 3.2.2, some measures for 2D representation of objects have been proposed. Assume that the three-way array $\underline{Y} \in \mathbb{R}^{n \times m \times l}$, where n is the number of objects, and m and l the number of features (related to a single object) in each of the other directions respectively; $\forall j = 1, 2, \ldots, m$ and $k = 1, 2, \ldots, l$. Then, the AMD measure [161] for two objects $\mathbf{Y}_a$ and $\mathbf{Y}_b$ of $\underline{Y}$ is defined as:

$$d_{AMD}(\mathbf{Y}_a, \mathbf{Y}_b) = \left(\sum_{k=1}^{l} \left(\sum_{j=1}^{m} (\mathbf{Y}_{a,j,k} - \mathbf{Y}_{b,j,k})^2 \right)^{p/2} \right)^{1/p} \tag{3.1}$$

The power p is used to emphasize either small or large differences between the elements, in dependence of the problem at hand. If $p < 1$, all the differences are reduced, thus the larger ones do not interfere much in the measure. On the other hand, if $p > 1$, the larger differences will be more pronounced, resulting in a heavy influence on the measure. This measure is a generalization of the Frobenius and Yang distance measures referenced in Section 3.2.2. When $p = 1$ in AMD, it is the same as the Yang distance, and for $p = 2$ is then the Frobenius distance.

These measures could be a good option when the spectral (functional) information can be assumed to be present in the data representation. However, this is not usually the case, as spectra are observed and recorded discretely. Consequently, they are analyzed with multivariate data analysis techniques that consider the spectrum as high-dimensional vectors of different but high-correlated features, instead of a continuous single entity. Therefore, when the information is not taken into account in the representation of the data, the dissimilarity measure has to take care of it.

Hence, considering the results obtained with the Shape measure (Manhattan distance on the first Gaussian derivatives) for simple spectra [88], we propose to make use of the derivatives into the AMD measure. In such a way, we can take the ordering information into account as well as the shape of the spectra. A principle of the DR approach is that, instead of a single representation of a problem, one may also consider either a complex representation, built from many dissimilarity representations, where different aspects of the data are described in various ways [94]. Based on this and the previously stated, and in a way that the information available in both directions of the 2D data can be taken into account, we define the 2DShape dissimilarity measure as follows:

1. Compute the matrix D^1

$$D^1_{a,b} = \left(\sum_{k=1}^{l} \left(\sum_{j=1}^{m} (\mathbf{Y}^{\sigma_1}_{a,j,k} - \mathbf{Y}^{\sigma_1}_{b,j,k})^2 \right)^{p_1/2} \right)^{1/p_1},$$

$$\mathbf{Y}^{\sigma_1}_{i,j,\cdot} = \frac{\mathrm{d}}{\mathrm{d}_j} G(j, \sigma_1) * \mathbf{Y}_{i,j,\cdot}$$

2. Compute the matrix D^2

$$D^2_{a,b} = \left(\sum_{j=1}^{m} \left(\sum_{k=1}^{l} (\mathbf{Y}^{\sigma_2}_{a,j,k} - \mathbf{Y}^{\sigma_2}_{b,j,k})^2 \right)^{p_2/2} \right)^{1/p_2},$$

$$\mathbf{Y}^{\sigma_2}_{i,\cdot,k} = \frac{\mathrm{d}}{\mathrm{d}_k} G(k, \sigma_2) * \mathbf{Y}_{i,\cdot,k}$$

3. Combine both dissimilarity matrices $D = \alpha_1 D^1 + \alpha_2 D^2$

The variables $\mathbf{Y}_{i,j,\cdot}$ and $\mathbf{Y}_{i,\cdot,k}$, stand for the k-th columns and the j-th rows of the i-th matrix (object); $\forall i = 1, 2, \ldots, n$. Their expressions correspond to the computation of the first Gaussian (that is what G stands for) derivatives of spectra. Thus, a smoothing (blurring) is done by a convolution process ($*$) with a gaussian filter and σ stands for a smoothing parameter [88]. The dissimilarities in step 1 and step 2 correspond to the first and second directions respectively, as indicated by the notation e.g. spectra and time. This measure can also be used in three-way data where there are no variations in shape in one of the directions. In this case, it is enough to use the AMD measure in step 1 or step 2 only, such that only the differences in area are compared. Moreover, different p values can be applied in the different directions. In the combination step, we included a weight for scaling. In this case, we defined $\alpha_c = \frac{D^c}{max(D^c)}$, to normalize the dissimilarity matrices.

If we analyze the computational complexity of all these measures, they are in the order of $\mathcal{O}(m.l)$, as our objects have two-dimensions (two types of features). In general, to compute the whole dissimilarity matrices, the computational complexity of all of them is $\mathcal{O}(n^2.m.l)$. We are assuming here the worst case, in which all training objects are used as prototypes. Nevertheless, if we really take into account the number of operations required for each measure, they are different. Between the Frobenius, Yang and AMD measures there are some slight differences in the number of operations required. Nevertheless, when we analyze the 2DShape measure, it requires 2 times the number of operations needed for the other measures, plus the last sum operation. Moreover, in case of applying the derivatives, these operations are also added, and it depends on the selected smoothing parameter σ. So, from this point of view, depending on the problem at hand, it is needed a trade-off between computational complexity and classification accuracy.

3.2.5 Materials and Methods

Two three-way spectral data sets will be studied in this paper. The first is a public domain data set and the description has been taken from the website [152, 124] for a better understanding of this paper. It consists of examples of red wine, produced from the same grape (Cabernet Sauvignon) and belonging to different geographical areas and producers. They were collected from local supermarkets and analyzed by means of HS-GC-MS (headspace gas chromatography/mass spectrometry). Separation of aroma compounds was carried out on a gas chromatography system (2700 columns from the scans of chromatographic profile). For each example, a mass spectrum scan (m/z: 5-204) measured at the 2700 elution time-points was obtained, providing a data cube of size $44 \times 2700 \times 200$ i.e. examples (objects) in first direction, elution time points in second direction and mass spectrum in third direction. The data set is composed of examples from 3 different geographical areas: South America (21 objects), Australia (12 objects) and South Africa (11 objects). For the two-way representation (1D representation of objects) of the data, the three-way data was unfolded in its second direction, obtaining a matrix of size 44×540000. All-zero columns were deleted in this representation (none of the objects have information in these columns), so the final data set has a size of 44×117060. The dissimilarity representation has a size of 44×44.

The second data set corresponds to seismic signals from the ice-capped Nevado del Ruiz volcano in the Colombian Andes, currently studied by the Volcanological and Seismological Observatory at Manizales. Signals were digitized at 100.16 Hz sampling frequency by using a 12 bit analog-to-digital converter. The data set for the experiments is composed of 12032-point signals of two classes of volcanic events: 235 of Long-Period (LP) earthquakes, and 235 of

Volcano-Tectonic (VT) earthquakes. A 2D time-frequency representation was computed with Short-Time Fourier Transform (STFT), obtaining a spectrogram from each signal [16]. To compute these spectrograms, trying to achieve a trade-off between time and frequency resolution, a 256-point (window size) STFT was calculated with 50% overlap. With this technique, it can be known what frequency intervals are present in a time interval of the signal and use it for the discrimination between classes. The concatenation of the spectrograms of the different signals (objects) will result in a $470 \times 93 \times 129$ three-way data i.e. events (objects) in first direction, time points in second direction and frequency components in the third direction. For the 1D (spectral) representation of each object, we have computed the spectrum by using a 12032-point Fast Fourier Transform (FFT), leading to a 470×12032 data. Consequently, the information in the whole signal is analyzed in both its 1D and 2D representation. The dissimilarity representation has a size of 470×100.

For both data sets, the related measures in Section 3.2.4 are used to obtain the DR from their 2D representation. This way, we can analyze which should be a suitable measure for each problem, and compare how the results behave in each case. The experiments are also made for several p values. Hence, we can also compare all the measures related before. For the wine data set, $p = [0.1, 0.5, 1, 2, 3]$ and also for both directions of the volcano data set. In the case of the 2Dshape measure, experiments were also ran by exchanging the p values of both directions. The σ parameter was optimized in a grid-search procedure with 10-fold cross-validation, for the different p values in both data sets. In the case of wine data, as the number of objects is so small, the optimization procedure was done with the whole data set. The best results were obtained for $\sigma = 5$. In the case of the seismic volcanic data, 170 objects (85 of each class) were used to optimize the σ parameter for each direction. The rest of the data was then used to evaluate the classification performances, by using the best σ values ($\sigma = 2$ for the time direction and $\sigma = 3$ for the frequency direction) overall p.

As mentioned previously, we will make a comparison between the classification accuracy by making use of the 2D structure or just the 1D. In the case of the 2D representation, the 2D measures explained in Section 3.2.4 will be used and compared. For 1D representation, we will use the shape measure introduced in [88]. It is defined as the Manhattan distance on the Gaussian derivatives of the spectra. The σ parameter in this measure was also optimized in a cross-validation procedure and the best results were achieved with $\sigma = 15$ for the volcanic data set and $\sigma = 20$ for the wine data set.

The Regularized Linear Discriminant classifier (RLDC) [43, 36] was built on the DR obtained from the different representations of the two data sets. In order to find the regularization parameters, an automatic regularization (optimization over training set by cross-validation) process was done. Experiments were repeated 10 times. Training and test objects were randomly chosen from the total data sets, in a 10-fold cross-validation process. For the wine data set, as the size of the training set is so small, we decided to use all the objects as prototypes. A random prototype selection was performed with several numbers of prototypes for the volcanic data set. The best results were obtained for 100 prototypes, which are the ones shown here. In both cases, the same training and test sets were used for all the representations, so the results can be comparable. Classifiers performances are evaluated in terms of the Average Classification Error (ACE), and the standard deviation from the different repetitions is taken into account.

The experiments were all performed in Matlab. For the computation of the spectrograms, we used the Signal Processing Toolbox from Matlab, and PRTools toolbox [149] for the computation of DR and classification of the data.

3.2.6 Experimental Results and Discussion

In this section we present several analyses. First, we want to compare for the spectral data sets, the results by applying the proposed measure and the other 2D measures of the literature.

This is in order to demonstrate the importance of selecting a suitable measure for the data at hand. An analysis of the influence of the parameter p in the results will also be performed. Afterwards, a comparison between the 2D and 1D representation of the objects will be done, to show the advantage of making use of the 2D structure data for a better discrimination of classes.

2DShape measure vs others measures

In chemometrics, in the case of the techniques combined with chromatography, in the chromatography direction (2nd), we will have the eluded peaks for all the components present in the substances. But in the spectral mode (third direction, mass spectra for GC-MS), each eluded component will only have one mass spectrum (mass fragments in which the molecule decomposes) independently of the class. This means that, with these techniques, if the classes differ because they have different components, the only thing we will see is the absence/presence of the peak or some differences in the concentration of the mass fragments.

Nevertheless, the other difference between classes that we can find is the relation between the eluded components in the chromatogram i.e. how the concentration of one of the peaks varies with respect to the others, for the different classes. In this case, it is important to take shape into account, because there is information in the ordering of the components (peaks) with different concentrations and also continuity.

Thus, for the wine data set we propose to adapt the 2Dshape measure defined in Section 3.2.4, which takes the information of both directions into account, to the specificities of the data. When computing the D^1 matrix for the chromatography direction, we will use the Gaussian derivatives to take into account the shape in the changes of concentration in the neighboring components. However, for the D^2 matrix from the mass spectra mode, the use of derivatives is meaningless, because there is no continuity between the mass fragments; just the differences between the concentration of the mass fragments will be computed.

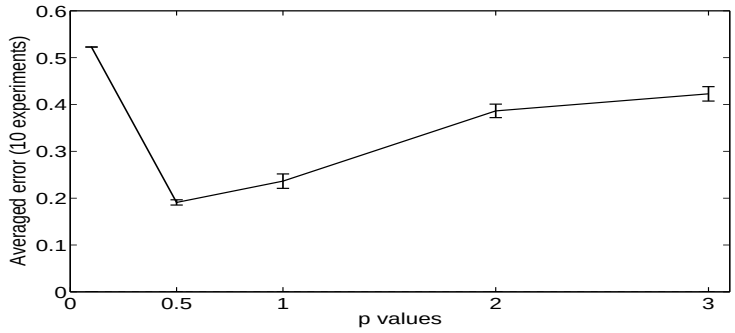


Figure 3.2: Average cross-validation error (with standard deviation) for the AMD measure for different p-values for Wine data set.

In Figures 3.2 and 3.3, the results for AMD and 2DShape measures are shown. Notice again that when $p = 1$, we are using the Yang distance and when $p = 2$, the Frobenius. This way, we are comparing the 3 measures related in Section 3.2.4. If we analyze Figure 3.2, we can see that the errors are in a range of $20\% - 50\%$ for the three measures (no functional information about the data is taken into account). Nevertheless, when we look at Figure 3.3, the highest error is around 28%, and coincides with the p-values for which the worst error was obtained with the AMD measure. It can be noticed that, by measuring the information in both directions and taking the functional e.g. shape, information into account, we have decreased the errors to around $12\% - 28\%$. Hence, from this data set we might conclude that, even when the p

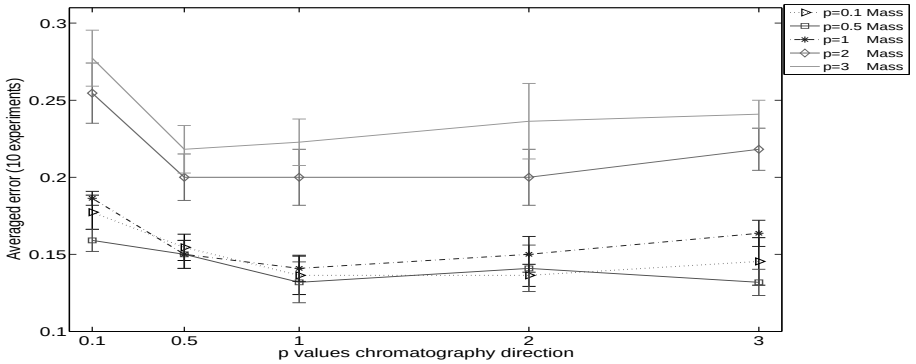


Figure 3.3: Average cross-validation error (with standard deviation) for the 2DShape measure for different p-values on both directions for Wine data set.

parameter is not optimized, by taking the spectral information into account better results can be achieved.

In Figure 3.4, the classification results for the DR from the different directions separately, are shown. We can see from this figure that, there is one direction for which better results are obtained. In this case, it is the second direction (chromatography), which makes sense because we are analyzing the relation of concentration of components (shape) for the different classes. However, the third direction (mass spectrum), although is not that informative can also help to discriminate. From this figure and Figure 3.3, we can notice that if we combine the dissimilarity matrices from both directions (using the p-values for which the best results are obtained in each direction separately) the independent results for each direction are outperformed. This corroborates our initial hypothesis of getting a better discrimination between classes, if we take the information of both directions into account. In this case, we can also see how the lowest error achieved in the combination 12%, was obtained by using $p = 1$ in the second direction and $p = 0.5$ in the third direction. These are the p-values for which the best results were obtained in the independent analysis.

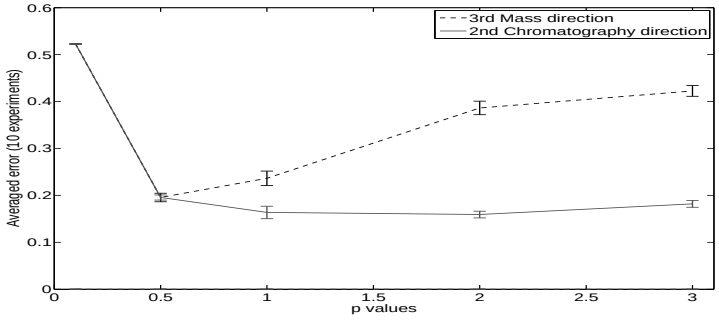


Figure 3.4: Average cross-validation error(with standard deviation) for the two directions separately in the 2DShape measure for Wine data set. Different p-values are analyzed.

A good example where the proposed measure in Section 3.2.4 can be useful is in this time-frequency representation of the second data set obtained by spectrograms. In this case, shape changes are present in the spectral (frequency) direction and connectivity in the time direction (due to the windows overlapping). In the next figures, the ACE are shown for the AMD (different p-values) measure in Figure 3.5 and 2DShape measure in Figure 3.6.

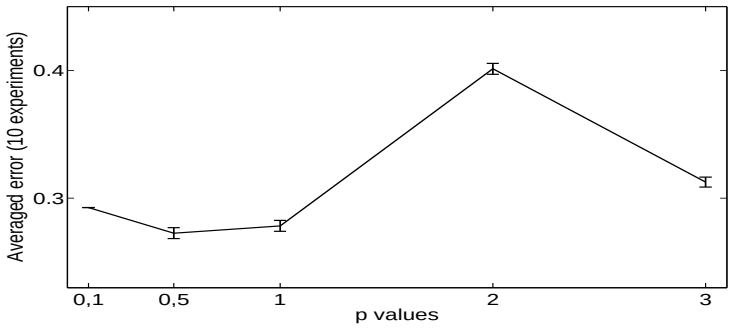


Figure 3.5: Average cross-validation error (with standard deviation) for the AMD measure for different p-values for Seismic volcanic data set.

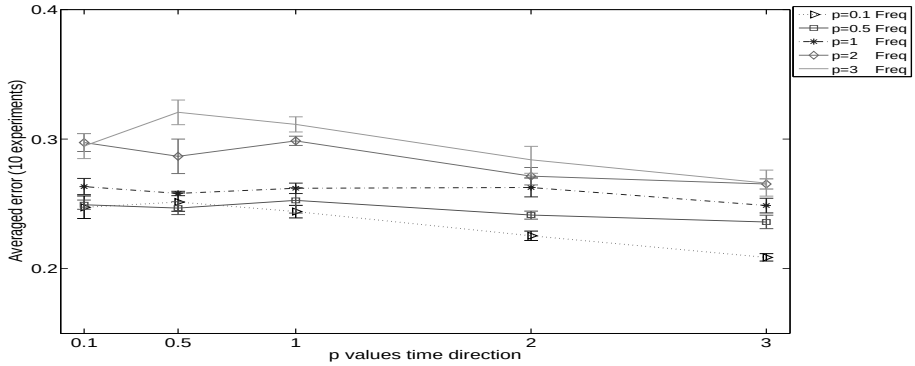


Figure 3.6: Average cross-validation error (with standard deviation) for the 2DShape measure for different p-values on both directions for Seismic volcanic data set.

From the figures, we can observe a similar behavior to the ones of the Wine data set. With the proposed measure, where the shape and continuity information is taken into account, we got to outperform the AMD from a lowest error of around 27% of ACE, to 21%. This suggests that the proposed 2D measure is capable of capturing the information needed. Also, we should take a look at the results (ACE) when analyzing the directions separately (See Figure 3.7). The results for the time direction (2nd) are not very good. The ones for the third direction (frequency) on the other hand, are much better. This also makes sense, and is corroborated by the fact that spectral-based classification is often used for this type of data, as spectral content of these signals allow the discrimination between the events (events do not change in time heavily). Nevertheless, when the information from both directions is taken into account, better results are achieved, as shown in the results in Figure 3.6. Moreover, we observe here that, in the case where the information from one of the directions is not sufficiently discriminative, the least we will get, are results similar to the ones for the discriminative direction.

We can also see from Figures 3.6 and 3.7 that, the best performances with the 2DShape measure, are achieved for the p-values for which the best results are obtained in each direction independently.

1D vs 2D object representation

In this section we make a comparison between the classification results when using the vectorial (1D) of objects and when making use of the 2D structure. In Table 3.1, the ACE for the DR from

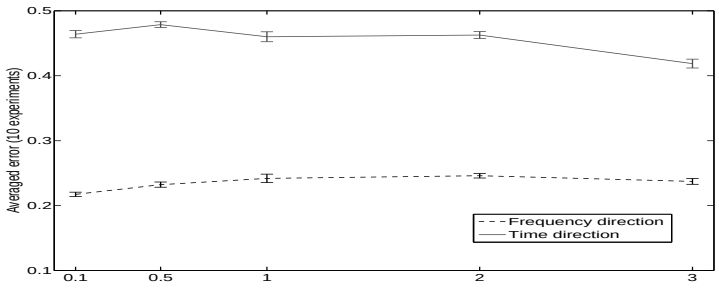


Figure 3.7: Average cross-validation error(with standard deviation) for the two directions separately in the 2DShape measure for Seismic volcanic data set. Different p-values are analyzed.

the 1D and 2D representations of both data sets can be observed. In Section 3.2.6, an analysis of different 2D measures was done, and in both data sets the best results were obtained when applying a version of the proposed 2DShape measure. Thus, in this section, the ACE values shown in Table 3.1 for the 2D representation of both data sets are those best values obtained in the previous section. For the Wine data set, the results achieved with the unfolding procedure are not very good. By applying the DR on this representation, the high-dimensionality of the obtained data (which is one of the main problems in this procedure) is reduced. This is because now in the new dissimilarity space, the dimensionality is given by the number of objects in the representation set (training set in this data set). Still, fictitious connections between the end point of the measurements in one direction and the start of the others are inserted. In any direction that we do the unfolding, the same phenomenon will happen; some relation will be lost or its benefit will not be used. However, these results ar notably outperformed when using the 2D representation.

Representations	Wine	Volcano
1D	36.1(1.4)	30.2(0.4)
2D	12.1(0.5)	20.9(0.2)

Table 3.1: Average cross-validation error in % (with standard deviation) for Wine and Seismic Volcanic data sets

In the case of the Seismic volcanic data set, the differences in 1D spectral content of a signal allow for the discrimination between the events. Nevertheless, with this representation we are not able to use the changes of frequency content in time to separate classes. It can be also observed in Table 3.1 that, the ACE on the dissimilarity space generated from just the spectral data (1D) is around 30%. Nevertheless, when we analyze the error of the DR from the 2D representation we see a significant improvement. This ratifies the fact that the time-frequency relation is more discriminative than just the spectral information.

Thus, it can be observed that it is usually advantageous to make use of the multi-dimensional structure of the data for the classification process. The information and relationship between the features of the different directions can be more discriminative, than if we just obtain a vector (1D) of features from its 2D representation, ignoring its original structure.

3.2.7 Conclusions

We introduced the use of the Dissimilarity Representation as a tool for classifying three-way data. In this approach, objects are analyzed in their matrix (2D structure) representation by using 2D measures. Moreover, information about the data that is missing in the original

representation e.g. shape, can be considered in it. We developed a new 2D dissimilarity measure that allows taking into account the shape and continuity information in the directions of three-way spectral data. Furthermore, the relationship between the different dimensions is taken into account in this measure. It was compared to other three measures of the literature, in order to proof the importance of the selection of a suitable dissimilarity measure for the problem at hand. We also showed that, even when there is more discriminative information in one direction of the three-way data than in other, the results usually improve by combining.

The good performance of classifiers on the DR from the 2D representation of the objects, compared with the one from the traditional 1D, shows that this approach can be a good solution for the classification of objects with a 2D structure. Although this paper was focused on the solution for three-way data, it can be extended to multi-way e.g. in chemometrics, molecular entities of a substance can be separated by size on a chromatographic system and then detected by fluorescence, leading to a four-way data of $(objects) \times (fractions) \times (excitation) \times (emission)$. The proposed measure, as the dissimilarity representation, could be extended to multi-way data; it is part of the future work. Moreover, this procedure could be used in other types of problems where shape changes help for discrimination, like in image processing e.g. for the classification of faces.

3.2.8 Acknowledgment

We acknowledge financial support from the FET programme within the EU FP7, under the project "Similarity-based Pattern Analysis and Recognition - SIMBAD" (contract 213250). We would also like to thank to the project Cálculo científico para catacterización e identificación en problemas dinámicos (code Hermes 10722) granted by Universidad Nacional de Colombia.

3.3 A study on the influence of shape in classifying small spectral data sets

This section is a modified version of the article published as 'A study on the influence of shape in classifying small spectral data sets', by D. Porro-Muñoz, R.P.W Duin, I. Talavera and M. Orozco-Alzate, in *Lecture notes in computer science vol. 7005, Springer Verlag, Berlin, 1st International Workshop on Similarity-Based Pattern Analysis and Recognition, Proc. SIMBAD2011 (Venice, Italy)*, 306-320, October 2011. In this new version, experiments were repeated several times in order to obtain smoother error curves. The sections of the paper were structured differently to make them fit thesis style. Notation was also changed in order to have a unified notation in the whole thesis.

Abstract

Classification of spectral data has raised a growing interest in may research areas. However, this type of data usually suffers from the curse of dimensionality. This deteriorates the performance of most statistical methods and/or classifiers. A recently proposed alternative that can help avoiding this problem is the Dissimilarity Representation, in which objects are represented by their dissimilarities to representative objects from each class. However, this approach depends on the selection of a suitable dissimilarity measure. For spectra, the incorporation of information on their shape, can be important in order to achieve a good discrimination. In this paper, we make a study on the benefit of using a measure that takes the shape of spectra into account. We show that the shape-based measure not only leads to better classification results, but that a lower number of objects suffices to achieve a good performance.

3.3.1 Introduction

Classification of unknown objects is one of the main problems in many research areas. Object representation plays an important role in this task. In practical classification problems, the number of training objects is usually very small, represented by a very large number of features that are not always the best to describe them. Many studies have been done on this issue; when only a certain number of objects is available, a peaking phenomenon occurs in the classification accuracy as the number of features is increased. This is known as the curse of dimensionality [45, 109, 38]. Hence, the ideal situation in order to obtain a good classifier would be to have at least as many objects as features. It appears to be difficult to achieve this in a number of real-world problems.

A type of data that has raised a growing interest in advanced approaches to its automatic analysis is the spectral data. It is due to the increasing possibilities of the different research fields e.g. chemometrics and signal processing, to obtain it, and the usefulness of the spectral information to describe and differentiate objects of different classes. This is the type of application where data sets are small because the cost to obtain them is very high, and they are usually much smaller than the dimensionality of the space. The traditional way of representing spectra is by sampling, as a sequence of individual observations made on the objects. The higher the sampling resolution, the more accurate the spectrum is described, which implies a representation in a high-dimensional space. However, this way of representation is not good for traditional procedures. It makes them suffer the curse of dimensionality. Furthermore, discriminative knowledge about spectra e.g. the continuity between the measured values, shape, is not taken into account in the traditional high-dimensional feature-based representation. Thus, it does not help avoiding the problem.

Recent works have studied alternative object representations instead of features, demonstrating that the curse of dimensionality can be avoided [38]. A recently developed alternative in pattern recognition is the Dissimilarity Representation (DR) [94]. It is based on the important role that pairwise dissimilarities between objects play. Classifiers may be built in the dissimilarity space generated by a representation set. In this way, the geometry and the structure of a class are determined by a user defined dissimilarity measure, in which application background information may be expressed. It is important to remark that, any traditional classifier that is defined on feature spaces can also be used in the dissimilarity space.

With the DR, the problem of building classifiers in high-dimensional spaces can be tackled, as the dimensionality will depend now on the size of the representation set (usually smaller or equal to the size of the training set). However, the main issue in this approach is the selection of a suitable measure for the problem at hand. The more discriminative information we take into account when designing the dissimilarity measure, the more compact the classes are. The centroid of the data should remain approximately the same and the average distance to this

mean should decrease or be constant [38], requiring less objects for its description and a good classification accuracy.

Due to benefits that the DR has shown, it has been explored in several applications like the discrimination of spectral data [94, 85, 88, 101]. In this paper, we will make an exhaustive experimental study on the DR for spectral data. We will focus on the usefulness of taking the shape of the curve into account in the dissimilarity measure. It will be shown that this can help achieving good classification results in small sample (in this case sample is used to refer to objects) size problems. Recently, the use of the DR was also extended to 2D spectral data i.e. objects represented by matrices, where two types of spectral features are described [99]. Thus, the study will be generalized to this type of data. We will use three one-dimensional spectral data set and a 2D spectral one. In the experiments, we compare the classification accuracy in measures that do not take shape into account with a measure that does. This analysis is done for several training set sizes and representation set sizes, to see how the measures influence the results. Moreover, for the measure that takes shape into account, we study the sensitivity of the results to the optimization of the parameters (Gaussian filter parameter). The paper is structured as follows. In Section 3.3.2, a brief introduction to the DR will be done. Also, the 1D and 2D measures to be used in the experiments are referenced. Following, the data sets and experiments will be described in Section 3.3.4. Finally, a discussion and the drawn conclusions will be presented in Section 3.3.5.

3.3.2 Dissimilarity Representation Approach

The Dissimilarity Representation (DR) [94] was proposed as a more flexible representation of the objects than the feature representation, with the purpose of having more information about the structure of the objects. It is seen as a link between the statistical and structural approaches, as both types of patterns can be described by the (dis) similarity measure. The DR is also based on the role that (dis) similarities play in a class composition. Objects from the same class should be similar and objects from different classes should be different (compactness property). Hence, it should be easier for the classifiers to discriminate between them.

Using the DR, classifiers are trained in the space of the proximities between objects, instead of the traditional feature space. Thus, in place of the feature matrix $\mathbf{X} \in \mathbb{R}^{n \times q}$, where n runs over the objects and q over the features, the set of objects is represented by the matrix $\mathbf{D}(\mathbf{X}, \mathbf{R})$. This matrix contains the dissimilarity values $d(x_i, r_j)$ between each object x_i of $\mathbf{X}$ and the objects r_j of the representation set $\mathbf{R}(r_1, ..., r_h)$. We build from this matrix a dissimilarity space. Objects are represented in this space by the column vectors of the dissimilarity matrix. Each dimension corresponds to the dissimilarities with one of the representation objects.

When an object is represented by a matrix $\mathbf{Y} \in \mathbb{R}^{m \times l}$, the theory of the DR is the same [99]. In fact, one of the advantages of the DR is that it can be generated from any representation of the objects e.g. vectors of numbers, graphs, as long as we have a proper dissimilarity measure. Hence, to obtain the dissimilarity space, a mapping $\phi(\cdot, \mathbf{R}) : \mathbb{R}^{m \times l} \to \mathbb{R}^h$ is defined, such that for every object $\mathbf{Y}$, $\phi(\mathbf{Y}, \mathbf{R}) = [d(\mathbf{Y}, \mathbf{R}_1), d(\mathbf{Y}, \mathbf{R}_2), ..., d(\mathbf{Y}, \mathbf{R}_h)]$. Classifiers are then built in this space, as in any feature space.

The elements of $\mathbf{R}$ are called prototypes, and have preferably to be selected by a prototype selection method [94]. These prototypes are usually the most representative objects of each class, $\mathbf{R} \subseteq \mathbf{X}$ or $\mathbf{X}$ itself, resulting in a square dissimilarity $\mathbf{D}(\mathbf{X}, \mathbf{X})$. $\mathbf{R}$ and $\mathbf{X}$ can also be chosen as different sets. The same holds when objects are represented by matrices or higher order arrays. As dissimilarities are computed to $\mathbf{R}$, a dimensionality reduction is reached if a good, small set can be found, resulting in less computationally expensive classifiers.

3.3.3 1D and 2D dissimilarity measures for spectral data

A general dissimilarity measure for all types of data does not exist. Thus, the selection of the suitable measure for the problem at hand is the key issue in the DR approach. In recent studies, some 1D [88, 97] and 2D [99] measures have been studied and proposed for spectral data. Such is the case of the very well known Manhattan (L1-norm) and Euclidean distances. However, although the previous dissimilarities are of the most used measures in the comparisons of chemical spectral data, the connectivity between the measurements and their shape, is not taken into account in neither of them. The features could be easily reordered and the same dissimilarity value is obtained.

In [88], the authors propose to compute the Manhattan measure on the first Gaussian derivatives (See Eq. 3.2) of the curves (Shape measure). Thereby, the shape information that can be obtained from the derivatives is taken into account:

$$d(x_1, x_2) = \sum_{j=1}^{m} |x_{1j}^{\sigma} - x_{2j}^{\sigma}|, \quad x^{\sigma} = \frac{\mathrm{d}}{\mathrm{d}_j} G(j, \sigma) * x \tag{3.2}$$

The expression of x^{σ} corresponds to the computation of the first Gaussian (that is what G stands for) derivatives of spectra. A smoothing (blurring) is done by a convolution process ($*$) with a gaussian filter and σ stands for the smoothing parameter. Good performances have been obtained for chemical spectral data with this measure [88, 97].

For the 2D representation of objects, generalizations of the Manhattan and Euclidean distances have also been proposed. Assume that two objects $\mathbf{Y}_a$ and $\mathbf{Y}_b \in \mathbb{R}^{m \times l}$, where m and l are the number of features in each of the two directions respectively; $\forall j = 1, 2, \ldots, m$ and $k = 1, 2, \ldots, l$. Then, the AMD measure [161] is defined as:

$$d_{AMD}(\mathbf{Y}_a, \mathbf{Y}_b) = \left(\sum_{k=1}^{l} \left(\sum_{j=1}^{m} (\mathbf{Y}_{a,j,k} - \mathbf{Y}_{b,j,k})^2 \right)^{p/2} \right)^{1/p} \tag{3.3}$$

The power p is used to emphasize either small or large differences between the elements, depending on the problem at hand. If $p < 1$, all the differences are reduced, thus the larger ones do not interfere much in the measure. On the other hand, if $p > 1$, the larger differences will be more pronounced, resulting in a heavy influence on the measure. This measure is a generalization of the Frobenius [156] and Yang [157] distance measures. When $p = 1$ in AMD, it is the same as the Yang distance, and for $p = 2$ is then the Frobenius distance.

These measures could be a good option when the spectral (functional) information can be assumed to be present in the data representation. However, this is not the case. Recently, considering the results obtained with the Shape measure for simple spectra, a new version for 2D spectral data (2Dshape measure) was introduced [99]:

1. Compute the matrix D^1

$$D_{a,b}^1 = \left(\sum_{k=1}^{l} \left(\sum_{j=1}^{m} (\mathbf{Y}_{a,j,k}^{\sigma} - \mathbf{Y}_{b,j,k}^{\sigma})^2 \right)^{p_1/2} \right)^{1/p_1}, \quad \mathbf{Y}_{i,j,\cdot}^{\sigma} = \frac{\mathrm{d}}{\mathrm{d}_j} G(j, \sigma) * \mathbf{Y}_{i,j,\cdot}$$

2. Compute the matrix D^2

$$D_{a,b}^2 = \left(\sum_{j=1}^{m} \left(\sum_{k=1}^{l} (\mathbf{Y}_{a,j,k}^{\sigma} - \mathbf{Y}_{b,j,k}^{\sigma})^2 \right)^{p_2/2} \right)^{1/p_2}, \quad \mathbf{Y}_{i,\cdot,k}^{\sigma} = \frac{\mathrm{d}}{\mathrm{d}_k} G(k, \sigma) * \mathbf{Y}_{i,\cdot,k}$$

3. Combine both dissimilarity matrices $D = \alpha_1 D^1 + \alpha_2 D^2$

The variables $\mathbf{Y}_{i,j,\cdot}$ and $\mathbf{Y}_{i,\cdot,k}$, stand for the k-th columns and the j-th rows of the i-th matrix (object); $\forall i = 1, 2, \ldots, n$. Their expressions correspond to the computation of the first Gaussian (that is what G stands for) derivatives of spectra, as in the 1D measure. The dissimilarities in step 1 and step 2 correspond to the first and second directions respectively, as indicated by the notation e.g. spectra and time. This measure can also be used in three-way data where there are no variations in shape in one of the directions. In this case, it is enough to use the AMD measure in step 1 or step 2 only, such that only the differences in area are compared. With this measure, the shape of the spectra can be taken into account. The previously mentioned measures, which have been used for spectral data, will be used for the purpose of this paper.

3.3.4 Experimental Section

For the purpose of this paper, a set of experiments were conducted on small sample size data sets in high-dimensional spaces. Only one of them does not suffer from this problem, but still we want to show how also in this case, with the selection of a suitable dissimilarity measure, a reduced number of training objects can be enough to obtain good classification results with the DR. All of them consist of two-class classification problems. The data sets are described in the following subsection.

Data sets

The first data set, named Tecator, originates from the food industry [151]. It consists of 215 near infrared absorbance spectra of meat examples, recorded on a Tecator Infratec Food and Feed Analyzer. Each observation consists in a 100 channel absorbance spectrum in the 850-1050 nm wavelength range. It is associated to a content description of meat example, obtained by analytic chemistry. The classification problem consists in separating 77 meat examples with a high fat content (more than 20%), from 138 examples with a low fat content (less than 20%). Original spectra are preprocessed, each spectrum is reduced to zero mean and unit variance.

The second data set is a real-world data set, which was obtained from a cooperation with the Oil Industry in Cuba. It consists of 31 fuel examples of Fourier Transform Infrared (FT-IR) transmittance spectra in a wavelength range of 600-4000 cm^{-1}. A base line correction and smoothing were performed on the data. The classification problem consists in determining the fuel type of the examples: regular gasoline (16 objects) and especial gasoline (15 objects).

The third data set is another fuel real-world data set of 40 examples measured at 127 wavelengths in a range of 275-220 nm, but this time measures have been taken by a Ultra-Violet Visible (UV) spectrophotometer. The classification problem consists also in determining the fuel type of the examples: regular gasoline (23 objects) and especial gasoline (21 objects).

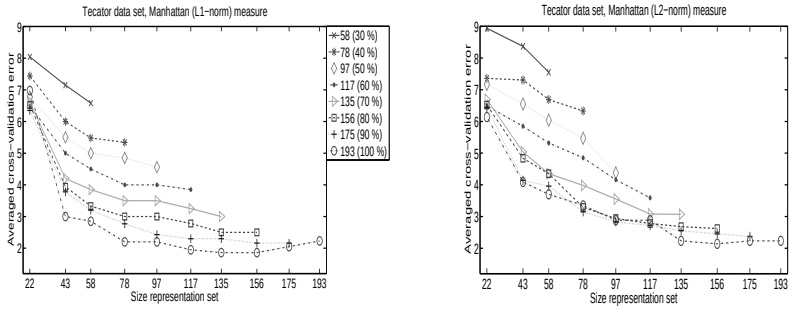
The fourth and last data set is a three-dimensional array, composed of objects naturally represented by 2D arrays. It is a public domain data set and the description has been taken from the website [152, 124] for a better understanding of this paper. It consists of examples of red wine belonging to different geographical areas and producers. They were analyzed by means of HS-GC-MS (headspace gas chromatography/mass spectrometry). Separation of aroma compounds was carried out on a gas chromatography system (2700 columns from the scans of chromatographic profile). For each example, a mass spectrum scan (m/z: 5-204) measured at the 2700 elution time-points was obtained, providing a data cube of size $33 \times 2700 \times 200$ i.e. examples (objects) in first direction, elution time points in second direction and mass spectrum in third direction. The data set is composed of examples from 2 different geographical areas: South America (21 objects) and Australia (12 objects).

Experiments and Discussion

A set of experiments are conducted on the four data sets. A Regularized Linear Discriminant Analysis classifier is built on the dissimilarity space obtained for the two dissimilarity measures that do no take shape information into account (Manhattan and Euclidean) and also for the Shape measure. For the later, several experiments are shown, with different values for parameter σ. In the figures, learning curves are shown for various sizes of training and representation sets. The main idea of this experimental set up is to show how the use of a suitable measure e.g. measures shape in spectral data, can influence not only in the classifiers accuracy, but on the sample size problem.

For each data set, subsets of different sizes $[30, 40, 50, 60, 70, 80$ and $90\%]$ from the total dissimilarity matrix were randomly chosen. Training and test objects were evaluated for each subset in a 10-fold cross-validation for Tecator data set and Leave-one-out (LOO) for the rest of the data sets. Experiments were repeated 50 times. Different sizes for the representation set were also randomly selected $[10, 20, 30, 40, 50, 60, 70, 80$ and $90\%]$. When using the Shape measure on the one-dimensional spectral data, the following values of σ were applied $[0.5, 1, 2, 3, 5, 7]$.

In the case of the Wine 2D spectral data, in the mass direction the classes only differ because they have different components. Therefore, the only thing we will see is the absence/presence of the peak or some differences in the concentration of the mass fragments. The other difference that we can find between these classes is related to the shape changes between the eluded components in the chromatogram i.e. how the concentration of one of the peaks varies with respect to the others, for the several classes. Thus, for the Wine data set we will use the 2Dshape measure. The D1 matrix will be computed for the chromatography direction. The Gaussian derivatives are applied to take into account the shape in the changes of concentration in the neighboring components. In this case, the following values of σ were applied $[1, 2, 5$ and $8]$. However, for the D2 matrix from the mass spectra mode, we will only compute the overall sum of the differences between the concentration of the mass fragments. The use of derivatives is meaningless, because there is no continuity between the mass fragments.



(a) (b)

Figure 3.8: Average cross-validation error (in %) for Tecator data set with (a) Manhattan and (b) Euclidean measures. The % of data is the size of the selected subset. A training and test set is selected in a 10-fold cross-validation for each subset.

In Figure 3.8, we can observe the same behavior for Manhattan (3.8(a)) and Euclidean (3.8(b)) measures. Classifiers may perform better sometimes in one or the other. However, for both of them the classification error usually decreases as the training set and representation sets increase. When the training sets are too small, the errors are far higher than for larger training sets. On the other side, for larger training sets, the classification accuracy sometimes does not differ that much for different sizes (taking the standard deviation into account). There is even a point, where results are better or the same with 90% of the data, than with the full data set.

Let us take a look at the results with the Shape measure (See Figure 3.9). Results have

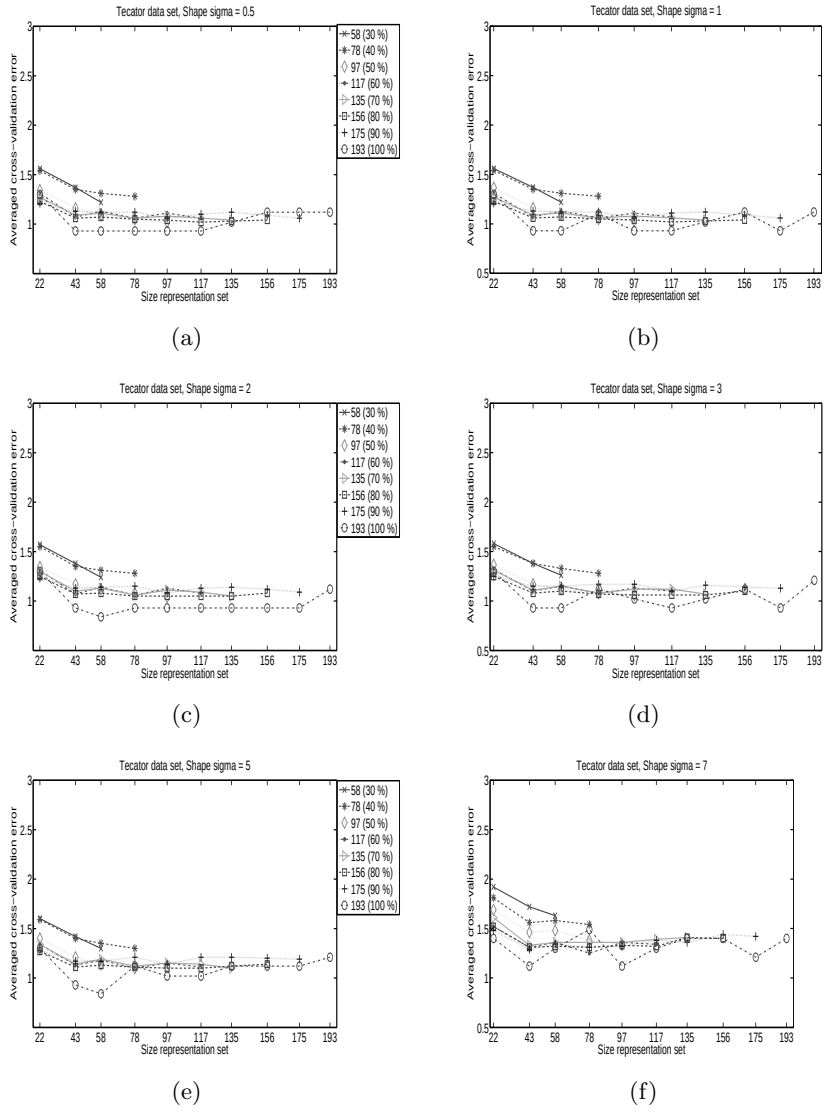


Figure 3.9: Average cross-validation error (in %) for Tecator data set with Shape distance and different values of sigma for (a) sigma=0.5, (b) sigma=1, (c) sigma=2, (d) sigma=3, (e) sigma=5 and (f) sigma=7. The % of data is the size of the selected subset. A training and test set is selected in a 10-fold cross-validation for each subset.

improved with respect to the dissimilarity measures that do not take the shape information into account. Of course, this is not for all values of σ. the optimization of the parameter does influence the results. From Figures 3.9(a) to 3.9(d), we can see that results are pretty much stable for all sizes of the training set. Here, if we take the standard deviation of all repetitions (around 0.5 the highest) into account, there is not much difference between using the 30% of the data and the largest amount of objects. Thus, it seems that with this measure, a small training set can be enough to reach even better results than with the other measures. However, curves are sometimes not very smooth, which could be a result of the random selection of the objects. For the last two values of σ, the results start increasing a bit, it can be due to the data is so smooth, that the measure starts failing. Thus, the importance of the optimization of the parameter.

With respect to the representation set, for all values of σ, the error is always higher with the smallest representation set. It seems that the representation set is not representative enough.

However, from that point on, the errors always start decreasing and keep quite stable until the representation set is the same size as the training set, where sometimes they start increasing again due to the peaking phenomena.

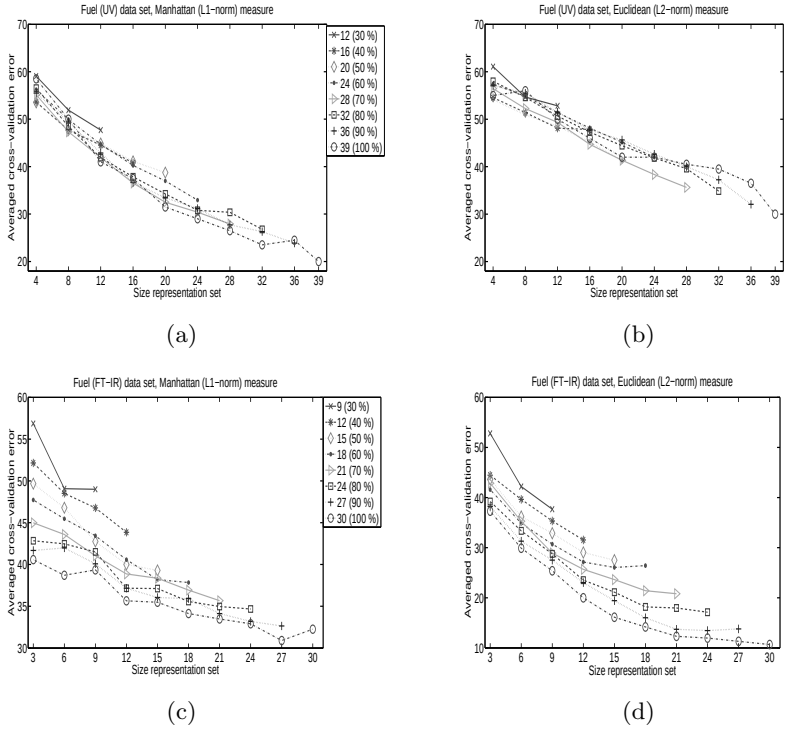


Figure 3.10: Average cross-validation error (in %) with Manhattan and Euclidean measures for Fuel (UV) data set (a) and (b) and Fuel (FT-IR)data set (c) and (d) correspondingly. The % of data is the size of the selected subset. A training and test set is selected in a LOO-fold cross-validation for each subset.

With the two fuel data sets, we are facing a very complicated classification problem: discrimination of special and regular fuel, thereby the classification accuracy is not very good. Moreover, these are both affected by the small sample size problem. In Figure 3.10, we can observe the same phenomena as in the first figure (See Figure 3.8). The errors decreasing while the size of the training set increases. In both cases, the results with the smallest sample size are far much higher; the number of objects is really small, therefore the more objects we add, there will always be a change.

For the Fuel data set from FT-IR (See Figure 3.12), we can see an improvement by using the Shape measure, while this is not the case for the UV data set (See Figure 3.11). This could be determined by the characteristics of the instrumental technique. It seems that the information obtained from the FT-IR spectra is more discriminative than that of UV-VIS. However, due to the small sample size problem, with the smallest training sets the results are still high even for the FT-IR data.

With respect to the representation set, for all values of σ, the error always increases while the training set decreases, errors always start high for the smallest representation set and decrease until reaching the size of the training set, where sometimes they start increasing again due to the peaking phenomena. There is no sufficient data, it would be too much to expect an improvement with little amount of objects. However, it can be observed that results for 70% of the data are very similar to those of the full data set. Again, for the best σ values, the errors for the smaller training sets decrease and errors for larger training sets, have a similar behavior (taking standard deviation).

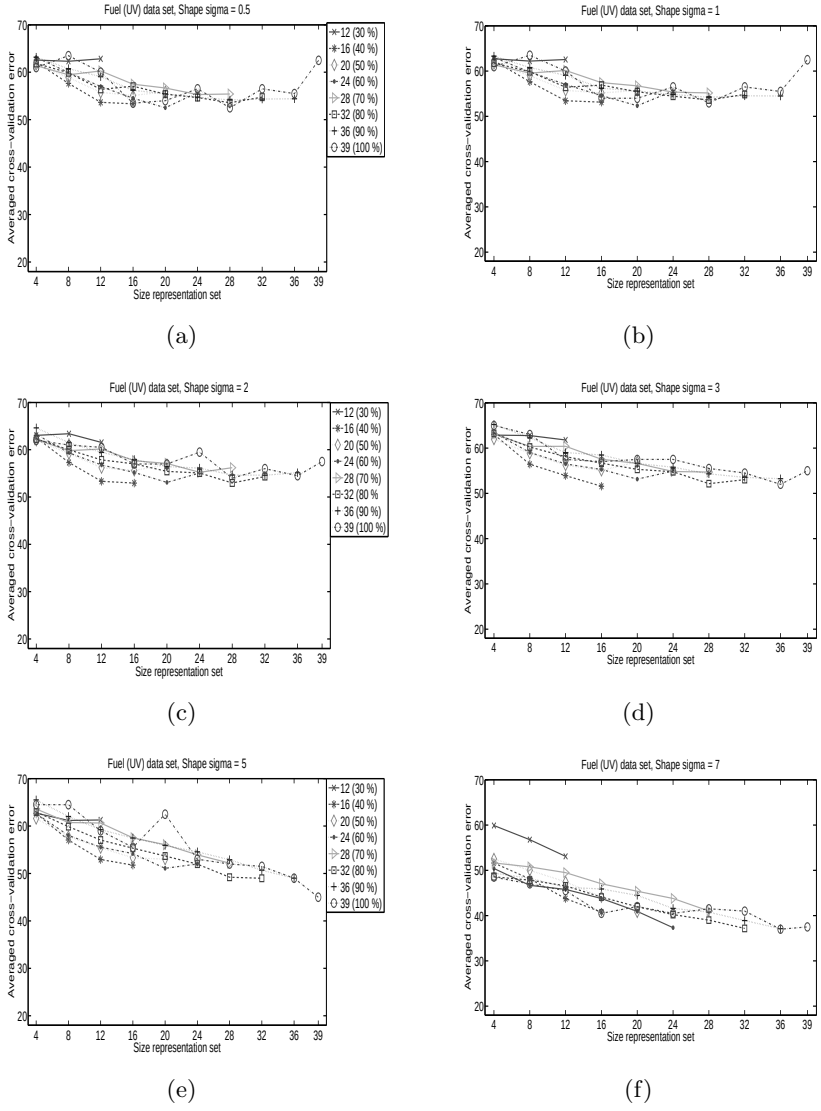


Figure 3.11: Average cross-validation error (in %) for Fuel (UV) data set with Shape distance and different values of sigma for (a) sigma=0.5, (b) sigma=1, (c) sigma=2, (d) sigma=3, (e) sigma=5 and (f) sigma=7. The % of data is the size of the selected subset. A training and test set is selected in a LOO-fold cross-validation for each subset.

The next data set is the three-way Wine data, where objects are represented by high-dimensional 2D matrices. In this case, we also compared the measures that take the shape information into account with does that do not. This is also a small sample size problem, in a high-dimensional space. When analyzing the AMD measure with different values of p, which are the homologous for the Manhattan and Euclidean measures for one-dimensional data, the behavior is similar(See Figure 3.13). Although the learning curves are sometimes a bit rough, we can see how the error decreases meanwhile the size of the training set increases.

In this case, although there is no shape information in both directions, it can be seen that the results also improve (See Figure 3.14) when taking this information into account (in the needed direction). If we take a look at the learning curve for all training set sizes, the best results are for $\sigma = 5$ and $\sigma = 8$. For this type of data, we can also observe, how by including certain discriminative knowledge in the measure i.e. shape, the results improve. With very small sample sizes the errors are still high (the available data is not enough to learn well). But,

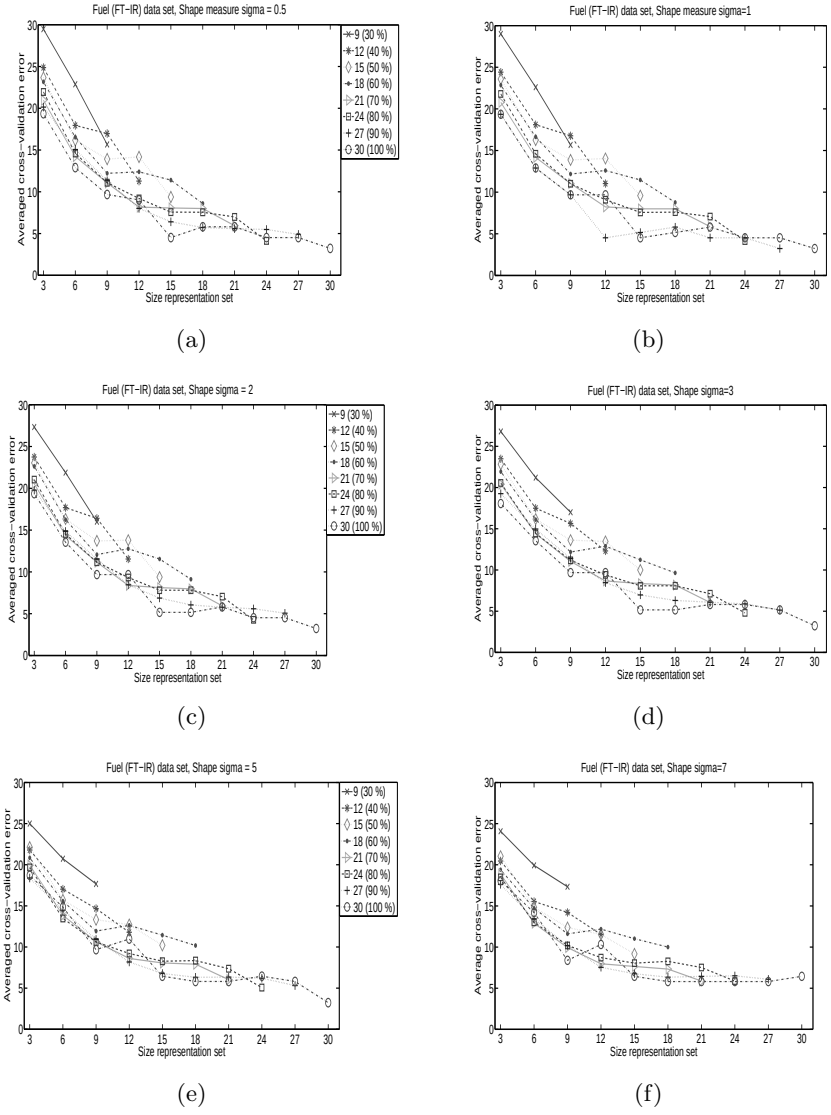


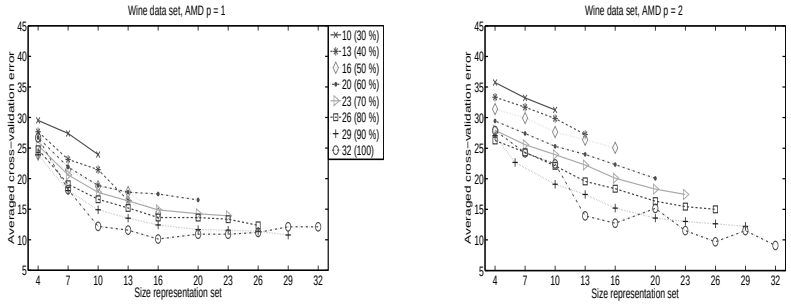
Figure 3.12: Average cross-validation error (in %) for Fuel (FT-IR) data set with Shape distance and different values of sigma for (a) sigma=0.5, (b) sigma=1, (c) sigma=2, (d) sigma=3, (e) sigma=5 and (f) sigma=7. The % of data is the size of the selected subset. A training and test set is selected in a LOO-fold cross-validation for each subset.

when the size of the training sets start increasing, the errors are similar for most sizes (taking standard deviation into account). In which seems to be the best value for σ, the best results are achieved again, with only 70% of the total data set.

3.3.5 Discussion and Conclusions

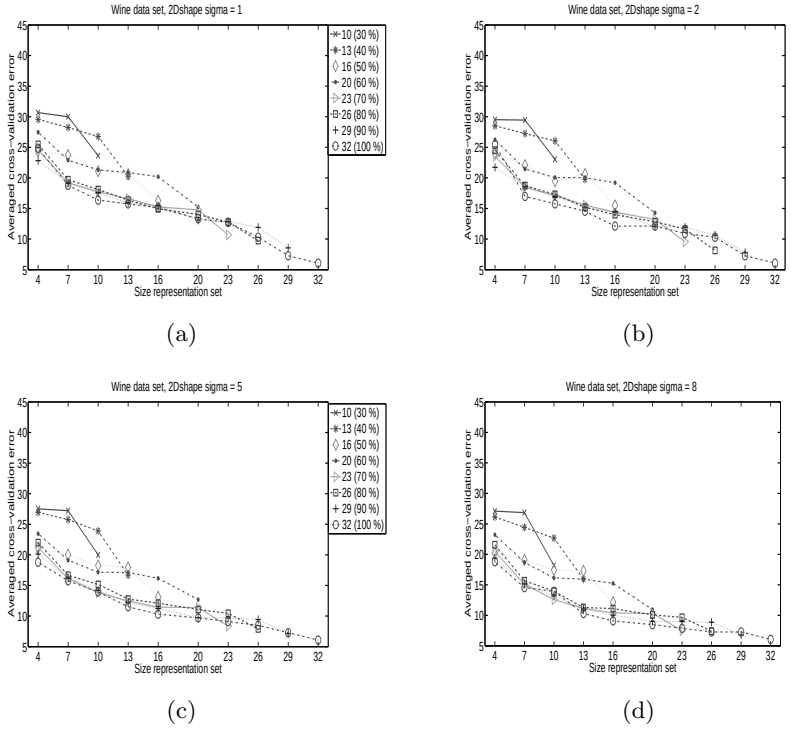
The small sample size problem in high-dimensional spaces is very common in spectral data. Many statistical methods and classifiers fail with this type of data. Alternative representations for such data, to improve classification accuracy, have been explored. Such is the case of the Dissimilarity Representation. However, the key issue of this approach relies on the selection of a suitable dissimilarity measure for the problem at hand. In the case of spectral data, a discriminative characteristic is the knowledge about the connection between the neighboring points and shape.

In our experimental study, we showed the importance of taking the shape of the curve into



(a) (b)

Figure 3.13: Average cross-validation error (in %) for Wine three-way data set with (a) Yang (AMD p=1) and (b) Frobenius (AMD p=2) measures. The % of data is the size of the selected subset. A training and test set is selected in a LOO-fold cross-validation for each subset.



(a) (b)

(c) (d)

Figure 3.14: Average cross-validation error (in %) for Wine three-way data set with Shape distance and different values of sigma for (a) sigma=1, (b) sigma=2, (c) sigma=5 and (d) sigma=8. The % of data is the size of the selected subset. A training and test set is selected in a LOO-fold cross-validation for each subset.

account for the success of the DR. Even when we are facing small sample size problems, if we use the shape information, a few objects are enough for classifiers to learn better. For all data sets, there is some size for the training set (usually smaller than the original data size), from which adding new objects will not make much of a difference. This was also experimented, in a not so small data set, and we reached the same conclusions. In this case, we also benefit from lowering the computational complexity of the classifier. This behavior is not the same for measures that do not consider discriminative information i.e. Manhattan, Euclidean or AMD for 2D data. In this case, the errors are smaller the larger the training set, so we are not solving a small sample size problem, as we would need more objects for the classifiers to learn better.

From the experiments with measures that take shape into account, i.e. Shape and 2Dshape, we can observe the influence of the optimization of the Gaussian filter parameter. There is always a value of σ for which the classification results are better than without measuring shape. It also stabilizes the learning curves of the different sizes of training set, which are around the same performance.

The representation set is also very important. In all experiments we can observe that even with the large data set (Tecator) the error always increases while the training set decreases, with the smallest representation set. It seems that the representation set is not representative enough. However, from that point on, for small data sets the errors always start decreasing, until reaching the size of the training set, where sometimes they can start increasing again due to the peaking phenomena. There are not sufficient objects to make the curves more stable. These experiments are all based on two-class classification problems; for multi-class problems, further studies should be done.

In conclusion, the incorporation of shape information in the dissimilarity representation is important for the discrimination of spectral data. It helps avoiding the curse of dimensionality problem, allowing classifiers to perform well in small sample size situations.

3.3.6 Acknowledgment

We acknowledge financial support from the FET programme within the EU FP7, under the project "Similarity-based Pattern Analysis and Recognition - SIMBAD" (contract 213250). We would also like to thank to the project Cálculo científico para caracterización e identificación en problemas dinámicos (code Hermes 10722) granted by Universidad Nacional de Colombia.

3.4 Optimizing Dissimilarities for the Classification of Three-way Chemical Spectral Data

This section is currently under second review after major revision as 'Optimizing Dissimilarities for the Classification of Three-way Chemical Spectral Data', by D. Porro-Muñoz, R.P.W Duin, I. Talavera, M. Orozco-Alzate and J. A. Pino-Alea, in *Chemometrics and Intelligent Laboratory Systems*. The notation was modified in order to have a unified notation in the whole thesis.

Abstract

Multi-way analysis of spectral data is gaining an increasing interest in chemometrical analysis. This representation aims at a more informative description of objects, allowing for a better interpretation of data and often resulting into better estimations. However, the continuous nature of spectra is hardly taken into account in their analysis, as it is usually based on vectors or matrices of individual features. Discriminative characteristics of spectra like their shape are thereby not considered. An alternative representation, the Dissimilarity Representation (DR), has been recently extended for the classification of multi-way spectral data, as a potential solution to this problem. Nonetheless, results can still be influenced by noisy and/or redundant information. In this paper, we investigate a modification of the DR based on selecting just a part of the spectral data (the peaks) for computing the dissimilarities. The mutual information of the detected peaks with the class labels is used as a criterion for selection. Results show that a reduction of the spectral peaks is possible without deteriorating the classification accuracy. Moreover, the computational cost to obtain the DR is reduced.

3.4.1 Introduction

The number of methods developed to date for multivariate analysis is very large. However, they are often designed for data represented in a two-way i.e. $objects \times features$ or matrix structure, e.g. a set of spectra, which are traditionally sampled and represented as vectors of individual observations. For some time now, potentially more informative descriptions of objects are considered. Examples are data obtained via hyphenated instruments like Gas-Chromatography combined with Mass Spectrometry (GC/MS). With this technique the molecules of each mixture are separated and they elute at different times (chromatogram); each molecule is then fragmented by the mass spectrometer and the ionized mass fragments are turned into an electrical signal (spectrum) [73]. Measurements obtained by these equipments do not fit in a single vector anymore, they should be then represented by a matrix $m \times l$ (or higher order arrays) where e.g. m is the number of elution time points and l the number of mass fragments. If all objects are gathered, multi-dimensional arrays are obtained i.e. $objects \times features1 \times features2$ [125, 73] (each direction is usually referred to as mode/direction/way) [60] and analyzed with tools that are capable of using this structure.

Fields like chemometrics [126], neuroinformatics [4], are very reach in instruments that provide this type of data sets, thereby multi-way data analysis [64, 126, 4] has become one of the main research topics in many areas. The introduced methods e.g. CANDECOMP/PARAFAC [53], N-PLS [126], make use of the complex multi-way structure of this type of data, allowing for its better interpretation and usually leading to better results. However, most efforts have been focused on the development of regression and exploratory analysis methods. Not much has been done in supervised classification.

For this task, it is sometimes common to unfold the multi-way array into a matrix and apply multivariate classifiers [115, 12, 11]. However, with this procedure, the multi-way structure is ignored and poor results could be obtained. The predominating procedures are based on modeling the data with a multi-way factor model e.g. PARAFAC, TUCKER3, with which dimensionality is also reduced by projecting objects on the lower dimensionality factor space (although it can be problematic with many irrelevant features). Afterwards, a traditional multivariate classification method is applied on the obtained scores [52, 62, 110]. Another procedure is the extension of the multivariate PLS-DA method to multi-way data, i.e. NPLS-DA [51, 11] and the recently proposed multi-way classification method : NSIMCA [37]. Like in the traditional SIMCA [154] method, a decomposition model is computed for each class before classification rules are built. This procedure tackles the problem of class modeling, as in the two-way version, for multi-way array data structure. It is still based, however, on a numerical analysis of the

individual data samples. Other characteristics of the nature of data that could be helpful for their better discrimination are not used. Such is the case for example of multi-way spectral data i.e. multi-way arrays provided by hyphenated instruments like GC/MS, as explained at the beginning of this section. In spite of the continuous functional nature of spectra, they are traditionally represented by sampling and are analyzed as a sequence of individual observations made on the objects.Therefore, the shape of the curve, which could be the main characteristic to discriminate objects by their spectrum is ignored.

In recent articles, a new approach that may contribute to this issue was proposed. It is the extension of the Dissimilarity Representation (DR) approach [94] for multi-way data [100], focused on spectral data. The DR is based on the important role that dissimilarities between objects play for discriminating between different classes (groups) of these objects. It proposes to represent objects by their dissimilarities to the most representative objects of each class. In this way, the geometry and the structure of a class are determined by the user defined dissimilarity measure, in which application background information may be expressed e.g. continuous nature of spectra. Classifiers may be built in this dissimilarity space. One of the main issues in chemometrics, the small number of objects in high-dimensional spaces, can also be tackled with the DR. It all depends on the selection of a suitable dissimilarity measure for the problem at hand. If the discriminative aspects of the data can be measured, a few objects should be enough for classifiers to learn well. Hence, non-linearly separable problems in the high-dimensional feature space can be turned into linear problems in the dissimilarity space [94].

A number of 2D (or second order) dissimilarity measures have been proposed to compare objects represented by matrices [78, 161]; a few can be found to compare 2D spectral data [34]. However, these measures do not take the nature of spectra into account. Recently, we proposed the 2Dshape measure [100] for this purpose. With this measure, objects are analyzed in their 2D structure and the information on their continuous nature, e.g. shape of the curves, is also taken into account. It is to be expected that by including the shape information in the representation of spectra, the performance of classifiers improves.

Nonetheless, there is still another issue to be aware of. In this kind of data, it is very common to have non-informative, noisy or redundant information. For example, in GC/MS systems, there are different sources of background noise e.g. contaminants in the ionisation chamber. Moreover, the majority of mass fragments are typically of no significance. Similar problems are found in other types of spectral data. This non-contributory information is usually counter-productive for classifiers. In the best case, when it does not influence their performance, the computational complexity increases. Thereby, it is very common to make a selection of the most contributive features for the modeling of the problem at hand.

Following up this idea, we wonder if the DR could benefit from performing a data reduction by feature selection prior to its computation. As it is now, the DR is based on a global comparison of spectra that enables to consider shape. Some studies for feature selection in multi-way data have been done. However, most alternatives are based on optimizing multi-way decomposition methods [155, 31, 7], they are dependent on them. Thus, they cannot be easily extended for other procedures. In this paper, we investigate what happens if the DR is just computed on selected parts (peaks) of the spectra. It will for sure diminish the computational cost. However, there still remains the question whether we are able to keep the shape information and compute proper dissimilarities.

With this purpose, we introduce a procedure to select discriminative features in both feature directions of a three-way spectral data set.

First, in each direction the full peaks are detected and isolated, such that the shapes of the peaks are respected when computing the DR. Next, the detected peaks are ranked according to their Mutual Information (MI) with the class variable. Finally, a DR is computed for the selected number of peaks, using shape information.

A comparison, in terms of classification, is carried out between the previously introduced DR for multi-way spectral data on the full data and on the reduced version according to the proposed procedure. Traditional approaches for classifying this type of data are also included in the comparison. Two types of three-way chemical spectral data, from different instrumental analysis techniques, are used. Both of them are of public domain. This comparison aims at showing the importance of taking the shape information of spectra into account, for their better discrimination. We will demonstrate the feasibility of a feature selection (both for improving results as well as for reducing the computational complexity), even when the representation is more consistent with its nature.

The paper is organized as follows. In Section 3.4.2, a brief introduction to the DR approach and its extension to multi-way data is done. The 2D Measure for the application of the DR in three-way spectral data is also presented. The introduced feature selection modification for multi-way data is explained in Section 3.4.3. Section 3.4.4 is dedicated to detail the specifications about the materials and methods applied in the experimental section. Following, the experimental results are presented and discussed. Finally, our conclusions are presented.

3.4.2 Dissimilarity Representation for Multi-way data

The Dissimilarity Representation [93, 94] was originally proposed as a more flexible representation of the objects than the traditional feature-based one. In this approach, new features are defined for the objects, such that they are represented by their proximities to a set of representative objects of each class. The fact (or property) that dissimilarities should be smaller for similar objects (the same class) and larger for different objects, suggests that they could be used as more discriminative features due to their crucial role in the class constitution. It aims at including more information about the characteristics and structure of the objects through a suitable (dis)similarity measure e.g. shape of spectra.

Let us define the Dissimilarity Space (DS) approach, given a t-way array $\underline{Y} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_t}$ where each object is represented by a (t-1)-dimensional array, a representation set $\underline{R}(\underline{R_1}, \ldots, \underline{R_h})$, where h is the number of prototypes, and a dissimilarity measure d [94, 100]. A mapping $\phi(\cdot, \underline{R}) : \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_{t-1}} \rightarrow \mathbb{R}^h$ is done, such that every object $\phi(\underline{Y_i}, \underline{R}) = [d(\underline{Y_i}, \underline{R_1}), d(\underline{Y_i}, \underline{R_2}), \ldots, d(\underline{Y_i}, \underline{R_h})]$ is associated by its dissimilarities to all objects in $\underline{R}$. Hence, a dissimilarity matrix $\mathbf{D}(\underline{Y}, \underline{R})$ is obtained, which is used to build a classifier in the correspondent dissimilarity space of dimension h, $/\mathbf{D} \subseteq \Re^h$. The prototypes are usually the most representative objects of each class, $\underline{R} \subseteq \underline{Y}$ or $\underline{Y}$ itself, resulting in a square dissimilarity matrix $\mathbf{D}(\underline{Y}, \underline{Y})$. Any traditional classifier can be built in the dissimilarity space as in the feature space.

Consequently, the relationships between all objects in the training and representation sets are used for classification. If a suitable measure is chosen, the compactness property (objects from the same class should be similar and objects from different classes should be different) of the classes should be more pronounced. Therefore, it should be easier for the classifiers to discriminate between them, such that linear classifiers in dissimilarity space may correspond to non-linear ones in the feature space. In general, any arbitrary classifier operating on features can be used [94].

2DShape measure for spectral data

As it was mentioned before, although spectra are continuous single entities, they are traditionally represented as a set of independent features. Therefore, discriminative information from their continuous nature, such as shape, is not taken int account.

A dissimilarity measure for three-way spectral data [100] was designed with this purpose. It takes into account the information contained in both feature directions. Additionally, it is defined in such a way that the differences in the shape of the surfaces (2D spectra) from the several classes can be measured. Assume that we have a three-way array $\underline{Y} \in \mathbb{R}^{n \times m \times l}$, where n

is the number of objects, and m and l are the number of features (related to a single object) in each of the other directions respectively; $\forall j = 1, 2, \ldots, m$ and $k = 1, 2, \ldots, l$. Then, the 2DShape dissimilarity measure, an extension of the Shape measure for simple spectra [88], was defined as follows:

1. Compute the matrix D^1

$$D^1_{a,b} = \left(\sum_{k=1}^{l} \left(\sum_{j=1}^{m} (\mathbf{Y}^{\sigma_1}_{a,j,k} - \mathbf{Y}^{\sigma_1}_{b,j,k})^2 \right)^{p_1/2} \right)^{1/p_1},$$

$$\mathbf{Y}^{\sigma_1}_{i,j,\cdot} = \frac{\mathrm{d}}{\mathrm{d}_j} G(j, \sigma_1) * \mathbf{Y}_{i,j,\cdot}$$

2. Compute the matrix D^2

$$D^2_{a,b} = \left(\sum_{j=1}^{m} \left(\sum_{k=1}^{l} (\mathbf{Y}^{\sigma_2}_{a,j,k} - \mathbf{Y}^{\sigma_2}_{b,j,k})^2 \right)^{p_2/2} \right)^{1/p_2},$$

$$\mathbf{Y}^{\sigma_2}_{i,\cdot,k} = \frac{\mathrm{d}}{\mathrm{d}_k} G(k, \sigma_2) * \mathbf{Y}_{i,\cdot,k}$$

3. Combine both dissimilarity matrices $D = \alpha_1 D^1 + \alpha_2 D^2$

The variables $\mathbf{Y}_{i,j,\cdot}$ and $\mathbf{Y}_{i,\cdot,k}$, stand for the k-th column and the j-th row of the i-th matrix (object); $\forall i = 1, 2, \ldots, n$. Their expressions correspond to the computation of the first Gaussian (that is what G stands for) derivatives of spectra. Thus, a smoothing (blurring) is done by a convolution process ($*$) with a Gaussian filter and σ stands for a smoothing parameter. With the application of the blurring, problems like peak shifting in chromatography could be tackled. For some values of σ, spectra can be sufficiently smoothed such that this shifting is not significant, and they can be compared. The dissimilarities in steps 1 and 2 correspond to the first and second directions respectively, as indicated by the notation e.g. spectra and time. This measure can also be used in three-way data where there is no continuity in one of the directions i.e. no variations in shape. In this case, it is enough with just computing the point-wise difference in that direction.

The power p is used to emphasize either small or large differences between the elements, in dependence of the problem at hand. If $p < 1$, all the differences are reduced, thus the larger ones do not interfere much in the measure. On the other hand, if $p > 1$, the larger differences will be more pronounced, resulting in a heavy influence on the measure.

3.4.3 Feature selection in three-way data

There are several approaches for feature selection in traditional multivariate data analysis. However, this is not the case for multi-dimensional data arrays. Our approach is based on defining the influence of the features from each direction (taking into account the information of the other direction) in the discrimination between classes. Hence, the first step of the proposed approach is to slice the three-way array, either in the second or the third mode.

Let us assume we are slicing in the second direction (See Figure 3.15 (a)). We will obtain a matrix $\mathbf{Y_j} \in \mathbb{R}^{n \times l}$, such that each object is represented by a vector like $y_{ij} = [y_{ij1}, y_{ij2}, ..., y_{ijl}]$, for $i = 1, 2, \ldots, n$. Taking into account the information in the representation of the objects in the $\mathbf{Y_j}$ slide (based on the behavior of the k features, $k = 1, 2, \ldots, l$, in this slide), we will find a

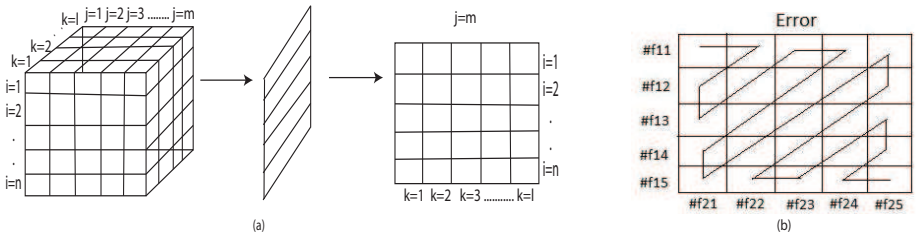


Figure 3.15: (a) Slicing of the cube to compute weights for each j feature (b) Zig-zag scan for the global minimum search in both directions

score (discriminative power) of $\mathbf{Y_j}$. The same procedure is applied to find the score of each $\mathbf{Y_k}$ slide, based on the information of the features in the second direction. Following this criterion, we can find two cases for feature selection in this type of data sets:

- Features in the analyzed direction are not continuous. Therefore, we can find the scores for each variable independently.
- Features in the analyzed direction have a continuous nature. In this case it is of our interest to maintain this continuity in the selection process in order to keep the discriminative information (the shape of the signal) in the later analysis of the data.

In the next subsections we will explain the feature selection procedure for each of these cases.

Selection of non-continuous features

For non-continuous data e.g. mass spectra, features can be scored independently, as there is no continuity/ordering in the measurements. Assuming we have a $\mathbf{Y_j}$ slide as explained in the previous section, there is a scoring function f_j associated to $\mathbf{Y_j}$, such that a score w_{jk} is assigned to each k feature, with $k = 1, 2, \ldots, l$, for the selected j-th feature:

$$f_j : y_{\cdot jk} \mapsto w_{jk} \tag{3.4}$$

where $y_{\cdot jk}$ is the vector of values of the k-th feature for all objects in slice Y_j. Afterwards, all w_{jk} scores are combined such that a final score w_j of the $\mathbf{Y_j}$ slide is obtained, according to its overall behavior in the features of the third direction:

$$w_j = \frac{\sum_{k=1}^{l} w_{jk}}{l} \tag{3.5}$$

If the nature of the third direction was also non-continuous, the same steps should be carried out to obtain a score w_k for the analyzed $\mathbf{Y_k}$ slide, based on the information of all features from the second direction.

Selection of continuous features

There are different measures to analyze the contribution of features to the classification process. However, they are usually designed such that the discriminatory power of each feature is analyzed independently. Consequently, for spectral data for example, there is the risk of selecting the most important features of each peak, but the shape of the curve is completely lost. This can affect the results as taking the shape into account would not make sense anymore.

Therefore, we propose to select full peaks instead of independent features in this case. With this purpose, we first make an average spectrum or chromatogram from each data set in order

to detect the peaks. What makes a peak is the fact that there is a high point with lower points around it. So, a point t is considered a maximum of a peak if for every point q in a neighborhood of t, $t - q :> \Delta$ holds [20]. The Δ value is selected by visual inspection; it is a difference between a maximum and its surrounding, and is used in order to ignore small peaks or noise in the signal. In the case of chromatographic signals, problems like peak shifts in the different signals may affect the peak detection process, therefore a previous alignment of peaks might be needed [128].

After peaks are detected, the selection procedure can be applied as follows. Assume we are selecting in the second direction and it has a continuous nature e.g. Ultraviolet spectra. Instead of having just a slide, this time we will take a portion of the cube $\underline{Y_p} \in \mathbb{R}^{n \times s \times l}$, which corresponds to a p-th peak and it is defined as the set of $\mathbf{Y_j}$ slides of the s features that conform the analyzed peak. In this case, we are not interested in just scoring an individual j-th feature i.e. $\mathbf{Y_j}$ slide, but the peaks found in the spectra. In order to compute the score for $\underline{Y_p}$ we have a scoring function g_p associated to it, such that a score w_{pk} is assigned to each feature k, with $k = 1, 2, ...l$, for the p-th peak:

$$g_p : Y_{\cdot pk} \mapsto w_{pk} \tag{3.6}$$

where $Y_{\cdot pk}$ is the matrix of values of the k-th feature for all objects in the sub-cube $\underline{Y_p}$. Afterwards, the scores of all k features are combined as in Equation 3.5, such that a final score w_p of the p-th peak is obtained, according to its overall behavior in the k features:

$$w_p = \frac{\sum_{k=1}^{l} w_{pk}}{l} \tag{3.7}$$

This procedure is repeated for all peaks in the selection direction. If the third direction would also have a continuous nature, the same steps should be followed to find the scores of the peaks in that direction.

Mutual Information on selected features

The Mutual Information (MI) [122] is a measure of predictability of one variable given the information of another. Let us assume that we have a variable $Z = [z_1, z_2, \ldots, z_n]$ and the target class $C = [c_1, c_2, \ldots, c_n]$. The MI represents how much uncertainty of C is lost if Z is known, an it is defined as:

$$I(C; Z) = \sum_{c \in C} \sum_{z \in Z} p\,(c, z) \log \frac{p\,(c, z)}{p\,(c)\,p\,(z)} \tag{3.8}$$

The MI is only zero when the analyzed variables are independent. If its value is large, it means that the known variable and the target class C are correlated. Therefore, it could be a good measure of the relevance of a feature for the class prediction [95, 114].

In our case, we also want to compute the relevance of a whole peak (set of features) for continuous data. The MI between a set of features $\mathcal{Z} = \{Z_1, Z_2, \ldots, Z_t\}$ (t is the number of features in the peak) and C can be then computed as [95]:

$$ID = \frac{1}{m} \sum_{Z \in \mathcal{Z}} I\,(C; Z) \tag{3.9}$$

In this paper, we use I as the f_j function to find the scores w_{jk} (see Equation 3.4) and ID as the scoring function g_p to compute the scores w_{pk} (see Equation 3.6.

Selecting in the dissimilarity measure

The computed scores should contain (depends on the used function) the information about the discriminative power of the features, according to their overall behavior. Therefore, we assume that a high score is indicative of a valuable feature. After scores for features/peaks in both directions are computed, they are sorted in a descending order. Based on this ranking, we propose to do a binary weighting, such that only those features with score values above a defined threshold (most significant features) will be included in the computation of the dissimilarity matrix, that is:

$$\delta_k = \begin{cases} 1, & \text{if } w_k > T_k \\ 0, & \text{otherwise} \end{cases}$$

$$\delta_j = \begin{cases} 1, & \text{if } w_j > T_j \\ 0, & \text{otherwise} \end{cases}$$

where T_k and T_j are the defined thresholds for selecting in each direction. In case of selecting peaks, if we were selecting in the second direction for example, all δ_j corresponding to the s features of the same peak will have the same value, as they all have the same score w_p. We can incorporate the binary weights into the computation of the dissimilarity measure as follows:

1. Compute matrices D^1 and D^2

$$D^1_{a,b} = \left(\sum_{k=1}^{l} \left(\delta_k \sum_{j=1}^{m} (\mathbf{Y}^{\sigma_1}_{a,j,k} - \mathbf{Y}^{\sigma_1}_{b,j,k})^2 \right)^{p_1/2} \right)^{1/p_1},$$

$$D^2_{a,b} = \left(\sum_{j=1}^{m} \left(\delta_j \sum_{k=1}^{l} (\mathbf{Y}^{\sigma_2}_{a,j,k} - \mathbf{Y}^{\sigma_2}_{b,j,k})^2 \right)^{p_2/2} \right)^{1/p_2}$$

The rest is maintained as in the previous definition. In this way, the features with low discriminatory power have no impact in the dissimilarities between spectra. Moreover, the cost of computing the dissimilarities related to these features will be avoided.

There are many feature selection methods in the literature, such that those most significant features are selected. However, the selection criterion is usually based on the minimization of the classification error by an exhaustive search of the best combination of features. This problem, for simple multivariate data is computationally complex (NP-Hard) and a heuristic is usually needed to find a solution. For three-way data, it is even worse. Therefore, we decided to use a more simple and intuitive way for selecting the features. For each direction, different feature sizes from the total number of features in the analyzed direction are fixed e.g. $10\%, 20\%, \ldots, 80\%$ of the features. Hence, experiments are conducted on the obtained feature sets for the specified number of features to be selected. Notice that according to this, the threshold T can be seen as the value for which the given percentage of features scorred higher than it. In order to select features in both directions at the same time, the combinations of the different feature sets are evaluated, and we select the best combination based the minimum classification error.

Depending on the problem at hand and the specialist knowledge, the feature selection could be done in only one of the two directions or in both of them. This procedure can be generalized to other models or types of data sets.

Selection of the number of features

In the selection process, the intuitive idea is to find the global minimum of the error function. As the features are ranked according to their significance, this minimum should indicate the optimal set of features. Nonetheless, when new features with redundant information are added, this global minimum could be found in one or more evaluations of the function. The information of the new features does not influence the results (only the computational cost).

If the selection is done through only one of the directions, the idea is to find the first minimum. However, applying this procedure for the selection in two directions is not that straightforward. In this case, the first minimum will depend on which direction the analysis starts, i.e. each feature in the first direction with all combinations of the second direction and vice versa. Hence, suppose we have an error matrix E. Each cell of E contains the classification error for the combination of a number of features in each of the two directions. We propose to do the selection by moving from the first combination, in a zig-zag scan, through the diagonals (See Figure 3.15 (b)). With this procedure, the fraction of analyzed features is similar in both modes. It goes from small to large amounts of features in the two directions. This is not the case for the method (common) previously mentioned. There is also the fact that the selection will depend on where the zig-zag scan starts i.e. from the first combination to the right or down. In any of the two methods, the specialist could select which direction is better to reduce for further analysis or decreasing the computational cost. However, if there is no prior knowledge about which direction should be preferably reduced, the proposed procedure is closer to the desired one.

There is another issue to be tackled in the selection. Although there could be only one global minimum, it may happen that for other combinations of features, very close values to that minimum are obtained. In classification problems, it is very common to find noisy objects that are on the border of two or more classes. Thus, according to the selected set of features, these objects may be classified differently. It also depends on the training sets e.g. one object is wrongly classified in two of ten folds in the cross-validation procedure. The error values for the different combinations of features may then just vary because of these few objects. In this case, the global minimum might not be the best option. It could increase the computational cost (number of selected features) depending on just the shape of those noisy examples. Therefore, it would be worth to have an error margin for the selection of the optimal combination of features and make an error/cost trade-off. That is: in a minimum search algorithm, a variable min is initialized with the first value of the array (vector or matrix). All values are then analyzed, and min is only updated when a value a lower than it is found i.e. $min = a$ if $a < min$. We propose to include a value e, such that we have an error margin related to an admissible number of wrongly classified objects e.g. two errors in 10 folds of a cross-validation. The selection of the global minimum will be defined then as: $min = a$ if $min - a \geq e$, e can be calculated as the percentage that the admitted number of classification errors represent from the total number of objects.

3.4.4 Materials and methods

Experiments are conducted on two chemical spectral public-domain data sets (from different instrumental analysis techniques). First, traditional approaches for classifying multi-way chemical spectral data are analyzed. They do not consider the shape information of spectra i.e. Unfolding + classification (UC), Multi-way decomposition model (PARAFAC, TUCKER) + classification (MDC), NPLS-DA [11] and the recently introduced NSIMCA method [37]. All these approaches will be compared with the previously introduced DR with 2D measure (DR2D) and the different variants of the proposed feature selection modification. The DR will also be computed on the unfolded data (UDC) to show the benefit of using shape.

Data sets

The first set consists of metabolite data containing High-Performance Liquid Chromatography (HPLC) measurements of commercial extract of St. John's wort. The following description has been taken from the literature [3] for a better understanding of this paper. HPLC-PDA (HPLC with photo-diodo array detection) profiles were obtained from 24 different examples of St. Johns' wort from several continents (classes): Africa (8 objects), Asia (6 objects), Europe (45 objects) and North America (30 objects). The number of examples from each continent varies between 2 and 12. HPLC-PD profiles replicates of three or four hour for each example were obtained. The chromatography was monitored between $190nm$ and $620nm$. Two regions of the chromatographic data were chosen for the analysis and reduced to steps of $3nm$ in the Ultraviolet-mode (UV) $(260 - 550nm)$ (97 features) and steps of 1.32 seconds in the retention time mode (549 features). The final three-way data has a size of $89 \times 97 \times 549$.

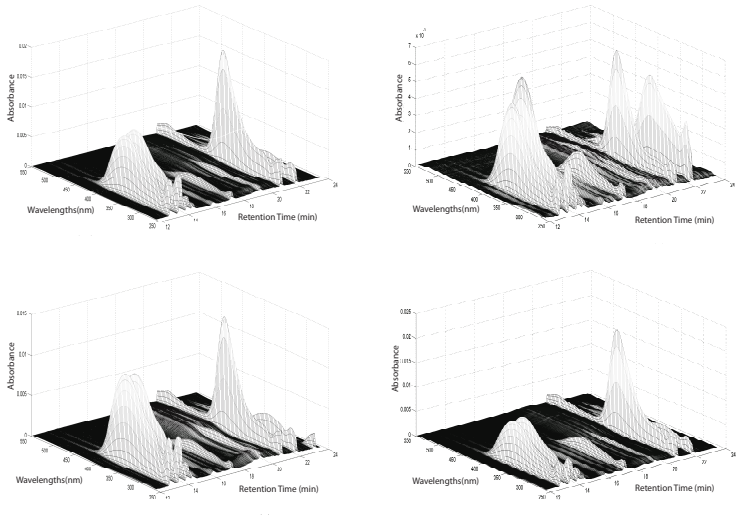


Figure 3.16: Average pattern for each class of St. John's data set (top-left) Africa, (top-right) Asia, (bottom-left) Europe and (bottom-right) North America.

The second data set consists of examples of red wine [124, 152], produced from the same grape (Cabernet Sauvignon) and belonging to different geographical areas and producers. They were collected from local supermarkets and analyzed by means of HS-GCMS (headspace gas chromatography/mass spectrometry). Separation of aroma compounds was carried out on a gas chromatography system (2700 columns from the scans of chromatographic profile). For each example, a mass spectrum scan (m/z : 5204) measured at the 2700 elution time points was obtained, providing a data cube of size $44 \times 2700 \times 200$ i.e. examples (objects) in first direction, elution time points in second direction and mass spectrum in third direction. The data set is composed of examples from three different geographical areas: South America (21 objects), Australia (12 objects) and South Africa (11 objects).

Software and optimization

The experiments were all performed in Matlab. For the case of N-way decomposition methods, the N-way toolbox [145] and a Parafac2 routine were used [147]. The PRTools toolbox [149]

was used for the DR and classification of the data. The Mutual Information toolbox was used for the computation of the MI. The experiments have been designed as follows:

1. For the two data sets, a double cross-validation procedure was performed such that all objects are used for training and test at some moment. In the first k-fold cross-validation, the training set of each fold is used to optimize the model (classifier) through and internal (LOO) cross-validation, and this model is then evaluated on the test set of the same fold. Performances are evaluated in terms of the Average Classification Error (ACE) from the k-folds, and the standard deviation is taken into account.

2. For the cases where a traditional classifier is used in a second step, two classifiers were used: Regularized Linear Discriminant Classifier (RLDC) and Support Vector Machine (SVM). To solve the multi-class problems, the one-versus-all classification scheme is applied. For the SVM classifier, the RBF kernel was applied in the two data sets. We want to show how linear classifiers can be sufficient on the DR when a suitable dissimilarity measure is used, not being the case for the other representations. The optimal kernel parameter and regularization parameter C were optimized as it was previously explained. The Linear Discriminant Classifier (LDC) assumes that the classes are described by multi-normal distributions with the same covariance matrices. Since for $n \times n$ dissimilarity representations the estimated covariance matrix S is singular, its inverse cannot be determined. Therefore, its regularized version is used instead (RLDC). Regularization takes care that the inverse operation is possible by emphasizing the diagonal values (variances) of the matrix S with reference to the off-diagonal elements (covariances) [89, 94]. To find the regularization parameters of RLDC, an automatic regularization (optimization over training set by cross-validation) process was done as explained in the previous item.

3. In the unfolding procedure, a set of chromatograms is obtained for both data sets by averaging in the spectral dimension for all objects.

4. In the MDC approach, the used methods and parameters are the ones reported in the literature: TUCKER3 for St. John's data [3] set and PARAFAC2 for Wine data set [8]. The number of factors of the NPLS-DA method were optimized in a LOO cross-validation in the training sets of each cross-validation fold for the classification. For the NSIMCA classifier, an alternative NSIMCA implementation was done. The number of factors for each class was determined in a k-fold cross-validation procedure by evaluating the estimation error in the decomposition methods. The decomposition methods used in NSIMCA are the same used for the MDC approach.

5. For the 2D measure, parameters were optimized in a grid-search procedure. Three versions of feature selection application are shown: 2D measure selecting in the first (FD) or second dimension (SD) only, and 2D measure selecting on both directions (BD). For each data set, different numbers of features are selected in a range of $[10\% - 80\%]$ of the data in steps of 10. As we are selecting peaks instead of independent features, the number of selected features does not correspond directly to the specified percentage. Peaks are concatenated until the total number of features (that comprise those peaks) correspond to that percentage or higher (never less), as the peaks will be fully taken into account. They cannot be cut such that the number of features correspond to the specified percentage, otherwise the shape and/or information of the peaks is lost. Thus, the final features are the total number of features from the selected peaks. If there are big peaks, it may happen that for different percentages, the same number of features is obtained. This number will be the one shown in the results. For the first step in the feature selection, only the folds from the first repetition of the 10-fold cross-validation are used. It is based on the evaluation of the classification error on the training sets.

The parameters used in each of the previously described approaches are summarized in Table 3.2. The ♯-folds column refers to the number of folds used to split the data into training and test for the classification.

Table 3.2: Optimized parameters for each of the compared procedures for the three data sets

Data Sets	Unfold dimensions	MDC [no. fact.]	NPLS-DA	NSIMCA [no. fact./class]	DR2D	♯-folds
St John's	89×549	TUCKER3 [4,3,5]	[10]	[4, 3, 5],[2, 3, 5] [2, 3, 5],[4, 3, 5]	$\sigma_1 = 2$,$\sigma_2 = 3$, $p_1 = 2$,$p2 = 1$	6
Wine	44×2700	PARAFAC2 [3]	[5]	[4, 5, 5]	$\sigma_1 = 5$, $p1 = 1$,$p2 = 0.5$	10

As data from the GS technique might have shifted peaks, in the MDC a PARAFAC2 model was applied for the GC-MS combined data sets. This method is an extension of the PARAFAC model, which can tackle this kind of problems. For the GC-MS technique (Wine), in the chromatography direction (2nd) we will have the eluded peaks for all the components present in the substances. In the spectral mode (third direction, mass spectra for GC-MS), each eluded component will only have one mass spectrum (not continuous mass fragments in which the molecule decomposes). In mixtures, the main difference between classes is given by the relation between the eluded components in the chromatogram i.e. how the concentration of one of the peaks varies with respect to the others, for the different classes. In this case, it is important to take shape into account, because there is information in the ordering of the components (peaks) with different concentrations and also continuity. Thus, for the Wine data set and GC-MS data in general, we propose to adapt the 2DShape measure (Section 3.4.2), which takes the information of both directions into account, to the specificities of the data. When computing the D^1 matrix for the chromatography direction, we will use the Gaussian derivatives to take into account the shape in the changes of concentration in the neighboring components. However, for the D^2 matrix from the mass spectra mode, the use of derivatives is meaningless, because there is no continuity between the mass fragments; just the differences between the concentration of the mass fragments will be computed.

3.4.5 Results and discussion

The classification results (ACE) for all data sets and the different approaches are shown in the following table:

Table 3.3: Averaged cross-validation error in % (with standard deviation) for the two data sets for different classifiers and three-way classification approaches.

Representations/Classifiers	St John's Wort	Wine
U(Chrom)-RLDA	7(0.2)	28.1 (0.5)
U(Chrom)-SVM	3.9 (0.2)	25.5 (0.4)
MDC-RLDA	24.4 (0.5)	43.2 (0.8)
MDC-SVM	4.5 (0.8)	35.6(1.2)
NPLS-LDA	5.6 (1.1)	25 (1.5)
NSIMCA	4.5 (0.2)	44.3 (0.6)
DR2D-RLDA	***2.3*** (0.1)	***12***(0.2)

In this table, it should be noticed that, the multi-way classification methods i.e. NPLS-DA and NSIMCA outperform the unfolding procedure (UC) with linear classifier, as expected. These methods exploit the complex and more informative structure of this data. However, when applying the multi-way decomposition methods i.e. MDC, results with linear classifiers are bad

for both data-sets. The parameters for the multi-way methods were optimized to model the structure of the data, but they do not model the classes. Therefore, there is still the need of complex non-linear classifiers to solve the classification problem.

The best results were obtained with the DR2D approach with the 2D measure for three-way spectral data. In the Wine data set, there is a large improvement when taking into account the multi-way structure and information of shape of the spectra. For the St John's data set there is also an improvement, although is not so remarkable. This classification problem is not that difficult, so there is not much to be improved. In this case, it is enough to use simple linear classifiers. This supports one of the purposes of this paper: showing the advantage of using the shape information in spectral data. It shows that in cases like these data sets, where there is functional information in one (GC-MS) or both directions, this approach could be a good option. Furthermore, it also demonstrates the advantage of the DR for small data sets. The use of a dissimilarity measure that describes the difference of classes as good as possible, helps to obtain a better separation of classes in the dissimilarity space. It shows that very good results can be obtained without the need of a complex non-linear classifier.

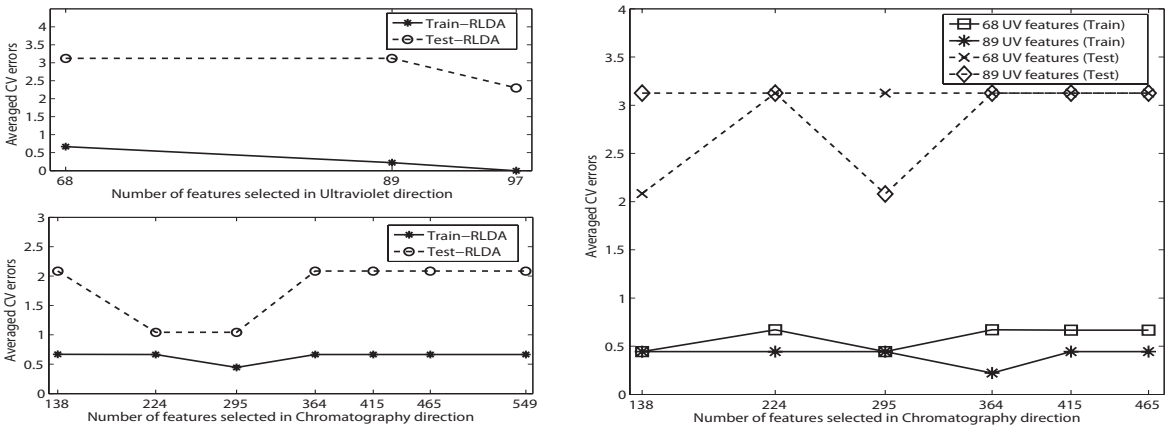


Figure 3.17: Feature selection with Mutual Information criterion on St John's data set: (top-left) Ultraviolet wavelengths selection only, (bottom-left) Chromatography retention times selection only and, (right) Selection in both directions at the same time. The results are shown in terms of the average classification error (in %) on the 6 training sets from the CV procedure.

In Figure 3.17, the results of the feature selection on St John's data set are shown. These are computed on the training sets and evaluated on the test sets from each fold in the CV procedure. Figure 3.17 (top-left) corresponds to results when selecting in the spectral mode only (UV). In this direction, there is one main big peak (related to the active component of the substances). Therefore, when selecting from $10 - 70\%$ of the features (68), they all correspond to that main peak. When including a second peak, we have then 80% (89) of the total number of features. If the training learning curve is analyzed, it can be seen that with 68 features (one peak), we already obtain very good results. There is an insignificant improvement when adding the features of the second peak. This is corroborated with the test learning curve. The ACE error is the same for both groups of selected features. This error value decreases a bit when using all features (See Table 3.3). The difference might be due to the selection process. It can happen that for a few objects, there are many chromatography peaks that have the same information if we throw out a part of the UV spectra. Therefore, it seems that a selection of those chromatography peaks that are better differentiated with the selected part of the UV spectra is needed.

In Figure 3.17 (bottom-left), the learning curve for the selection in the chromatography

direction only is shown. It can be noticed that results are very good since they start(for both training and test sets) i.e. with 10%(138) of the features. There is some point in the training learning curve (295 features-40%) for which the results improve, but it is not very significant. If we evaluate the test set by just using these features, a slightly better result than that for the full feature set (see also Table 3.3) is obtained.

Results for the selection in both directions at the same time are presented in Figure 3.17 (right). Each learning curve corresponds to a percentage of features selected in the UV mode. Each point of the curves are related to a percentage of features selected in the chromatography mode. We can observe that results are very similar to those when selecting in one direction only. For example, the training learning curve when using $10-70\%$ of the features in the UV direction and selecting in the chromatography direction, is very similar to that of Figure 3.17 (bottom-left). Thus, there are definitely redundant or not contributive features in both directions. They could be ignored without worsening the results. In this data set, for all possible variants of selection, we have seen that all results have a similar behavior in the training sets. It seems that there is a noisy example, which is classified differently depending on the shape it takes according to the selected features. Therefore, the shape of the training learning curves are dependent on this object. Following the procedure described in Section 3.4.3, we can include an error margin of, e.g. 1 wrongly classified object in the whole CV procedure, i.e. 1 out of 438 objects (73 objects $\times$ 6 folds). This way, we could avoid an over-fitting because of this problematic example. According to our definition of global minimum, it leads us to selecting 68 features ($10-70\%$) on the UV direction and 138 (10%) in the chromatography direction. If the selected features are now used to evaluate the test sets, we can see that we obtain an ACE= 2.1% (See Figure 3.17). Analyzing the test learning curves, it appears that the selection on the training set results is reasonable. Let us compare the results of the DR2D on the full (See Table 3.3) and reduced data. In terms of accuracy, there is not really a remarkable improvement. Results with the full data are already good, so it could be difficult to beat them. Although it is not very significant, better results are obtained by a selection just based on the chromatography (See Figure 3.17 bottom-left). In this case, it is up to the user to make a cost-accuracy trade-off.

If we consider the computational cost, there is a noticeable reduction. The computational complexity of a basic 2D dissimilarity measure should be in the order of $\mathcal{O}(n^2.m.l)$. The cost reduction of the training process is not clear. The number of operations for the feature selection has to be added to it. Nonetheless, once we have the most discriminative features, the number of operations in the test set is noticeably reduced. Let us make the calculations on the St John's data set for example, with the original size and the reduced size. In the test set, there is a reduction from $16*89*97*549$ to $16*89*68*138$ operations (5 times reduced). Especially in this data set, this procedure would then be very beneficial with respect to the computational cost.

Figure 3.18 shows part of the results for Wine data set. The figures are related to the selection in one of the directions and keeping all features in the other. If we take a look at the first figure (left), it can be observed that the ACEs for the training and test sets are quite large in the beginning. It can be due to the number of selected peaks, which is not sufficient for a good discrimination. Moreover, there might be noise introduced in the mass direction (which is very common), or there are many non-discriminative mass fragments. Thus, the information in the selected peaks with all mass information might not vary sufficiently to be discriminative. It can be also observed that, from the selection of 40% of chromatography peaks and on, the ACE values start decreasing until the full data size is reached. In Figure 3.18 (right), the results by just selecting in the mass mode are presented. In this direction, as there is no shape information to be taken into account, the MI weight is computed independently for each mass fragment. In this figure, if all chromatography peaks are used, by just keeping a 10% of the mass features the results are already very good for the training set. A slight improvement can be noticed from

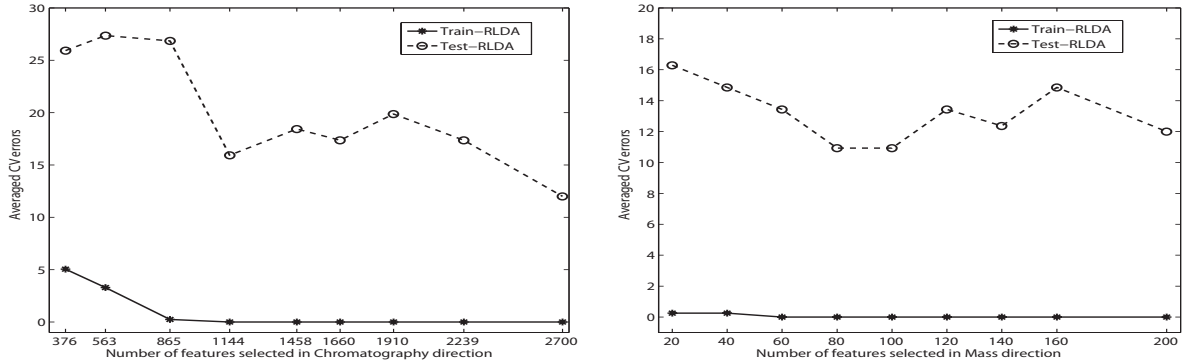


Figure 3.18: Feature selection with Mutual Information criterion on Wine data set: (left) Chromatography retention times selection only and (right) Mass (m/z) selection only.

the selection of 30% of mass features until the full size is reached.

If we compare the curves for training and test, there is a large difference between them which may point to overtraining. It should be noted that, the performance on the training set is serving as an optimization criterion for the feature selection. Therefore, due to the small size of the used training sets and the good separability of the classes in this representation, very often a perfectly separable training set is encountered, which might not be very useful. However, this difference between the curves can be explained by realizing that in a 10-fold cross-validation the test sets have a size of just 4, 5 or 7 objects at most in one of the folds. Thereby, the large error value of test results corresponds to just 0, 1 or 2 errors averaged.

Nevertheless, if the learning curve for the averaged error on the test sets (from CV) is analyzed, a similar behavior can be observed. Results for a small amount of mass fragments are not very good. There might be too little information on the mass spectra as to make the large amount of chromatography peaks different and informative enough. However, when using 50% (80 features) or more of the mass spectra ACE values are better than those with the full data. It can be observed that the curve varies somewhat afterwards. This could be due to sometimes noisy and other times redundant information is added, making the ACE to become worst or not. According to the previous analysis, a feature reduction in both directions at the same time seems to be more reasonable. Hence, a good combination (trade-off) of the informative/discriminative features in both directions can be found.

The ACE for selecting in both directions is shown in Figure 3.19. We decided to show the learning curves for all combinations as a surface plot, for a better appreciation of the results. The same phenomenon as in the previous figures is observed. However, the most important fact is that, the ACE when reducing in both directions at the same time is similar to that when reaching the total number of features. It seems that there is a lot of redundant information, which is not contributive at all. Therefore, let us apply the procedure explained in Section 3.4.3: with an error margin of one wrongly classified object. The global minimum corresponds to the combination of 20(10%) features in the mass mode and 376(10%) features in the chromatography mode. When evaluating the test sets with the selected features (See Figure 3.19 (right)), an ACE= 8.4% is obtained. If results on training (Figure 3.19 (left)) and test set (Figure 3.19 (right)) are compared, they are very consequent. It can be seen that, although in different scales, the shape (with some difference in proportion) of both surfaces is similar for most part of the figures. So, the results are consistent with the selection on the training set. It can also be noticed that for the test set, when reaching the full size of the data (top-left corner of Figure 3.19 (right)) the ACE value is very similar to that of Table 3.3 on the full data set. Thus, the proposed procedure has shown to be advantageous for this data set in two senses. First, the

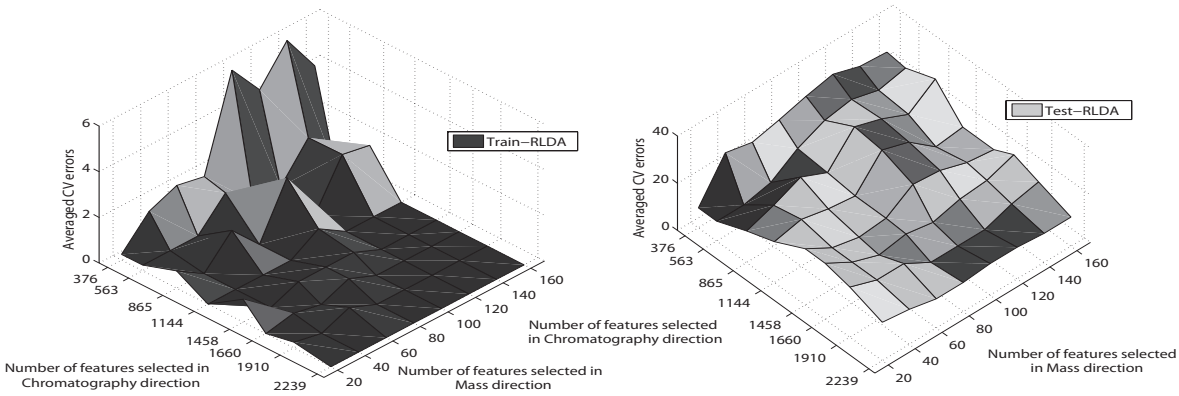


Figure 3.19: Feature selection with Mutual Information criterion on Wine data set: Surface for all combinations of features from both directions at the same time on (left) training and (right) test sets. The results are shown in terms of the average classification error (in %) on the 6 training sets from the CV procedure.

classifier accuracy was improved when ignoring what seemed to be noisy features. It shows that there can be small, good sets of features. Moreover, the computational cost is greatly reduced (72 times).

In this work we have shown the benefits of feature selection to optimize the DR, but another aspect most be added. Features are initially ranked according to their discriminative power. Thus, it gives the possibility of doing an exploratory analysis. Those chemical components (peaks) that are contributing the most to the classification problem can be determined.

3.4.6 Conclusions

We have investigated the use of an alternative representation for multi-way chemical spectral data. It makes use of the physical knowledge of the spectral background of the data by modeling their relations in a dissimilarity representation (DR). Moreover, it makes use of the information in both directions (multi-way structure) of the three-way data. In the comparison with the traditional methods for classifying multi-way spectral data, this procedure improves the results in the presented examples. It shows the benefit of taking into account discriminant characteristics of spectra like shape.

Inside the DR based procedure, we studied the possibility of optimizing the dissimilarity measure by using just a part (possibly most discriminative peaks) of the data. The selection based on the training set error appeared to be a reasonable approach. In the study, we showed that the introduced procedure can be very beneficial for high-dimensional data sets e.g GC-MS. It may be much more cost-effective as less measurements are needed and it may even yield better results.

3.5 Continuous Multi-way Shape Measure for Dissimilarity Representation

This section is a modified version of the article published as 'Continuous Multi-way Shape Measure for Dissimilarity Representation', by D. Porro-Muñoz, R.P.W Duin, I. Talavera and M. Orozco-Alzate, in *Proceedings of the 17th Iberoamerican Congress on Pattern Recognition, CIARP2012, LNCS*. In this version, more details about the existent gradient convolution kernels and related works are added. These were skipped in the original version due to space problems. Notation was also changed in order to have a unified notation in the whole thesis.

Abstract

For many applications, a straightforward representation of objects is by multi-dimensional arrays e.g. signals, spectroscopic data. However, there are only a few classification tools that make a proper use of this complex structure to obtain a better discrimination between classes. Moreover, they do not take into account background knowledge information of the application that can also be very beneficial in the classification process. Such is the case of multi-dimensional continuous data, where there is a connectivity between the points in all directions, a particular (differentiating) shape in the surface of each class of objects. The dissimilarity representation has been recently proposed as a tool for the classification of multi-way data, such that the multi-dimensional structure of objects can be considered in their dissimilarities. In this paper, we introduce a dissimilarity measure for continuous multi-way data. It allows using the information on how the shape of the surface varies in all directions. Experiments show the suitability of this measure for classifying continuous multi-way data.

3.5.1 Introduction

Representation of objects by matrices or higher-order arrays has become very popular in many application areas. In some cases, different types of features are arranged together and related in a single structure. This aims at extracting relationships and patterns that can be used for other analysis e.g. a three-way array defined as: $users \times queries \times webpages$ in order to improve personalized web searches [127]. In other cases, this multi-dimensional representation can be obtained directly from acquisition equipments. Such is the case of some spectroscopy equipments like surface excitation-emission autofluorescence [11] measurements and time-frequency representation of signals [100], etc.

In any case, tools that make a proper use of the multi-way structure are needed. Traditional multivariate methods are not suitable for it. Data would have to be re-arranged in a vector, thus the information of the multi-way structure is lost and huge-dimensional problems are created. Therefore, multi-way data will not be analyzed optimally.

Several methods have been proposed for multi-way data analysis, mainly in the psychometrics and chemometrics fields. However, most of them aim for regression e.g. N-PLS or exploratory analysis e.g. PARAFAC, TUCKER3 [126], to extract hidden structures and capturing underlying correlations between features in the multi-way array. The development of tools for multi-way classification is rather poor in comparison with the large amount of methods for other purposes. There are basically three main multi-way classification approaches, namely: NPLS-DA [11], building traditional multivariate classifiers on the scores of the multi-way decomposition methods like PARAFAC [2], and NSIMCA [37]. These methods succeed in making use of the complex multi-way structure. However, this might not be enough for a proper analysis of the data. Continuous data, for example, are just numerically analyzed as a set of sampled, individual features. Features are in this paper, the measurements/observations in the different modes of a multi-way data set. Important discriminative information like their shape is not considered in the analysis.

Recently, the Dissimilarity Representation (DR) approach was introduced for multi-way data classification [100]. The DR approach consists in representing objects by their proximities with respect to other objects [94]. As classes of objects are determined by how (dis)similar they are, the authors advocate that for classification, a representation based on dissimilarities between objects may be more informative than the traditional feature based representation.

One of the goals and advantages of this approach is the possibility of introducing discriminative context information in the representation of objects by the dissimilarity measure. In the case of 2D or any-dimensional continuous data, e.g. aligned images, spectroscopic data, it could be important for their analysis to consider their functional (continuous) nature: changes

in the shape of the surface of the different classes. Several 2D measures have been proposed for image comparison; however most of them are just based on the pairwise comparison of objects, ignoring the continuous nature of images e.g. AMD [161], Frobenius [157]. Recently, the 2Dshape [100] measure was introduced for this purpose. This measure takes into account the information on both directions of the array. In order to reflect the shape of the data, the comparison of objects is based on the differences between the first Gaussian derivatives in each direction. However, it is based on the combination of 1D dissimilarities; hence it does not analyze the combined 2D shape changes [1] It was mainly thought (can be adapted) for data sets with features of different nature in the different directions e.g. data with a continuous and a non-continuous direction, like in Gas Chromatography/Mass Spectrometry [8] or in social network analysis: $users \times keywords \times time\ samples$.

In this paper, we propose a new dissimilarity measure, the Continuous Multi-way Shape (CMS) measure, that exploits the information on the whole structure of multi-dimensional continuous data. It is based on the differences between the gradients of objects, thus the shape changes of the surfaces are considered. The point connectivity in all directions is used simultaneously. The computation of the gradient components is usually based on convolution kernel operators. We also introduce a gradient kernel operator where each partial derivative is computed as the derivative of a polynomial fitted to the analyzed points. The proposed measure will be compared to the 2Dshape method [100], and to two traditional measures [161], which are not based on shapes of continuous data types. It is shown that considering the continuous nature of data can be beneficial to improve its classification.

The paper is organized as follows. Fundamentals of the dissimilarity representation approach are presented in Section 3.5.2. In Section 3.5.3 and Section 3.5.4 the new measure and its generalization are detailed. Experiments and results are discussed in Section 3.5.5. Conclusions are presented in Section 3.5.6.

3.5.2 Dissimilarity Representation for Multi-way Data

In the DR [94] approach, new features are defined for the objects, such that they are represented by their proximities to a set of representative objects of each class. The fact (property) that dissimilarities should be smaller for similar objects and larger for different ones, suggests that they could be used as more discriminative features, if a suitable measure is used.

Thus, in this approach, given a set of training objects $\mathbf{X} = \{x_1, x_2, ..., x_l\}$, a representation set (a set of prototypes for each class) $\mathbf{R} = \{r_1, r_2, ..., r_p\}$, and a dissimilarity measure; the distance between each object $x_i \in \mathbf{X}$ to each object $r_h \in \mathbf{R}$ will be defined as $d(x_i, r_h)$. The representation set $\mathbf{R}$ can be a subset of $\mathbf{X}$, $\mathbf{R} \subseteq \mathbf{X}$ or $\mathbf{X}$ itself, being then $\mathbf{D}(\mathbf{X}, \mathbf{X})$ a square dissimilarity matrix, or $\mathbf{R}$ and $\mathbf{X}$ can be completely different sets. There are a number of approaches to select prototypes of the representation set [94].

Let us assume we have an n-dimensional array $\underline{Y} \in \mathbb{R}^{I_1 \times I_2 \times ... \times I_n}$ (training set), where each object is $\underline{Y}_i \in \mathbb{R}^{I_1 \times I_2 \times ... \times I_{n-1}}$. To build the dissimilarity space, the mapping $\phi(\cdot, \underline{R}) : \mathbb{R}^{I_1 \times I_2 \times ... \times I_{n-1}} \rightarrow \mathbb{R}^h$ is defined, such that for every object $\phi(\underline{Y}_i, \underline{R}) = [d(\underline{Y}_i, \underline{R}_1), d(\underline{Y}_i, \underline{R}_2), ..., d(\underline{Y}_i, \underline{R}_h)]$. We need then a ($n$-1)-dimensional dissimilarity measure for its computation. Classifiers are built in this space, as in any feature space. Consequently, the relationship between all objects in the training and representation sets is used for classification. If a suitable measure is chosen, the compactness property (objects from the same class should be similar and objects from different classes should be different) of the classes should be more pronounced. Therefore, it should be easier for the classifiers to discriminate between them, since linear classifiers in the dissimilarity space may correspond to non-linear classifiers in the feature space. In general, any arbitrary classifier operating on features can be used [94].

[1] The naming of the procedure given in paper [100], 2Dshape, can be confusing.

3.5.3 Continuous Multi-way Shape Measure

A measure that somehow respects the multi-way structure of the data (considers the relationship of different directions) is needed. It should also make use of the continuity and shape information of the multi-dimensional continuous data.

Given an $n-1$ dimensional object, each point in the multi-dimensional surface could be analyzed with an $n-1$ dimensional window, such that shape changes in all directions can be taken into account. Thus, the comparison between two objects should be based on the differences of their multi-way shape, considering the connectivity that exists between the neighboring points in the different directions. In the case of 1D continuous data e.g. spectral data, derivatives are the commonly used tool to evaluate shape changes. For multi-dimensional functions, the gradient is the natural extension of the derivative concept. It is defined as a vector of the derivatives of the function with respect to the different coordinate axes (partial derivatives).

Although data may have a continuous nature, they are captured by the sampling procedures of sensors as a collection of discrete values. As derivatives are undefined for discrete functions, they need to be estimated somehow to be used on these data. A widely used method for approximating the derivative of a discrete function is the application of linear filters by convolution.

Given the multi-dimensional discrete functions $\underline{Y}$ and $\underline{H}$, a convolution operation is defined as follows:

$$\underline{Y}' = \underline{Y} * \underline{H} \tag{3.10}$$

where $*$ is the convolution operator [49]. Thus, given two objects $\underline{Y_a}$ and $\underline{Y_b}$, they can be compared by computing the difference between the gradients of the surfaces that represent these objects. As derivatives and therefore the gradient are very noise sensitive, data should be smoothed before performing these operations. A common way to smooth data is by convolving it with a Gaussian filter. Following the previous ideas, the Continuous Multi-way Shape (CMS) measure is defined as:

Definition 1 *Let $\underline{Y}$ be an n-way data set and let $\underline{Y_a}$, $\underline{Y_b}$ be two objects from this data set. The dissimilarity between $\underline{Y_a}$ and $\underline{Y_b}$ can be computed as:*

$$d_G(\underline{Y_a}, \underline{Y_b}) = \left\| \sum_{i=1}^{f} \underline{Y_a} * \underline{G_\sigma} * \underline{H_i} - \underline{Y_b} * \underline{G_\sigma} * \underline{H_i} \right\|_F \tag{3.11}$$

where $\|\cdot\|_F$ is the Frobenius norm for tensors [68], $\underline{G_\sigma}$ a Gaussian convolution kernel to smooth the data first, $\underline{H_i}$ is a partial derivative kernel and f is the amount of partial derivatives in the different directions in order to obtain the gradient.

Gradient kernels

The Prewitt operators [49] are 2D filters and they are based on linear filters that compute the average gradient components of three adjacent lines and columns to overcome the noise sensitivity. The horizontal, vertical and the two diagonal (main and secondary) Prewitt kernels are defined as follows:

$$H_x^P = \begin{array}{|c|c|c|}\hline -1 & 0 & 1 \\ \hline -1 & 0 & 1 \\ \hline -1 & 0 & 1 \\ \hline \end{array} \quad H_y^P = \begin{array}{|c|c|c|}\hline -1 & -1 & -1 \\ \hline 0 & 0 & 0 \\ \hline 1 & 1 & 1 \\ \hline \end{array} \tag{3.12}$$

$$H_{md}^P = \begin{array}{|c|c|c|}\hline -1 & -1 & 0 \\ \hline -1 & 0 & 1 \\ \hline 0 & 1 & 1 \\ \hline \end{array} \quad H_{sd}^P = \begin{array}{|c|c|c|}\hline 0 & -1 & -1 \\ \hline 1 & 0 & -1 \\ \hline 1 & 1 & 0 \\ \hline \end{array} \tag{3.13}$$

The kernels for Sobel [49] operators are very similar to those of Prewitt. They differ in that central weights for the smoothing are higher. Horizontal,vertical and the two diagonal Sobel kernels are then defined as:

$$H_x^S = \begin{array}{|c|c|c|}\hline -1 & 0 & 1 \\ \hline -2 & 0 & 2 \\ \hline -1 & 0 & 1 \\ \hline \end{array} \quad H_y^S = \begin{array}{|c|c|c|}\hline -1 & -2 & -1 \\ \hline 0 & 0 & 0 \\ \hline 1 & 2 & 1 \\ \hline \end{array} \tag{3.14}$$

$$H_{md}^S = \begin{array}{|c|c|c|}\hline -2 & -1 & 0 \\ \hline -1 & 0 & 1 \\ \hline 0 & 1 & 2 \\ \hline \end{array} \quad H_{sd}^S = \begin{array}{|c|c|c|}\hline 0 & -1 & -2 \\ \hline 1 & 0 & -1 \\ \hline 2 & 1 & 0 \\ \hline \end{array} \tag{3.15}$$

The previous gradient operators are based on approximating a partial derivative at a point p by computing the slope of the line that fits the previous and next point of p in the direction of the derivative. Extensions of these filters to diagonal directions and to higher dimensions can be found [58], and the same idea of a line fitting is kept.

3.5.4 Gradient Polynomial-Based Kernel for the CMS Measure

We propose to approximate each partial derivative in point p as the derivative of the polynomial of degree t, which is obtained by interpolating p and its t nearest points in the direction of the derivative. This technique can also be wrapped in a linear filter.

For a 2D kernel of size $[3 \times 3]$, this approach is equivalent to applying the Prewitt operator. Thus, the particular case of a 2D kernel of size $[5 \times 5]$ will be explained here. For 2D objects, we want to analyze the derivatives in four directions (horizontal, vertical and the two diagonals). Without loss of generality, let us see the case for the horizontal direction. Assume there are five points $p_0 = (-2, y_0)$, $p_1 = (-1, y_1)$, $p_2 = (0, y_2)$, $p_3 = (1, y_3)$ and $p_4 = (2, y_4)$, so a fourth degree polynomial should be approximated to compute the derivative in each direction. We chose these values for x just for simplicity, but it really does not matter, the shape of the polynomial is the same:

$$P(x) = ax^4 + bx^3 + cx^2 + dx + e \tag{3.16}$$

Its derivative is the cubic polynomial $P'(x) = 4ax^3 + 3bx_2 + 2cx + d$ and, evaluated in point p_2, it is reduced to $P'(0) = d$. If the system of equations from evaluating all points in $P(x)$ is solved, it is obtained:

$$P'(0) = d = \frac{y_0 - 8y_1 + 8y_3 - y_4}{12} \tag{3.17}$$

which can be wrapped in a filter kernel. We do not want to use the exact magnitude of the objects gradient for a further analysis. It is just needed to perform a comparison between objects, thus the division by twelve (in this case) or any number obtained in the computed derivative expression does not interfere in this analysis.

As for Prewitt and Sobel operators, where adjacent lines are averaged, an approximated average polynomial of the adjacent set of points can be obtained. Then, the following convolution kernels for horizontal, vertical, main diagonal and secondary diagonal are defined respectively:

$$H_x^L = \begin{array}{|c|c|c|c|c|}\hline 1 & -8 & 0 & 8 & 1 \\ \hline 1 & -8 & 0 & 8 & 1 \\ \hline 1 & -8 & 0 & 8 & 1 \\ \hline 1 & -8 & 0 & 8 & 1 \\ \hline 1 & -8 & 0 & 8 & 1 \\ \hline \end{array} \quad H_y^L = \begin{array}{|c|c|c|c|c|}\hline 1 & 1 & 1 & 1 & 1 \\ \hline -8 & -8 & -8 & -8 & -8 \\ \hline 0 & 0 & 0 & 0 & 0 \\ \hline 8 & 8 & 8 & 8 & 8 \\ \hline -1 & -1 & -1 & -1 & -1 \\ \hline \end{array} \tag{3.18}$$

$$H^L_{md} = \begin{array}{|c|c|c|c|c|}\hline 0 & 0 & 0 & 1 & 1 \\ \hline 0 & 0 & -8 & -8 & 1 \\ \hline 0 & 8 & 0 & -8 & 0 \\ \hline -1 & 8 & 8 & 0 & 0 \\ \hline -1 & -1 & 0 & 0 & 0 \\ \hline \end{array} \quad H^L_{sd} = \begin{array}{|c|c|c|c|c|}\hline 1 & 1 & 0 & 0 & 0 \\ \hline 1 & -8 & -8 & 0 & 0 \\ \hline 0 & -8 & 0 & 8 & 0 \\ \hline 0 & 0 & 8 & 8 & -1 \\ \hline 0 & 0 & 0 & -1 & -1 \\ \hline \end{array} \tag{3.19}$$

The same idea can be generalized to larger windows, but higher order polynomials are used. These filters can also be extended to n-way arrays. Polynomials will be determined in the same manner according to the size of the window, but now there will be more directions to be analyzed. For example, in the case of a 4-way array where objects are 3D, if we use a $[3 \times 3 \times 3]$ window, 13 directions can be analyzed as in Figure 3.20.

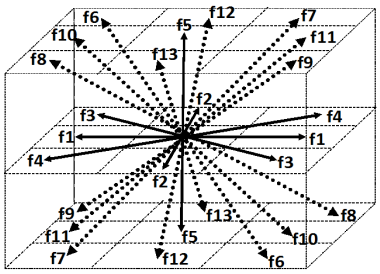


Figure 3.20: The 13 directions to be analyzed in a 3D object

The proposed CMS measure can be seen as a generalization of the idea of the 2Dshape [100] measure for 2D objects. Both measures are based on smoothing surfaces with a Gaussian kernel and their partial derivatives are compared in order to take the shape of the functions into account. However, in the case of the 2Dshape measure, a 1D Gaussian derivative is computed independently for every feature in the horizontal and vertical directions of a 2D object, treating that feature in the selected direction as a 1D signal, without taking into account the relationship between the features in the two directions. Differences in each direction are then combined.

In contrast, the CMS measure already considers 2D information in the smoothing process by applying a 2D Gaussian kernel. Moreover, with 2D gradient kernel operators, more global information (relationship between features) is considered. As partial derivatives are computed on the average line obtained from a number of rows/columns (depending on the kernel size), information from the neighboring signals is used. The most remarkable aspect of the CMS measure is that it is not restricted to measure the shape in two directions only. Although CMS is based on the idea of the gradient of an image, which is mathematically defined by a vector of 2 components i.e. horizontal and vertical, this measure allows analyzing other directions e.g. diagonals. This way, more accurate approximations of the information on the shape of objects and dependencies in the different directions are computed for the comparison. With the introduced polynomial-based kernels, the partial derivatives can be approximated as the derivative of higher-degree polynomials instead of a simple line. Consequently, it is to be expected that the multi-way shape information can be better modeled with the proposed CMS measure, leading to a better discrimination of these types of objects. At last, the CMS measure has been defined for continuous multi-way objects in general, while 2Dshape can only be applied to 2D objects.

3.5.5 Experimental Setup and Discussion

In this section, we present experiments with a Regularized Linear Discriminant classifier based on dissimilarities. Experiments are conducted on five 2D continuous data sets of different

sources. Our aim is to compare the performance of classifiers on the two shape-based measures, 2Dshape and the proposed CMS measure for multi-way continuous data. These performances will also be compared with the non-shape based measures Frobenius and Yang, which are versions of the AMD [161] distance with weights $p = 2$ and $p = 1$, respectively.

For the different data sets, experiments were carried out differently. For small data sets (Parma ham and St John's), classification errors were obtained in a 10 times k-fold cross-validation (CV). The Enzyme data comes with training and test, so they were evaluated. In the case of Colon and Volcano data, 10 different training and test sets were randomly chosen and the error values were averaged. For the three bigger data sets, i.e. Enzyme, Colon and Volcano, a part of the data was used to optimize the measures parameters in a CV procedure. The rest was then used as explained before. For the other two, as they are too small, the parameters were optimized with the whole data sets.

The first data set is private and it comes from 1200 patches of 1024×1024 pixels of 36 colon tissue slides from Atrium hospital in Heerlen, The Netherlands. Patches were filtered with Laplace filters in 90 different scales using $\sigma = 2.\hat{}[0.1 : 0.1 : 9]$. The log-squares of the results are summarized in 60 bin normalized histograms with bin centers $[-50 : 1 : 9]$. Thus, a 90×60 array is obtained for every patch, leading to a three-way array of $1200 \times 90 \times 60$. The patches are labeled in two classes: Normal and Tumor. A representation set of 550 prototypes was randomly chosen from the training set.

The second data set corresponds to seismic signals from Nevado del Ruiz volcano in the Colombian Andes. The data set is composed of 12032-point signals of two classes of volcanic events: Long-Period earthquakes, and Volcano-Tectonic earthquakes. A 2D time-frequency representation was computed for each signal with a 256-points (window size) Short-Time Fourier Transform (STFT), with 50% overlap. The concatenation of the obtained spectrograms results in a $470 \times 93 \times 129$ three-way array. The dissimilarity matrix has a size of 470×100.

The third and fourth data sets are from public domains and they are both obtained by Fluorescence spectroscopy. The first of them consists of a training set of size $323 \times 15 \times 15$ and a test set of $53 \times 15 \times 15$. The two feature directions correspond to excitation and emission wavelengths, respectively. The classification problem consists in determining the quality (Low or High) of a process according to the enzyme activity [96]. A representation set of 100 prototypes was randomly chosen from the training set, thus the dissimilarity matrices have a size of 323×100 and 53×100 for training and test sets, respectively. The other data set has a size of $67 \times 11 \times 13$ and the purpose is to determine the age range of a Parma ham example: raw (0 months), salted (3 months), matured (11 and 12 months) and aged (15 and 18 months) [37].

The last data set consists of 108 examples of carrot juice, which have been crystallized, with the aim of describing their quality (Good/Bad) [26]. Images of size 528×528 of each biocrystallized example were taken. Gabor filters with a bank of 128 filters from 16 orientations was applied, resulting in a four-way data set of $108 \times 528 \times 528 \times 128$. Thus, 128 dissimilarity matrices were computed with 2D measures on the 528×528 matrices of each filter and latter averaged. All objects were used as representation set in the DR.

Table 3.4: Classification error with different measures: CMS measure with Prewitt, Prewitt in 4 directions (including diagonals), Sobel, Sobel in 4 directions and Polynomial filter, 2Dshape measure, Frobenius and Yang.

	CMS						No shape	
Data	Prew.	Prew.(4d)	Sob.	Sob.(4d)	Polyn.	2Dshape	Frob	Yang
Colon cancer	11.0	11.5	11.2	12.0	**9.5**	12.7	13.3	13.3
Volcano	28.0	25.6	28.2	23.4	23.4	**20.9**	40.0	28.7
Enzyme	9.4	**5.7**	9.4	9.4	9.4	13.2	9.4	9.4
Parma ham	3.7	**2.4**	3.7	2.5	3.7	2.9	4.5	4.3
Carrot juice	7.2	**6.0**	7.2	6.3	7.1	8.3	9.8	10.7

Results are shown in Table 3.4. It can be seen that as expected, measures that take the continuous information of data into account give the best results. The CMS measure, with most filters, outperforms the results obtained with the 2Dshape measure in general, corroborating our previous analysis.

The selection of the kernel to be applied should depend on the problem at hand and how rough shape changes are. Larger kernels should be able to capture better the changes in the surface when these are not so sudden. However, if there are shape changes in small regions, they might be averaged in a large window. Thus, small windows should work better in these cases. It is shown that results are improved by using the diagonal directions in the Prewitt and Sobel operators. This supports the previously discussed argument that if more directions are analyzed, there can be more information that contributes to a better discrimination of the classes.

3.5.6 Conclusions

We introduced a multi-dimensional dissimilarity measure for multi-way continuous data based on the computation of the gradient. This was proposed with the aim of applying the DR approach as a classification tool for this type of data. The new measure allows taking into account the complex multi-dimensional structure, such that the shape information of the surfaces (objects) can be considered in the dissimilarity representation of the objects. The way the measure has been defined, allows to use different gradient convolution kernels, according to the problem at hand. This measure was compared to the 2Dshape measure and other non-shape based measures for the classification of 2D objects. Results have corroborated the presented argument that considering the continuous multi-way nature of these types of data in their analysis can lead to better results. Moreover, it is shown that by taking into account the information in more directions, results can be improved.

Chapter 4

Missing values in dissimilarity-based classification of multi-way data

This chapter is an extended version of the paper accepted for publication as 'Missing Values in Dissimilarity-Based Classification of Multi-way Data', by D. Porro-Muñoz, I. Talavera and R.P.W Duin, in *Proceedings of the 18th Iberoamerican Congress on Pattern Recognition, CIARP2013, LNCS*. In this version, there is a more extensive study on the state of the art methods for treating missing values in multi-dimensional data. It includes the analysis of three more approaches for dealing with missing values in the dissimilarity-based classification of multi-way data. A third pattern of missing values in multi-way data was also added to the analysis. The full chapter was structured as a journal article and is currently under review in the *International Journal of Pattern Recognition and Artificial Intelligence*.

Abstract
Missing values can occur frequently in many real world situations. Such is the case of multi-way data applications, where objects are usually represented by arrays of 2 or more dimensions e.g. biomedical signals that can be represented as time-frequency matrices. This lack of attributes tends to influence the analysis of the data. In classification tasks for example, the performance of classifiers is usually deteriorated. Therefore, it is necessary to address this problem before classifiers are built. Although the absence of values is common in these types of data sets, there are just a few studies to tackle this problem for classification purposes. In this paper, we make a study of different methods for reconstructing or dealing with multi-way incomplete data. Performance of dissimilarity-based classifiers has been evaluated for these methods on different patterns of missing values. Some of the analyzed techniques have shown to be suitable for the problem at hand; favoring the good performance of classifiers even for large amounts of missing attributes (up to 70%).

4.1 Introduction

Classification problems are very common in most research areas. Thus, class modeling and labeling of unknown objects is one of the main tasks in pattern recognition. Object representation plays an important role in this task. However, even when a proper representation is found, problems like the absence of values for some of the measured features can affect the accuracy of classifiers.

The so-called missing values are part of almost all research areas. There can be several reasons for data to be missing. It can be due to equipments malfunctioning, noise, loss information, incomplete experiments, experiments are too costly, data were not entered correctly or data just do not exist for some objects. For example, in biomedical signal processing or chemometrics, equipments may have some problems and signals are lost or have some irregularities in the measurements. In other cases, missing values are not actually present in the data obtained directly from the equipment. Nonetheless, they are inserted as a post-processing in order to make the data more suitable to be described for some specific models [10].

Depending on the reason for the existence of missing values, they may have different distributions. If they are scattered without showing and specific pattern, they can be categorized as missing completely at random or just missing at random. When they have a pattern throughout the array because e.g. some data could not be collected, then they are considered as not at random [120]. In many cases missing values are not treated, but omitted in the analysis by skipping the incomplete objects or the missing attribute. The advantage of this approach is application-dependent. When deleting features, specially for random missing values, important discriminative features might be discarded just because there was one value missing. Another common approach for handling missing values in multivariate data is the imputation approach. Missing data is estimated by using information from the data set. Examples of common imputation approaches are: mean, hot deck [74], regression based-substitution like support vector regression [56], maximum likelihood-based methods [119, 64], multiple imputation [119, 121].

For many applications e.g. neuroinformatics, chemometrics, psychometrics, web mining, data sets can have a multi-dimensional structure e.g. $objects \times frequencies \times time - points$, instead of the simple vector representation. These structures are often richer in information, thus advantageous for many purposes as classification. Therefore, it is important to employ proper tools in order to analyze them.

As in the two-dimensional case, these types of data may be affected by the presence of missing values. However, as there is more information, data analysis should be easier and estimations more reliable. For multi-way data, different behaviors for missing values can also be observed [129, 64] (See Figure 4.2). The simplest case (maybe not so common) is when missing values are randomly scattered without any pattern, denoted as RMV in [129]. Another common pattern is when complete fibers i.e. rows or tubes (See Figure 4.1) are missing at random (RMF). This can be the case when objects are analyzed with a spectrometer and it stops functioning for a moment. Measurements are not taken and a whole spectrum might be missing. A third pattern that may be identified is when missing values are systematic for all objects (SMV) i.e. the same columns are missing for all objects. An example is the excitation-emission fluorescence data. With this particular technique, signals registered at emission wavelengths lower that excitation wavelength do not exist physically, so those values are set to missing [129].

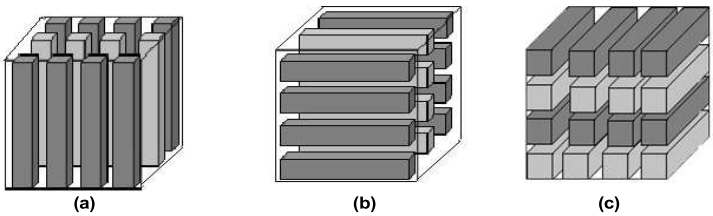


Figure 4.1: Definition of (a) columns, (b) rows and (c) tubes in three-way data

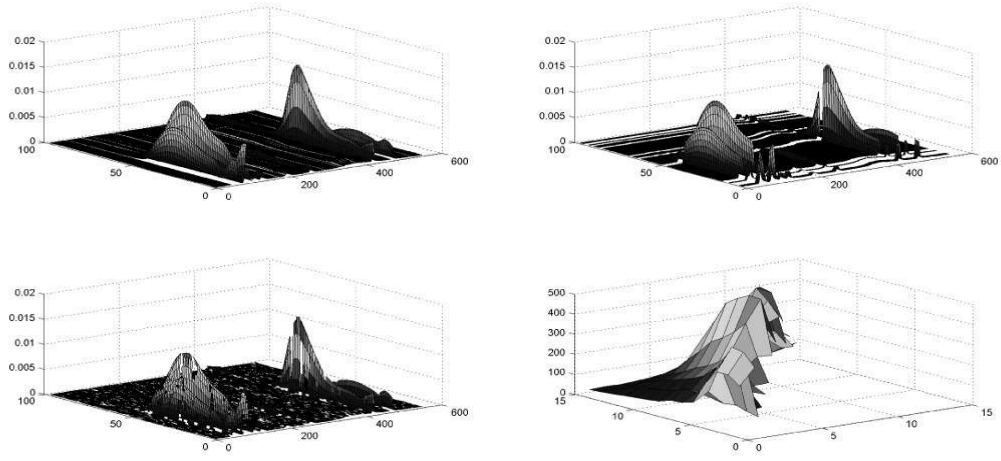


Figure 4.2: Different patterns of missing values in a 2D object: Randomly Missing Fiber (RMF) (a) Rows and (b) Tubes, (c) Randomly Missing Value (RMV) and (d) Systematically Missing Values (SMV) with an example of excitation-emission fluorescence data.

In contrast with the two-way case, there is a limited research addressing the problem of missing values in multi-way data. Most of the related studies are dedicated to the robustness of factorization methods. Examples of the most common methods are PARAFAC algorithms based on Expectation Maximization - Alternating Least Squares [129, 9] and based on the

Levenberg - Marquadt method known as INDAFAC [129]. A more recent development is the CP-WOPT algorithm [5], which is also proposed as an improved PARAFAC algorithm for handling missing data. Most of these methods have been developed by the psychometrics and chemometrics communities, where multi-way analysis has been one of the main research topics. Other extensions of the existent multi-way methods like TUCKER3 for dealing with missing values can be found in [141, 142, 118, 126, 64]. However, these methods are based on seeking accuracy in the obtained factors.

In this paper, we make a study of approaches for dealing with missing values in multi-way data, such that the error function in the classification is minimized. We will use the Dissimilarity Representation (DR) [94] approach recently extended for the classification of multi-way data [100]. Roughly speaking, in this approach, (dis) similarities between objects are used as new features to describe them in a new space. Classifiers can be used in this space as in the traditional feature space. One of the options to deal with missing data in this case could be to reconstruct the data before the computation of the dissimilarity matrix. For this purpose, we will compare different techniques that can be used for the estimation of missing values in multi-way data. Namely, simple imputation with fiber mean, triangulation-based interpolation, PARAFAC and CP-WOPT. In this case, if one of the last two methods is applied, once the factorization is done, it can be used to reconstruct the incomplete data. Another variant for dealing with missing data, particularly for the DR approach, consists in modifying the dissimilarity measure. With this purpose, a modification of the dissimilarity measure that will be used here, was also introduced in the comparison.

Summarizing, we made a study of different methods that can be used for dealing with missing values in multi-way data for classification purposes. Classification is based on the DR approach. Some of the methods used for imputation have been previously proposed for other types of problems, but they can also be fitted for this particular problem. All of them were analyzed for the different patterns of missing values in multi-way data. Advantages and disadvantages of each of them are discussed. An experimental study is also carried-out on 4 data sets, by varying the amount of missing values and their patterns.

The rest of the paper is organized as follows. In Section 4.2, the DR approach is briefly explained. A description and comparative analysis of the studied methods is presented in Section 4.3. Section 4.4 is dedicated to the experiments and discussion. Conclusions are finally presented in Section 4.5.

4.2 Dissimilarity Representation

The Dissimilarity Representation (DR) [94] approach has been introduced for classification purposes. It consists in a representation of objects by their (dis) similarities to a set of prototypes of each class identified in the problem at hand. Thus, every object is represented by a vector of dissimilarities to other objects, instead of attributes as in the traditional feature space. One of the advantages of this approach is that it can be obtained from any representation of objects e.g. graphs, multi-dimensional objects, as long as a suitable measure is used. Moreover, this approach allows introducing discriminative context information that helps for a better discrimination of objects

So, let us define the Dissimilarity Space (DS) approach, given a t-way array $\underline{Y} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_t}$ where each object is represented by a (t-1)-dimensional array, a representation set $\underline{R}(\underline{R_1}, ..., \underline{R_h})$ where h is the number of prototypes, and a dissimilarity measure d [94, 100]. A mapping $\phi(\cdot, \underline{R}) : \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_{t-1}} \rightarrow \mathbb{R}^h$ is done, such that every object $\phi(\underline{Y_i}, \underline{R}) = [d(\underline{Y_i}, \underline{R_1}), d(\underline{Y_i}, \underline{R_2}), ..., d(\underline{Y_i}, \underline{R_h})]$ is associated by its dissimilarities to all objects in $\underline{R}$. Hence, a dissimilarity matrix $\mathbf{D}(\underline{Y}, \underline{R})$ is obtained, which is used to build a classifier in the correspondent dissimilarity space of dimension h. The prototypes are usually the most representative objects of each class, $\underline{R} \subseteq \underline{Y}$ or $\underline{Y}$ itself,

resulting in a square dissimilarity matrix $\mathbf{D}(\underline{Y},\underline{Y})$. Any traditional classifier can be built in the dissimilarity space as in the feature space.

Few work has been done to treat missing data in the DR approach. In [79], two alternatives for dealing with missing values in the dissimilarity representation-based classification are proposed. However, this work is only based on 2D data, where objects are represented by vectors in the feature space. It does not fit multi-way data.

In this paper, we propose two alternatives for classifying incomplete multi-way data by using the dissimilarity representation. The first approach is based on completing the multi-way data before computing the dissimilarity matrix. Three different imputation options are discussed in Section 4.3. The second alternative consists in adapting the dissimilarity measure, such that the dissimilarities between objects are obtained from the available information only.

The data sets to be studied here have a continuous nature. The characteristic shape of the surfaces for each class of objects is an important discriminative property of these type of data. Moreover, the information from the multi-way structure should be taken into account. Recently, the Continuous Multi-way Shape (CMS) [103] was introduced with this purpose and it will be used here.

The CMS measure consists in the comparison of multi-way objects based on the differences of their multi-way shape, considering the connectivity that exists between the neighboring points in the different directions. With this purpose, differences between the gradients of the surfaces of these objects are computed, based on the application of linear filters by convolution.

Thus, given $\underline{Y_a}$, $\underline{Y_b}$ two multi-way objects from a multi-way data set $\underline{Y}$, the dissimilarity measure CMS can be defined as:

$$d_G(\underline{Y_a},\underline{Y_b}) = \left\| \sum_{i=1}^{f} \underline{Y_a} * \underline{G_\sigma} * \underline{H_i} - \underline{Y_b} * \underline{G_\sigma} * \underline{H_i} \right\|_F \tag{4.1}$$

where $\|\cdot\|_F$ is the Frobenius norm for tensors [68], $*$ is the convolution operator [49], $\underline{G_\sigma} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_{t-1}}$ a Gaussian convolution kernel to smooth the data first, $\underline{H_i} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_{t-1}}$ is a partial derivative kernel and f is the amount of partial derivatives in the different directions in order to obtain the gradient.

Details on how the CMS measure can deal with missing values will be given in the next Section.

4.3 Dealing with missing values in multi-way data

4.3.1 Simple imputation: Averaging

Imputation is one of the most common approaches to deal with missing values. It is the process in which missing data are estimated such that a complete data set is obtained. Mean-based imputation is one of the simplest and frequently used techniques in 2D data, although considered among the less accurate. Recently, this technique was extended for missing value imputation of three-way DNA microarray analysis [72]. Given the three-way array $\underline{Y}$ (as defined in Section 4.2), the value in position $\underline{y}_{ijk}$ can be estimated in three different ways. It can be obtained by averaging the information in the kth row or jth tube of the ith object. A problem of this technique is that by averaging all values in the selected direction, values that are very far (which can be very different) from the position to be estimated have the same influence as the closest (related) values. Hence, the accuracy of the method is affected. A second way to perform this approach is by averaging the values of position (j,k) for all objects. This is also quite problematic. As it is unsupervised, values of objects from all classes are averaged. There can be too much variability introduced in the averaged values, therefore the final result can be something far from its real value. Another problem is that in the case where a random tube/row

is missing, the averaging by moving in the contrary direction i.e. averaging values of all tubes for missing row k, does not work, as all values will be missing for that position.

4.3.2 Factorization-based estimation

Factorization methods are very common for the analysis of the traditional multivariate data and multi-way data. They are used to extract and model the underlying structure i.e. finding patterns, relationships between the features in the different dimensions, of the high-dimensional data. These methods are of course affected by the presence of missing values, as data can be improperly analyzed. Therefore, creating robust methods to missing data has been one of the main tasks in the development of factorization methods [64]. Such is the case of the PARAFAC [129, 64] method, which is one of the most used multi-way data analysis approaches.

Given the three-way array $\underline{Y}$ of dimensions $I \times J \times K$, the PARAFAC model (decomposition) [53, 126] can be expressed by the factor matrices $A(I \times F)$, $B(J \times F)$ and $C(K \times F)$, such that:

$$\underline{y}_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf}, \tag{4.2}$$

where $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$, $k = 1, 2, \ldots, K$ and R is the number of selected factors.

In principle, factorization methods handle missing data with the aim of obtaining the most accurate data model. However, once the factorization has been computed, the resulting factor matrices can be used to reconstruct the original data and missing values are then estimated by Equation (4.2). Thus, a three-way array (multi-way) without missing values $\underline{\hat{Y}} \approx \underline{Y}$ can be obtained, by making use of the information from the whole multi-way structure.

There are two main algorithms that have been studied for the computation of the PARAFAC model with missing data. The first one is based on single imputation (which is considered in the category of Expectation-Maximization approach for incomplete data) [129]. It is combined with the PARAFAC-Alternating Least Squares (ALS) algorithm, with which factors are updated until convergence. This algorithm has shown to work well for small amounts of missing data and it is very simple and fast. However, it may suffer of slow/no convergence as the amount of missing values increases [129] and it also depends on the patterns of the missing values. The second algorithm is the least squares approach known as INDAFAC (Incomplete Data PARAFAC), and it is based on fitting the PARAFAC model to the available data only by means of the computationally more expensive Levenberg-Marquadt method (modification of Gauss-Netwton algorithm for non-linear least squares problems) [75, 129]. This algorithm performs similar to PARAFAC-ALS. However, in previous studies it has shown to be more accurate and faster for large amounts of missing data problems and for SMV patterns. Recently, another least squares algorithm was proposed for PARAFAC models, CANDECOMPT/PARAFAC Weighted OPTimization (CP-WOPT) [5]. It is a scalable algorithm, which is based on direct non-linear optimization to solve the least squares problem. This algorithm has shown to work well even with 70% of missing data and it is faster than the INDAFAC method. However, this method has not been study for missing data with a SMV pattern.

The traditional PARAFAC-ALS algorithm and CP-WOPT are used in this work as means of estimation of missing values for the classification of incomplete multi-way data sets.

4.3.3 Triangulation-based interpolation imputation

Interpolation is another option for the imputation of missing values. Given a finite set of points, an interpolating function can be defined such that it should pass exactly through each point.

This function can be used to estimate values of unknown points inside the range of the originally available points.

In the case of three-way data, objects have a regular gridded 2D structure. However, when data is incomplete, this structure becomes more irregular as the number of missing values increases. One useful approach to interpolate irregularly gridded data is triangulation. When a data set is triangulated, a network or mesh of triangles is constructed with the data points at the vertices of the triangles. The mesh of triangles defines a piecewise-planar interpolating function; that is, each triangle is a piece of a plane surface. A common approach for constructing the triangulation is the Delaunay triangulation method [83]. It has favorable geometric properties that produce good results. Roughly speaking, it looks for the most equilateral-shaped triangles, ensuring that the interpolated values are influenced by sample points in the neighborhood of the query location (empty circumcircle property).

After triangulation is computed, there are different methods to obtain the interpolating function. Three of them will be analyzed and applied here.

Triangle-based Nearest Neighbor:

Nearest neighbor interpolator (proximal interpolation) is a simple method for multivariate interpolation. This algorithm searches inside the Delaunay triangulation and selects for each position the value of the nearest point (vertex) [143]. The algorithm is very simple to implement and is commonly used. Another advantage of this method is that it does not have a restricted domain as a nearest neighbor to a query point always exists. Therefore, extrapolation for points outside the initial set of points is very easy. However, the resulting interpolator is a piecewise constant function that will be continuous in those constant parts but discontinuous in the others (unless the function takes the same value in all points). The data we are dealing with here has a continuous nature. A discontinuous function will not approximate it accurately and the estimation of the missing values will be influenced by it. Therefore, the reconstructed surface will be non-smooth and this effect will be more pronounced the more missing values have to be estimated.

Triangle-based Linear Interpolation:

Linear interpolation is the most popular and widely used reconstruction method. It is simple and easy to implement, fast and results are usually not bad. After the vertices that define each triangle are obtained by Delaunay triangulation, triangles are created by drawing lines between data points [143]. Each triangle defines a plane with the tilt and elevation of the triangle determined by these three vertices. The result is a patchwork of triangular faces. Although the interpolating function is continuous, the connection of triangles by straight lines does not contribute to its smoothness. Thus, the reconstructed surface will not be very smooth. However, when applying the CMS measure, as the first step consists in smoothing the surface with a Gaussian filter, this problem can be lighten. This approach works best when data points are evenly distributed.

Triangle-based Spline Interpolation:

In this approach a smooth piecewise two-dimensional cubic (i.e. a bicubic) function is fitted to the 2D triangulated polygonal data [143, 66]. In practice, there are many approaches to accomplish this [66]. In this case each original triangle is subdivided into three sub-triangles using the centroid of the main triangle. Separate bicubic patches are then fitted to each of the three sub-triangles, which are smoothly connected providing a continuous spline-like surface approximation at all points (Clough-Tocher interpolation). This procedure is exact and smooth, with interpolated values inside the triangle having values that may be above or below

the maximum and minimum values of the vertex z-values. However, the 'nearest'and 'linear' methods are computationally more efficient than the 'spline' method.

A disadvantage of the last two interpolation methods is that they do not extrapolate to points that fall outside the convex hull from the initial set of points. Therefore, after the reconstruction of the initially incomplete data, there might be still missing points. In the case of the RMV pattern, it should not be very problematic though. Some points in the extremes of the grid could be missing but they could be reconstructed by other of the imputation methods. Another option is to use the adapted measure for missing values (see next Subsection). In this case, the measure would only have to deal with those missing values with which the interpolation method could not deal. In the case of the RMF it is more difficult, as complete rows or columns could be missing from the extremes of the grid. It could be solved as in the previous case. Moreover, the original CMS measure could be applied on the reconstructed sub-grid that falls inside the convex hull. This way, only commonly available data will be used for the computation of dissimilarities between two objects. There can be two main difficulties if the problem is solved this way, when the amount of missing values is very large. The first is that two objects do not have a coincident complete part, thus there is not information to compute dissimilarities. The other is that the amount of available columns or rows are smaller than the size of the filter, so the measure cannot be applied. However, this only happens in very rare cases.

4.3.4 Ignoring missing values in DR: adjustment of CMS measure

An alternative for dealing with missing values in the DR approach is to compute proximities on available data only. This way, there is no need for a previous pre-processing that would lead to an increase of the computational cost. However, this approach depends on the dissimilarity measure to be used i.e. the definition of each measure has to be adapted for this purpose, which is not always straightforward. In this paper, the adaptation of the CMS measure will be explained. Although the CMS measure was proposed for multi-way data in general, in this paper we will focus on three-way data only.

In this measure, missing values will be treated in the first step i.e. Gaussian filter. The idea is to use a filter that will only process the non-missing values in the analyzed window. In practice, if we have a matrix $\mathbf{Y}$ and a 2D filter kernel $\mathbf{G}$, the result of applying the filter $\mathbf{G}$ (or any other filter) at each position of matrix $\mathbf{Y}$ is defined as:

$$\mathbf{Y}'(u,v) = \sum_{k=-P}^{P} \sum_{l=-P}^{P} \mathbf{Y}(u-k, v-l) \cdot \mathbf{G}(k,l) \tag{4.3}$$

where $2P+1$ is the size of the filter in both the horizontal and vertical directions of the convolution kernel $\mathbf{G}$.

So, suppose we are analyzing a part of the data with q missing values. The filter is only applied to the $(2P+1)^2 - q$ non-missing values. In such case, as the number of summed values are less, the filtering result S for the analyzed position will not correspond to that if the data was complete. Therefore, a normalization should be applied. This normalization can be done in the following way:

$$S' = \frac{S}{(2P+1)^2 - q} \cdot (2P+1)^2 \tag{4.4}$$

If S' is used as the filtering result, instead of S, we are doing an implicit estimation of the missing values. That is, we are assuming that each missing value contributes in the filtering result S' with a value of $\frac{S}{(2P+1)^2-q}$. However, this can be considered a drawback of this approach, since the implicit estimation of the missing value can change according to the position of the filter on the 2D matrix.

When the amount of missing values is large, it could happen that all values in the window of the analyzed point are missing. In such situation, the previous adaptation does not work, it assigns NaN to the analyzed point. In this case, the idea is then to omit these points when objects are compared.

4.3.5 Setting missing values to zero

Setting missing values to zero used to be a common approach for dealing with missing data. However, it is very dangerous/inaccurate for the analysis of data. It is the same as if any random value, unrelated to the nature or origin of the data, was used to fill the incomplete data. Nevertheless, this also depends on the patterns of the missing points and the purpose of the analysis. In our case, in the presence of a SMV pattern, which is the same for every object, the missing part should not be treated in the computation of the dissimilarity value. Theoretically, the measurements were not taken because they make not sense, therefore there is no reason to impute them. Thus, if they are set to zero or any constant value c, it will not affect the dissimilarity values. It is like all objects are equal in that part/it just does not exist, so it does not influence the dissimilarities. Not special treatment should be needed in this case. This method will be applied here for data with these type of pattern of missing values, i.e. SMV.

4.4 Experiments and discussion

The main goal of the following experiments is to evaluate how the explained imputation methods and the proposed adaptation of the CMS measure contribute to the DR-based classification of incomplete multi-way data. Moreover, we study how the methods behave for different patterns of missing data. With this purpose, 4 different three-way continuous data sets are used. Data sets and the experimental setup will be described in the next subsections.

4.4.1 Data sets

The first data set is private and it comes from 1200 patches of 1024×1024 pixels of 36 colon tissue slides from Atrium hospital in Heerlen, The Netherlands. Patches were filtered with Laplace filters in 90 different scales using $\sigma = 2.\hat{}[0.1 : 0.1 : 9]$. The log-squares of the results are summarized in 60 bin normalized histograms with bin centers $[-50 : 1 : 9]$. Thus, a 90×60 array is obtained for every patch, leading to a three-way array of $1200 \times 90 \times 60$. The patches are labeled in two classes: Normal and Tumor. A representation set of 550 prototypes was randomly chosen from the training set.

The second data set consists in metabolite data containing HPLC measurements of commercial extract of St. John's wort. The description has been taken from the literature [3] for a better understanding of this paper. HPLC-PDA (HPLC with photo-diodo array detection) profiles were obtained from 24 different examples of St. Johns' wort from several continents (classes): Africa (8 objects), Asia (6 objects), Europe (45 objects) and North America (30 objects). The number of objects from each continent varies between 2 and 12. HPLC-PD profiles replicates of three or four hour for each example were obtained. The chromatography was monitored between $190nm$ and $620nm$. Two regions of the chromatographic data were chosen for the analysis and reduced to steps of $3nm$ in the Ultraviolet-mode (UV) $(260-550nm)$ (97 features) and steps of 1.32 seconds in the retention time mode (549 features). The final three-way data has a size of $89 \times 97 \times 549$.

The third and fourth data sets are from public domains and they are both obtained by Fluorescence spectroscopy. Fluorescence excitation-emission measurements are used because they are known to reflect important properties of the fermentation process. For the first data set, several examples from many batches were obtained and measured on an at-line multi channel

fluorescence detection system. Each fluorescence landscape from the sensor is obtained using 15 excitation filters in the range from 270 to 550 nm with a spectral resolution of 20 nm, and 15 emission filters range from 310 to 590 nm, with a spectral resolution of 20 nm too. The enzyme activity is related to the quality of processes [96]. Thus, the classification problem consists in determining the quality (Low or High) of a process according to that enzyme activity. As the fluorescence data is a function of both excitation and emission filters, each object is a 2D matrix. When several objects are analyzed, they will form the third direction of a three-way array. This data set is then composed of a $323 \times 15 \times 15$ training set and a $53 \times 15 \times 15$ for test.

The other data set comes from raw pork from fresh hams obtained in the local market, whereas Parma hams were from a processing plant in Parma, Italy. Parma ham ages ranges from salted (3 months) to matured (11 and 12 months) and further to aged (15 and 18 months). A total of 67 meat examples were submitted to duplicate measurements of surface autofluorescence spectroscopy. Examples were measured at 15 excitation wavelengths (270550 nm), and 15 emission wavelengths (310590 nm), both with a step of 20 nm. The emission wavelengths were shifted by 40 nm from each excitation wavelength applied. Before analysis of the data, the excitation wavelengths above 470 nm, and the emission wavelengths below 350 nm were removed. Thus, the dimension of the data set was reduced from $67 \times 15 \times 15$ to $67 \times 11 \times 13$ ($objects \times excitation \times emission$) [82].

4.4.2 Experimental setup

Experiments were designed as follows:

- In the first two data sets (Colon Cancer and St Johns) there are no missing values, but these were generated artificially to test the methods. For each data set, 10 new data sets were first created by inserting various amounts ($1-5, 10, 20, 30, 50, 70\%$) of missing values in the whole data set. Thus, all objects have the same probability of having missing values and the amount per object is completely random. This procedure was done for RMV and FMV (rows and tubes) patterns. So, at first we have generated in total 30 new data sets from the original ones. To avoid that results are influenced by a specific random pattern and determine a general behavior, we repeated the previous configurations 5 times for each of the two data sets.

- In the case of autofluorescence data i.e. Enzyme and Parma, missing values were originally introduced as part of a pre-processing. Before analysis, values for emission wavelengths below excitation wavelengths should be deleted. Those measurements have no sense. Therefore, there is a systematic pattern of missing values for all objects, corresponding to those of the specified excitation/emission wavelengths.

- Most imputation methods and the modified CMS measure explained in Section 4.3 are evaluated on all data sets. For non-systematic patterns of missing data, it makes no much sense to fill the holes with zeros (see Section 4.3.5). Therefore, this will only be evaluated on data with SMV pattern, in this case autofluorescence. In the experiments for data with missing rows or tubes (Colon and St Johns), the original CMS measure is applied on the reconstructed sub-grid that falls inside the convex hull. This way, the experiments are in the same conditions as for the other reconstruction methods.

- There are different methods for the selection of the number of components in the factorization-based methods [126]. However, as in our case the interest is to reconstruct the original data as best as possible we will use the residuals evaluation criteria. This consists on trying to find a minimum sum of squares of errors in the approximation of the non-missing values. In all cases, classifiers performed better for those models that fulfilled the previous criteria.

- Results are given in terms of classification error. For the different data sets, experiments were carried out differently. For small data sets (Parma ham and St John's), classification errors were obtained in a 10 times k-fold cross-validation (CV). The Enzyme data comes with training and test set. In the case of Colon data, 10 different training (90%) and test (10%) sets were randomly chosen and the error values were averaged. Experiments for the 5 repetitions of each of the configurations were averaged.

- The Regularized linear discriminant classifier was used on the dissimilarity space. The Linear Discriminant Classifier (LDC) assumes that the classes are described by multi-normal distributions with the same covariance matrices. Since for $n \times n$ dissimilarity representations the estimated covariance matrix S is singular, its inverse cannot be determined. Therefore, its regularized version is used instead (RLDC). Regularization takes care that the inverse operation is possible by emphasizing the diagonal values (variances) of the matrix S with reference to the off-diagonal elements (covariances) [94]. To find the regularization parameters of RLDC, an automatic regularization (optimization over training set by cross-validation) process was done.

- In the DR approach, for the small data sets, the representation set has the same size of the training set obtained in each fold of the cross-validation procedure. For the Colon data set, the representation set was randomly chosen for each generated training set. A total of 550 prototypes were used, leading to a dissimilarity matrix of 1000×550. In the case of the Enzyme data set, a representation set of 100 prototypes was also randomly chosen from the training set, thus the dissimilarity matrices have a size of 323×100 and 53×100 for training and test sets, respectively.

The following tables (Tables 4.1 - 4.6) summarize the classification errors of two of the data sets after reconstructing the data with the methods explained in Section 4.3. The classification errors on the complete data sets are used as a baseline for the comparison.

Table 4.1: Classification errors of Colon Cancer data set after treatment of missing values with different methods. Results for different percents of random missing data are shown. The baseline error with the complete data is 0.095

Methods	Random missing values (%)									
	1	2	3	4	5	10	20	30	50	70
Averaging rows	0.11	0.12	0.14	0.14	0.14	0.15	0.2	0.23	0.27	0.31
Averaging cols	0.11	0.13	0.13	0.13	0.13	0.16	0.21	0.22	0.29	0.32
Averaging objects	0.11	0.11	0.11	0.11	0.12	0.12	0.13	0.13	0.15	0.18
PARAFAC	0.27	0.3	0.32	0.32	0.32	0.33	0.34	0.32	0.36	0.39
CP-WOPT	0.18	0.2	0.22	0.22	0.26	0.26	0.28	0.28	0.36	0.36
Interp Linear	0.12	0.12	0.12	0.12	0.13	0.14	0.14	0.14	0.15	0.19
Interp KNN	0.11	0.12	0.12	0.13	0.13	0.13	0.15	0.15	0.18	0.22
Interp Spline	0.12	0.12	0.12	0.13	0.13	0.13	0.14	0.14	0.16	0.19
Adapted CMS directly	0.12	0.13	0.14	0.14	0.14	0.17	0.19	0.21	0.28	0.30

Table 4.2: Classification errors of Colon Cancer data set after treatment of missing values with different methods. Results for different percents of complete tubes missing are shown. The baseline error with the complete data is 0.095

	Complete tubes missing (%)									
Methods	1	2	3	4	5	10	20	30	50	70
Averaging rows	-	-	-	-	-	-	-	-	-	-
Averaging tubes	0.11	0.13	0.13	0.13	0.13	0.15	0.16	0.19	0.22	0.28
Averaging objects	0.24	0.11	0.23	0.21	0.21	0.25	0.22	0.26	0.26	0.29
PARAFAC	0.30	0.30	0.31	0.31	0.31	0.31	0.31	0.32	0.33	0.40
CP-WOPT	0.2	0.2	0.2	0.26	0.22	0.24	0.24	0.26	0.28	0.36
Interp Linear	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.11	0.14	0.18
Interp KNN	0.12	0.12	0.12	0.12	0.12	0.12	0.13	0.13	0.14	0.17
Interp Spline	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.13	0.17
Adapted CMS directly	0.14	0.14	0.14	0.14	0.13	0.16	0.17	0.19	0.24	0.29

Table 4.3: Classification errors of Colon Cancer data set after treatment of missing values with different methods. Results for different percents of complete rows missing are shown. The baseline error with the complete data is 0.095

	Complete rows missing (%)									
Methods	1	2	3	4	5	10	20	30	50	70
Averaging rows	0.13	0.13	0.14	0.14	0.14	0.15	0.15	0.17	0.19	0.23
Averaging tubes	-	-	-	-	-	-	-	-	-	-
Averaging objects	0.11	0.12	0.1	0.12	0.13	0.13	0.13	0.15	0.15	0.17
PARAFAC	0.31	0.31	0.31	0.32	0.32	0.32	0.32	0.34	0.38	0.4
CP-WOPT	0.19	0.24	0.2	0.22	0.24	0.22	0.22	0.28	0.28	0.34
Interp Linear	0.11	0.12	0.12	0.12	0.11	0.12	0.12	0.14	0.16	0.21
Interp KNN	0.12	0.13	0.13	0.13	0.13	0.14	0.16	0.16	0.19	0.22
Interp Spline	0.11	0.11	0.11	0.11	0.11	0.11	0.12	0.14	0.16	0.22
Adapted CMS directly	0.14	0.15	0.15	0.15	0.15	0.14	0.15	0.18	0.19	0.25

Table 4.4: Classification errors of St Johns data set after treatment of missing values with different methods. Results for different percents of random missing data. The baseline error with the complete data is 0.02

	Random missing values (%)									
Methods	1	2	3	4	5	10	20	30	50	70
Averaging rows	0.05	0.07	0.10	0.15	0.17	0.25	0.28	0.30	0.33	0.36
Averaging cols	0.07	0.14	0.21	0.23	0.26	0.28	0.34	0.37	0.41	0.43
Averaging objects	0.02	0.02	0.03	0.03	0.03	0.03	0.04	0.06	0.15	0.22
PARAFAC	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
CP-WOPT	0.03	0.03	0.03	0.03	0.04	0.04	0.10	0.10	0.10	0.12
Interp Linear	0.04	0.06	0.10	0.10	0.10	0.10	0.10	0.11	0.11	0.11
Interp KNN	0.02	0.02	0.02	0.02	0.02	0.04	0.06	0.09	0.12	0.15
Interp Spline	0.04	0.07	0.10	0.10	0.10	0.10	0.10	0.11	0.11	0.12
Adapted CMS directly	0.02	0.02	0.02	0.03	0.03	0.04	0.05	0.07	0.17	0.26

Analyzing the overall behavior of all methods in the first six tables we can appreciate that:

As explained before, the averaging method is applied in three different directions: rows, tubes and objects (columns). It can be observed that for small amounts of missing values, when averaging is performed by rows or tubes, the accuracy of the classifier does not deviate much from the baseline error. However, the error increases gradually as the amount of missing data is larger. This behavior is to be expected due to the formulation of the method itself. The fact that a missing point can be approximated by averaging very far away values, increases the possibilities of introducing noise in the data. Therefore, affecting the performance of classifiers. As it was previously mentioned, this method fails when averaging in the contrary direction for FMV i.e. rows or tubes, problems. If results when averaging by objects are analyzed, it can be observed that most of the time they perform better than the other averaging version. This

Table 4.5: Classification errors of St Johns data set after treatment of missing values with different methods. Results for different percents of complete tubes missing are shown. The baseline error with the complete data is 0.02

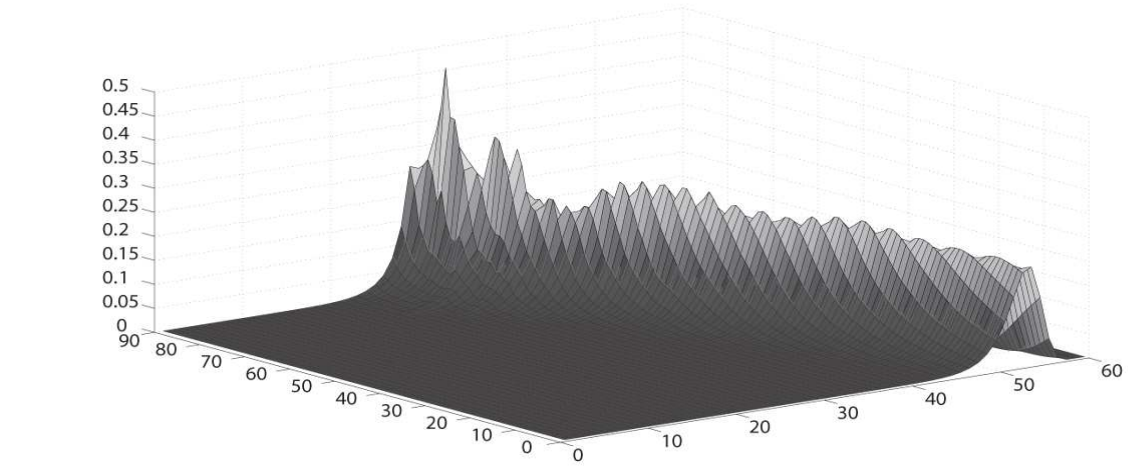
	Complete tubes missing (%)									
Methods	1	2	3	4	5	10	20	30	50	70
Averaging rows	0.04	0.04	0.07	0.08	0.11	0.16	0.25	0.29	0.32	0.33
Averaging tubes	-	-	-	-	-	-	-	-	-	-
Averaging objects	0.02	0.02	0.02	0.02	0.03	0.04	0.07	0.11	0.20	0.26
PARAFAC	0.05	0.05	0.05	0.05	0.05	0.06	0.05	0.05	0.05	0.06
CP-WOPT	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.05
Interp Linear	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.03
Interp KNN	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.06	0.13
Interp Spline	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.03
Adapted CMS directly	0.03	0.03	0.03	0.04	0.03	0.03	0.04	0.08	0.24	0.41

Table 4.6: Classification errors of St Johns data set after treatment of missing values with different methods. Results for different percents of complete rows missing are shown. The baseline error with the complete data is 0.02

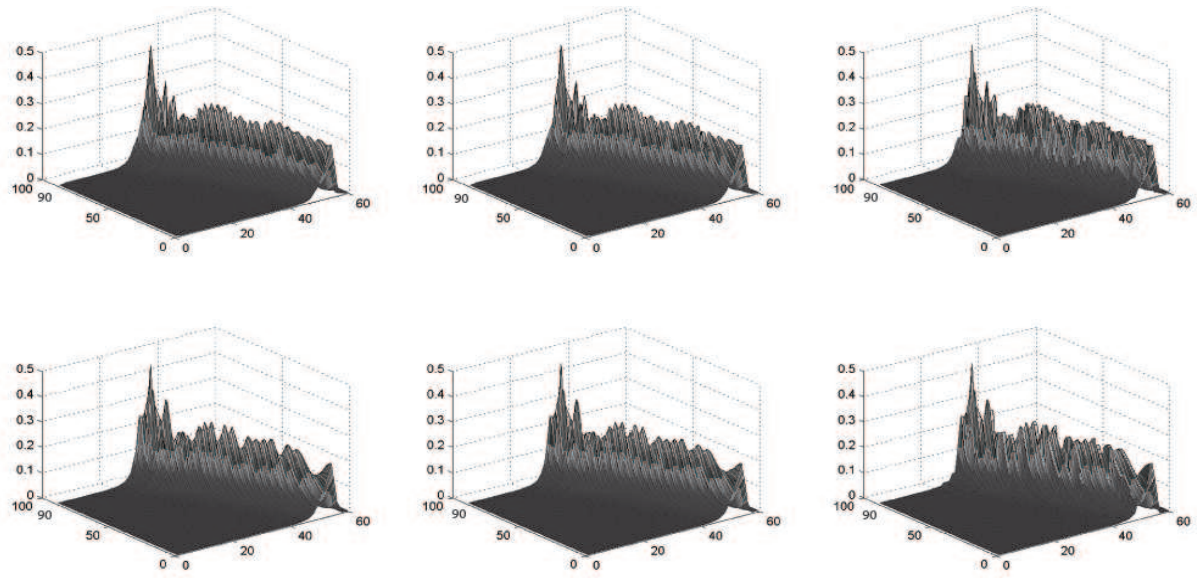
	Complete rows missing (%)									
Methods	1	2	3	4	5	10	20	30	50	70
Averaging rows	-	-	-	-	-	-	-	-	-	-
Averaging tubes	0.04	0.07	0.11	0.15	0.18	0.26	0.27	0.27	0.3	0.29
Averaging objects	0.02	0.02	0.0	0.03	0.03	0.04	0.07	0.12	0.18	0.26
PARAFAC	0.05	0.05	0.05	0.04	0.05	0.06	0.07	0.11	0.13	0.23
CP-WOPT	0.04	0.04	0.04	0.04	0.05	0.07	0.07	0.09	0.12	0.19
Interp Linear	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.04	0.15
Interp KNN	0.02	0.02	0.02	0.02	0.02	0.02	0.04	0.05	0.10	0.22
Interp Spline	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.04	0.14
Adapted CMS directly	0.02	0.03	0.03	0.03	0.03	0.06	0.13	0.22	0.32	0.46

might seem a bit counterintuitive because values from different classes are used to approximate the missing value. However, this behavior can be frequent due to the nature of the analyzed data, where patterns of classes differ by small variations in some features. Moreover, objects can consist of surfaces with very large peaks and sudden valleys. Therefore, the values of a position (j, k) for all objects in the three-way array could be more similar than the values of feature j or k for one object.

Factorization-based methods seem to work well when they converge, but this is not the case when convergence is not reached. It can be observed in Colon data set that the performance of the classifier is bad for all patterns of missing values. It is actually the worst result. In this case, both algorithms took long to converge and for large amounts of missing data convergence was never reached. However, when these methods converge, like in the case of St John's, they perform well in general. There is a slight improvement of CP-WOPT based results over those of PARAFAC-ALS, specially for large amounts of missing data, as expected. It has to be noticed that for large amounts of missing values these methods are stable. Nonetheless, even when the stability of these methods (when they converge) for different amounts of missing data is very attractive, their slow/no convergence problem is a strong drawback when comparing methods to reconstruct the missing data.



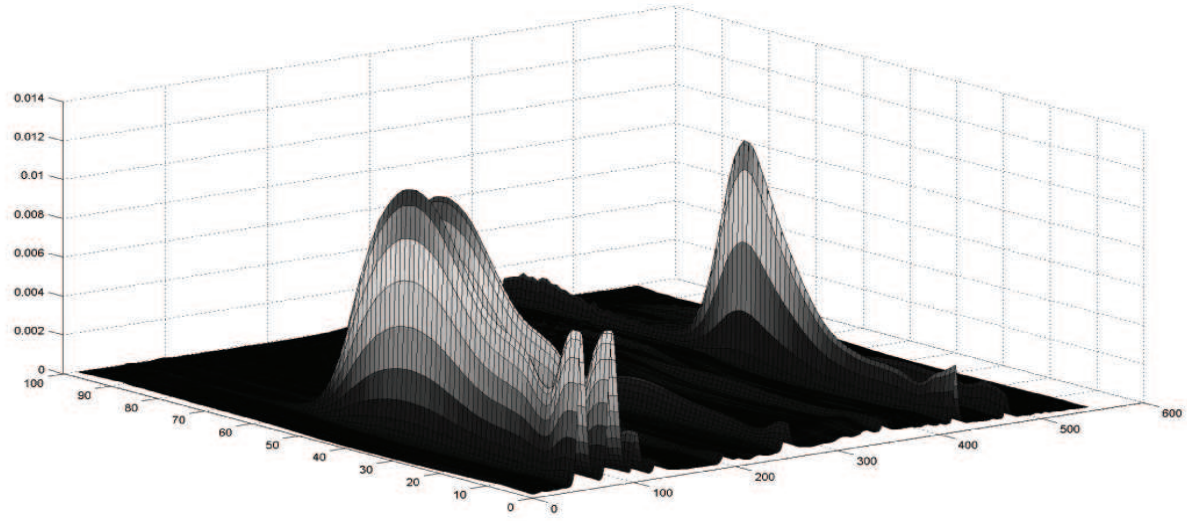
(a)



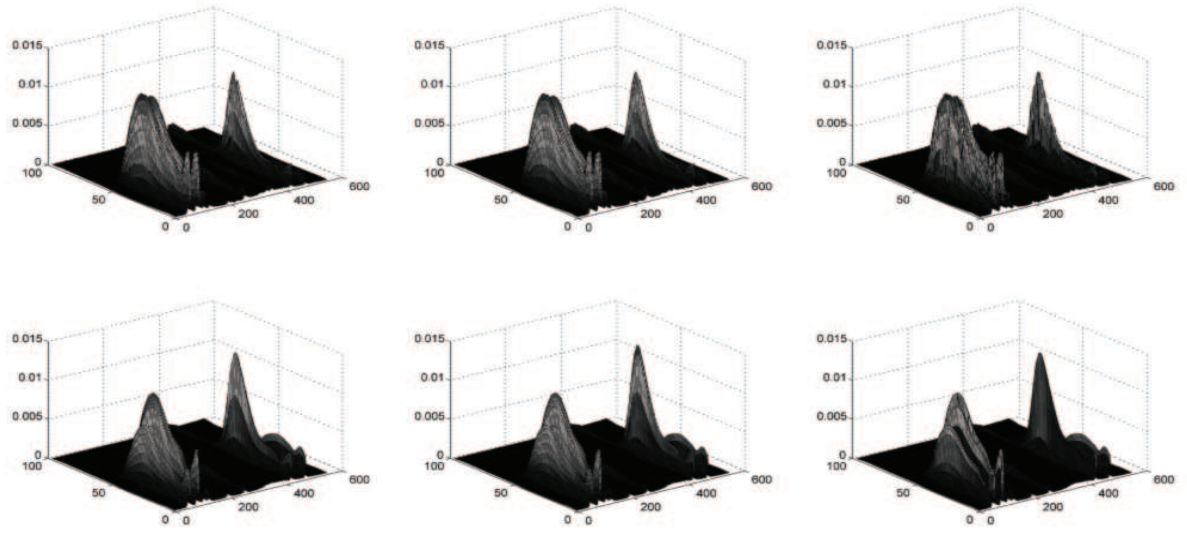
(b)

Figure 4.3: Surfaces of a complete and reconstructed object (30% missing data) from Colon data set. Figure (a) corresponds to the complete surface of an object of this data set. Figure (b) corresponds to the reconstructed surfaces of the object by the three triangle-based interpolation methods: Linear Interpolation, Spline interpolation and KNN interpolation, which appear in this same order. The top and bottom figures of (b) correspond to the reconstruction of RMV and RMF, respectively.

As an overall behavior, performance of classifiers is good after computing dissimilarities on the reconstructed data by interpolation. The minimum classification error in the tables is usually achieved by applying one of these methods. This error increases slowly as the amount of missing values is larger, but in general they are more stable than other methods e.g. averaging, modified CMS. It can be observed that even for 70% of missing values, the lower classification errors are achieved by applying the linear and spline interpolation methods. This could be because triangulation can efficiently represent the same surface as other methods do but with fewer data points [66]. KNN interpolation however, seems to be less stable; comparing it with the other interpolation methods, KNN's accuracy sometimes decreases for more than 30% of



(a)



(b)

Figure 4.4: Surfaces of a complete and reconstructed object (30% missing data) from St John's data set. Figure (a) corresponds to the complete surface of an object of this data set. Figure (b) corresponds to the reconstructed surfaces of the object by the three triangle-based interpolation methods: Linear Interpolation, Spline interpolation and KNN interpolation, which appear in this same order. The top and bottom figures of (b) correspond to the reconstruction of RMV and RMF, respectively.

missing data. This is to be expected, as with large amounts of missing data, there is less information to use for the approximation. Therefore, the nearest neighbor of the analyzed point can be somehow far from it. An interesting issue is that results for spline and linear interpolation are similar, although we expected that there would be a superiority of the spline method. In this case, the two used data sets have a surface with few rough changes and large flat areas. Therefore, the linear method can interpolate this surface as accurate as the spline method (See Figures 4.3 and 4.4). In these figures it can be observed how for 30% of missing data, the KNN interpolator is starting to generate more discontinuous-flattened surfaces. It

should also be noticed that, despite linear and spline interpolation methods are not capable of imputing values outside the convex hull, and that the measure was only applied on the data inside this convex hull, the accuracy of classifiers was not very affected. However, it should not be the case if important discriminative features are left out from the defined convex hull.

Let us analyze the adapted CMS measure directly applied on the incomplete data. It has to be remarked that for small amounts of missing data $(1 - 5\%)$, the performance of the classifier is comparable with that of the best imputation methods. In fact, for St John's data set, the baseline classification error is reached. A very attractive characteristic of the modified measure is that without a pre-processing step i.e. imputation, approximation, it has shown to work well with small amounts of missing data. Therefore, it could be a good option for these types of problems. Good performances can be obtained without the need of the extra computational cost of the imputation process. However, when the amount of missing values is large (usually above 10%), the classifier seems to lose stability and its performance worsens drastically as the number of missing values increases. This could be explained by the fact that when there are many contiguous windows missing, the Gaussian filter cannot deal with it. Too much information is lost and the idea of derivatives is kind of pointless. In these cases, the use of an imputation method e.g. interpolation is recommended.

In general, for the two analyzed patterns of missing values, that is RMV and FMV (rows or tubes), all methods behaved similar. It can be observed that the type of pattern of the missing points did not have a strong influence in the performance of the methods. The main disturbing factor in the performance of the methods was the amount of missing values.

In Table 4.7, the results for Enzyme and Parma ham data sets are shown. Theoretically, it does not make sense to impute missing values with this pattern for the computation of the dissimilarity measure (See Section 4.3 for more). As these values do not actually exist, the imputed values can introduce a considerable amount of noise, such that performance of classifiers is affected. However, we applied the imputation methods on the Autofluorescence data sets to show that this problem holds in practice. The best performance of classifiers is obtained with the modified CMS measure. As explained in Section 4.3.4, the measure does not take into account the missing values in this case; which is how it should be done according to the nature of the missing data.

Table 4.7: Classification errors of Enzyme and Parma ham data set after treatment of systematic missing values with different methods.

Methods	Data sets	
	Enzyme	Parma ham
Averaging rows	0.15	0.04
Averaging columns	0.13	0.06
PARAFAC	0.09	0.08
CP-WOPT	0.09	0.09
Interp KNN	0.09	0.04
Filling with zeros	0.06	0.024
CMS directly	0.06	0.023

It can be observed that similar results where obtained by filling the holes with zeros. This is to be expected. With this approach, the only zeros that are considered in the computation of the measure are those which are close to the borders of the complete part. The rest are canceled in the process. Therefore, this is almost the same as applying the adapted CMS measure along. In such case, it will be less computationally expensive to apply the modified measure directly on the data.

4.5 Conclusions

We have investigated two main approaches with the aim of dealing with the problem of missing values in the classification of multi-way data. The study was based on the Dissimilarity Representation approach, which consists in building classifiers on a space where objects are represented (introducing context information) by their dissimilarities. As a first attempt, imputation techniques were applied to reconstruct the data before dissimilarities were computed. These were all evaluated on four data sets, where missing attributes had different patterns. Most methods performed well for small amounts of missing values. However, triangulation-based interpolation methods have shown to be the most stable. When applying these methods, classifiers performed relatively well for large amounts of missing data; in some cases even 70% of the data was missing.

Although these methods seem to be suitable for the problem at hand, they imply an extra computational cost. Therefore, we studied as a second approach, the possibility of computing dissimilarities with the available data only. In this paper, as we experimented on continuous multi-way data, the Continuous Multi-way Shape measure was used. In order to deal with the presence of missing attributes, a modification of this measure was proposed. This approach, with this particular measure, has shown to be the best option for systematic missing data problems. Moreover, for the other patterns of missing values, it works well when they are present in small amounts (up to 10%). From that point on, classifiers performance deteriorates increasingly. We can then conclude that this approach is suitable for small amounts of missing data. However, it has the disadvantage that a modification of the applied dissimilarity measure is usually needed and it is not always straightforward. Although experiments were carried out on three-way data only, both techniques can be extended to higher-order representations of objects. In spite that the studied imputation techniques were evaluated for dissimilarity-based classification, they can be used for classification purposes in general, or any other type of analysis that requires a minimum reconstruction error.

Chapter 5

Future Perspectives

5.1 Towards the application in other research areas

The main approaches introduced in this thesis are oriented to the classification of multi-way spectral/continuous data sets in general. However, most experimentations are based on chemical spectral data. Data sets with similar characteristics can be found in many other research fields e.g. biometrics [48], in neuroinformatics [2, 80], physics [84]. Therefore, it would be very interesting to investigate the generality of the proposed approach in these applications. Specific background knowledge of each problem could be added to the 2D dissimilarity measures developed in this thesis, in order to have a better description of the problem at hand.

5.2 Clustering for multi-way data

The unsupervised classification of individuals/objects into groups is mainly done by cluster analysis. It differs with respect to the supervised classification in that a number of objects is available, but there is no previous knowledge on what the classes are. Thus, the goal of cluster analysis is to partition the data set such that similar objects are grouped together and dissimilar objects fall into different groups [57].

An important step in cluster analysis is the selection of a distance measure. These methods start by computing distances or similarities between objects in order to cluster them, thus the applied measure will influence the shape of clusters. Therefore, if distances are defined such that users can supply their expert knowledge e.g. shape information for spectral data analysis, the clustering procedure could be tailored and "better groupings" could be obtained. Moreover, although traditional clustering algorithms are based on analyzing distances directly, it is intriguing to investigate clustering analysis on the dissimilarity space.

In the case of two-way data, many procedures exist for cluster analysis. However, there is a limited number of techniques for clustering three-way profile data [3, 64]. With the CMS (Continuous Multi-way Shape) measure [103] proposed in this thesis for multi-way continuous data, a new alternative for clustering multi-way objects can be created. The CMS measure takes into account the multi-dimensional structure of objects, but still allows for the use of traditional clustering algorithms e.g. k-means, expectation-maximization, hierarchical clustering [57].

There are, however, data sets for which there is only continuity in one of the directions/ways of the object e.g. chromatography-mass spectrometry data. For this type of data it makes no sense to use the CMS measure. Therefore, we proposed the 2Dshape measure [100], which uses the information on the two directions of 2D objects, but it can be adapted to only measure continuity in the continuous direction. In this case, traditional clustering algorithms could be used on three-way data as with the CMS measure.

Another alternative can be to slice the three-way data in the non-continuous direction, and

use the first part of the 2Dshape measure or a 1D shape-based measure for the continuous direction. Afterwards, m clusterings or clustering algorithms can be applied on the objects represented by the m slices and use a clustering ensemble method to obtain a consensus clustering [136].

Slicing three-way data can be understood as having different representations for the same objects. Therefore, the clustering ensemble approach fits to this problem. Indeed, a previous study [137] proposes a similar idea (not using shape-based measures for continuous data) and combined dendrograms resulting from hierarchical algorithms. Nevertheless, the clustering ensemble approach offers a large variety of alternatives that could be used in this problem.

5.3 Dissimilarity Representation for Regression

Regression and classification can be seen as similar problems. For both of them, a set of examples comprising n observations are given. The goal is to obtain a model, such that we are able to predict the value of a target variable C, given the n-dimensional vector of a new example x. The main difference between these two tasks is that in regression, the target variable is continuous, and for classification C is categorical i.e. classes instead of numbers. Due to this, it has not been difficult to find that in some areas, classification methods have been adapted to be used for regression problems and vice versa [130, 81, 13, 19]. Therefore, we think that the advantages of the Dissimilarity Representation approach for classification can also be extended to regression problems; the introduction of context/data structure information should also help for a better mathematical modeling of the relation between the independent variables and the properties of the examples. In a recent bibliography search, we realized that there are already some initial studies on this issue[159].

5.4 Dissimilarity Representation for non-continuous multi-way data

Existent and potential multi-way data can be found in numerous disciplines, with diverse types and characteristics. Examples of these applications are Psychometrics [65], Chemistry [47], Social Networks [127], Bioinformatics [160]. For example, with the goal of classifying types of users and to improve personalized web searches, a three-way array can be defined as: $users \times queries \times webpages$. Such a structure, containing the relation between the different factors can be more informative, allowing for a better analysis of the data. The use of the DR approach in these data sets would bring the same benefits as for continuous multi-way data. The key issue, once again, is to define a suitable dissimilarity measure for the problem at hand. As much context information is included, the better representation could be obtained. Just the fact of having to design a dissimilarity measure, for any of the possible multi-way data, will constitute research topics themselves.

A quite common data type, different to the three-way profile data, are known as three-way (dis) similarities data sets. This type of data is sometimes easier to find, as for some applications it can be more natural to define a (dis)similarity between objects than to define features explicitly. For some research areas like psychology, food research, marketing, it is common to ask subjects to judge the similarity between two or different types of stimuli (objects) e.g. the taste of some types of food, the visual similarity between two posters. If there are several judges and stimuli, we have a three-way array with the form $stimuli \times stimuli \times judges$ [64]. Sometimes these analyses are done by just a single expert but in different occasions along time, places, thus the third mode changes.

In some literature, this type of data sets is also referred to as quantification matrices. They are used as a way of simultaneously analyzing variables of different nature but measured on the

same objects. Objects are represented by their dissimilarities according to each variable. Once all objects are represented in terms of (dis) similarities, all variables can be analyzed together despite they may have a different nature [59]. In general, this type of data sets have been analyzed with three-way multi-dimensional scaling techniques [27, 21] with exploratory analysis purposes.

Alternatively, if we are in presence of a classification problem, the dissimilarity space may be an interesting approach. However, for such data the issue is not to find a (dis) similarity measure as in the previous cases. The problem now is how to go from the three or multi-way dissimilarity data to a single matrix such that a dissimilarity space can be built. Would it be enough to just concatenate or do a weighted combination of the dissimilarity matrices obtained for each variable along the third direction? In cases where there is an ordering/time information in the third direction, how could this be taken into account when obtaining the final dissimilarity matrix?

This thesis is an initial study on the potentialities of the dissimilarity representation for the classification of continuous multi-way data, but it does not exhaust the benefits and application of this approach to other types of problems in multi-way data analysis. The possibility of including context information on the problem at hand should be as important as to analyze the multi-way data structure (dependencies between the modes). We think that these issues, together with the development of new approaches for the classification of multi-way data, will receive much more attention in the future.

Bibliography

[1] Abraham C, Biau G, Cadre B. On the kernel rule for function classification. Annals of the Institute of Statistical Mathematics 2006; **58**(3): 619–633.

[2] Acar E., Aykut-Bingol C., Bingol H., Bro R., Yener B. Multiway analysis of epilepsy tensors. Bioninformatics 2007; **23**: i10–i18.

[3] Acar E., Bro R., Schmidt B. New exploratory clustering tool. J.Chemom. 2008; **22**: 91–100.

[4] Acar E., Yener B. Unsupervised Multiway Data Analysis: A Literature Survey. IEEE Transactions on Knowledge and Data Engineering 2009; **21**: 6–20.

[5] Acar E, Dunlavy D. M, Kolda T. G, Mrup M. Scalable Tensor Factorizations for Incomplete Data. Chemometrics and Intelligent Laboratory Systems 2011; **106**(1): 41–56.

[6] Aguilera A.M, Ocaña F.A, Valderrama M.J. An approximated Principal Component Prediction model for continuous time stochastic processes. Appl. Stochastic Models & Data Anal. 1997; **13**: 61–72.

[7] Allen G. Sparse Higher-Order Principal Components Analysis. In: Proceedings of the 15th International Conference on Artificial Intelligence and Statistics AISTATS 2012, vol 22, Journal of Machine Learning Research: W& CP 22: La Palma, Canary Islands, 2012, 27–36.

[8] Amigo Jose M., Skov T., Bro R., Coello J., Maspoch S. Solving GC-MS problems with PARAFAC2 Trends in Analytical Chemistry 2008; **27**(8): 714–725.

[9] Andersson C. A, Bro R. The N-way toolbox for MATLAB. Chemometrics and Intelligent Laboratory Systems 2000; **52**(1): 1–4. http://www. models.kvl.dk/source/nwaytoolbox/.

[10] Andersen C. M, Bro R. Practical aspects of PARAFAC modeling of fluorescence excitation-emission data. J. Chemom. 2003; **17**: 200–215.

[11] Arancibia J.A., Boschetti C.E., Olivieri A.C., Escandar G.M. Screening of oil samples on the basis of excitation-emission room-temperature phosphorescence data and multiway chemometric techniques. Introducing the second-order advantage in a classification study. Anal Chem. 2008; **80**: 2789–2798.

[12] Ballabio D., Consonni V., Todeschini R. Classification of multiway analytical data based on MOLMAP approach. Analytica Chimica Acta. 2007; **605**(2): 134–146.

[13] Barker M., Rayens W. Partial Least Squares for discrimination. J.Chemom. 2003; **17**: 166–173.

[14] Bartosch T., Seidl D. Spectrogram analysis of selected tremor signal using short-time Fourier transform and continuous wavelet transfrom. Annali di Geofisica 1999; **42**(3): 497–506.

[15] Becker H., Comon P., Albera L., Haardt M., Merlet I. Multiway space-time-wave-vector analysis for source localization and extraction. In: EUSIPCO 10, XVIII European Signal Processing Conference, Aalborg, Denmark, 2010, 23–27.

[16] Benbrahim M., Daoudi A., Benjelloun K., Ibenbrahim A.. Discrimination of seismic signals using artificial neural networks. In: C. Ardil (Ed.), WEC (2), Enformatika, anakkale, Turkey, 2005, 4–7.

[17] Benítez M. C., Ramírez J., Segura J. C., Ibáñez J. M., Almendros J., García-Yeguas A., Cortés G.. Continuous HMM-based seismic-event classification at Deception Island, Antarctica. IEEE Transactions on Geoscience and remote sensing 2007; **45**(1): 138–146.

[18] Biau G, Bunea F, Wegkamp M.H. Functional classification in hilbert spaces. IEEE Transactions on Information Theory 2005; **51**(6): 2163–2172.

[19] Bibi S., Tsoumakas G., Stamelos I., Vlahavas I. Regression via classification applied on software detect estimation. Expert Systems with Applications 2008; **34**: 2091–2101.

[20] Billauer E. Peak detection method for Matlab. http://billauer.co.il/peakdet.html/ [2011].

[21] Borg I., Groenen P. Modern multidimensional scaling: Theory and applications (2nd ed.). Springer, New York, 2005.

[22] Xu Y, Gong F, Dixon S, Brereton R, Soini H, Novotny M, Oberzaucher, E, Grammar, K, Penn, D. Application of dissimilarity indices, principal coordinates analysis and rank tests to peak tables in metabolomics of the gas chromatography mass spectrometry of human sweat. Anal. Chem. 2007; **79**(15): 5633–5641.

[23] Brereton R.G. Chemometrics for Pattern Recognition Wiley, 2009.

[24] Bro R. Multi-way Analysis in the Food Industry. Models, Algorithms, and Applications. PhD Thesis: Amsterdam, Netherlands, 1998.

[25] Schölkopf B. Support Vector Learning. PhD thesis: Munich, Germany, 1997.

[26] Busscher N, Kahl J, Andersen J. O, Huber M, Mergardt G, Doesburg P, Paulsen M, Ploeger A. Standardization of the Biocrystallization Method for Carrot Samples. Biological Agriculture and Horticulture 2010; **27**: 1–23.

[27] Carroll, J.D., Chang J. Analysis of individual differences in multidimensional scaling via an N-way generalization of EckartYoung decomposition. Psychometrika (1970); **35**: 283–319

[28] Cardot H, Ferraty F, Sarda P. Functional linear model. Statist. Probab. Lett. 1999; **45**(1): 11–22.

[29] Cattell, R.B. The three basic factor-analytic research designs-their interrelations and derivatives. Psychological Bulletin (1952); **49**: 499–452

[30] Cérou F, Guyader A. Nearest neighbor classification in infinite dimension. ESAIM: Probability and Statistics 2006; **10**: 340–355.

[31] Cichocki, A., Zdunek R., Phan A., Amari S. Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation Wiley, 2009

[32] Cover T.M., Hart P.E. Nearest neighbor pattern classification. IEEE Transactions on Information Theory (1967); **13**(1): 21-27.

[33] Curilem G., Vergara J., Fuentealba G., Acuña G., Chacón M., Classification of seismic signals at Villarrica Volcano (Chile) using neural networks and genetic algorithms. Journal of Volcanology and Geothermal Research 2009; **180**(1): 1–8.

[34] Daszykowski M., Walczak B. Methods for the exploratory analysis of two-dimensional chromatographic signals.

[35] Dennis J., Tran H. D., Li H. Spectrogram Image Feature for Sound Event Classification in Mismatched Conditions. IEEE SIGNAL PROCESSING LETTERS 2011; **18**(2): 130–133.

[36] Duda R.O, Hart P.E, Stork D.G. Pattern Classification. Wiley, 2001.

[37] Durante C., Bro R., Cocchi M. A classification tool for N-way array based on SIMCA methodology. Chem. and Intell. Lab. Syst. 2011; **106**: 73–85.

[38] Duin, R.P.W. Classifiers in almost empty spaces. In: 15th International Conference on Pattern Recognition. Volume 2., Barcelona, Spain, IEEE Computer Society (2000).

[39] Dyrby M, Engelsen S, Nørgaard L, Bruhn M, Lundsberg Nielsen L. Chemometric quantitation of the active substance in a pharmaceutical tablet using Near Infrared (nir) Transmittance and nir ft Raman spectra. Applied Spectroscopy 2002; **56**(5): 579–585.

[40] Ebrahimi D., Li J., Hibbert D. B. Classification of weathered petroleum oils by multi-way analysis of gas chromatography-mass spectrometry data using PARAFAC2 parallel factor analysis. J Chromatogr A. 2007; **1166**(1-2): 163–70.

[41] Fearn T, Riccioli C, Garrido-Varo A, Guerrero-Ginel J. E. On the geometry of SNV and MSC. Chemometrics and Intelligent Laboratory Systems 2009; **96**: 22–26 .

[42] Ferraty F, Vieu P. Nonparametric Functional Data Analysis: Theory and Practice (Springer Series in Statistics). Springer-Verlag New York, Inc., 2006.

[43] Friedman J. H. Regularized discriminant analysis. Journal of the American Statistical Association 1989; **84**(405): 165–175.

[44] Fukunaga K. Introduction to Statistical Pattern Recognition (2nd edn.) Computer Science and Scientific Computing. Academic Press Professional, Inc.,1990.

[45] Fukunaga, K., Hayes, R.: Effects of sample size in classifier design. IEEE Transactions on Pattern Analysis and Machine Intelligence 1991; **11**(8): 873–885.

[46] Geladi P, MacDougall D, Martens H. Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat. Appl. Spectrosc 1985; **39**(3): 491–500.

[47] Gemperline, P. J., Miller, K. H., West, T. L., Weinstein, J. E., Hamilton, J. C., Bray, J. T. Principal component analysis, trace elements, and blue crab shell disease. Analytical Chemistry (1992); **64**: 523–531.

[48] Geng X, Smith-Milesa K, Zhou Z. H, Wang L. Face image modeling by multilinear subspace analysis with missing values. In: Proceedings of the 17th ACM International Conference on Multimedia MM'09 2009; 6290–632.

[49] Gonzalez R. C., Woods R. E. Digital Image Processing (3rd edition). Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2006.

[50] Grill C. P., Rush V. N. Analysing spectral data: comparison and application of two techniques. Biological Journal of the Linnean Society 2000; **69**:121–138.

[51] Hall G.J., Clow K.E., Kenny J.E. Estuarial fingerprinting through multidimensional fluorescence and multivariate analysis. Environ. Sci. Technol. 2005; **39**: 7560–7567.

[52] Hall G. J., Kenny J. E. Estuarine water classification using EEM spectroscopy and PARAFAC-SIMCA. Analytica Chimica Acta 2007; **581**: 118–124.

[53] Harshman R.A. Foundations of the Parafac procedure: models and conditions for an explanation multi-modal factor analysis. UCLA Working Papers in Phonetics, Los Angeles 1970; **16**: 1–84.

[54] Hernández N, Biscay R.J, Talavera I. Support vector regression methods for functional data. In: CIARP, Viña del Mar, Chile 2007; LNCS, Springer, **4756**: 564–573.

[55] Hernández N, Talavera I, Biscay R.J, Porro D, Ferreira M. C. Support vector regression for functional data in multivariate calibration problems. Analytica Chimica Acta 2009; **642**(1-2): 110–116

[56] Honghai F, Guoshun C, Cheng Y, Bingru Y, Yumei C. A SVM Regression Based Approach to Filling in Missing Values. LNCS - Knowledge-Based Intelligent Information and Engineering Systems 2005; **3683**: 581–587.

[57] Jain A. K., Murty M.N.,Flynn P.J. Data Clustering: A Review ACM Computing Surveys (CSUR) (1999); **31**(3): 264–323.

[58] Jetto L, Orlando G, Sanfilippo A. The Edge Point Detection Problem in Image Sequences: Definition and Comparative Evaluation of Some 3D Edge Detecting Schemes. In: Proceedings of the 7th Mediterranean Conference on Control and Automation (MED99): Haifa, Israel, 1999: 2161–2171.

[59] Kiers H. K. L. Three-way methods for the analysis of qualitative and quantitative two-way data DSWO Press, University of Leiden, Netherlands 1989.

[60] Kiers H.A.L. Towards a standardized notation and terminology in multiway analysis. J. Chemom. 2000; **14**: 105–122.

[61] Kim S. W., Duin R. P. W. On combining dissimilarity-based classifiers to solve the small sample size problem for appearance-based face recognition In: CAI '07: Proceedings of the 20th conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence, Volume 4509 of LNCS, 2007, 110–121.

[62] Khosravi A., Meléndez J., Colomer J. Classification of sags gathered in distribution substations based on multiway principal component analysis. Electric Power Systems Research 2009; **79**: 144–151.

[63] 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 2007; **20**: 2–13.

[64] Kroonenberg P. M. Applied Multiway Data Analysis. Hoboken, NJ:Wiley, 2008.

[65] Kroonenberg, P. M., Roder, I. Situational dependence of emotions and coping strategies in children with asthma. A three-mode component analysis. New trends in psychometrics. Universal Academic Press (2008): 191–198.

[66] Lai M.-J., Schumaker L.L. Spline Functions On Triangulations, Encyclopedia of Mathematics and its Applications. Cambridge University Press (UK), 2007.

[67] Latchoumane D. , Charles Francois V., Vialatte A. , Cichocki F., Jeong J. Multiway Analysis of Alzheimers Disease:Classification based on Space-frequency Characteristics of EEG Time Series. In: Proceedings of the World Congress on Engineering WCE 2008, Vol. II, 2008.

[68] Lathauwer L, De Moor B. From matrix to tensor: Multilinear algebra and signal processing. In: Proceedings of the 4th International Conference on Mathematics in Signal Processing: Warwick, UK, 1996, **I**: 1–11.

[69] Lesage P. , Glangeau F. , Mars J. Applications of autoregressive models and time-frequency analysis to the study of volcanic tremor and long-period events. Journal of Volcanology and Geothermal Research 2002; **114**: 391–417.

[70] Leurgans S.E, Moyeed R.A, Silverman B. Canonical correlation analysis when the data are curves. Journal of the Royal Statistical Society 1993; **Ser. B**(55): 725–740.

[71] Li T., Sidiropoulos N. D., Giannakis G. B. PARAFAC STAP for the UESA Radar. In: Proceedings of ASAP2000, MIT Lincoln Laboratory, Lexington, Mass, 2000.

[72] Li Y, Ngom Alioune, Rueda L. Missing Value Imputation Methods for Gene-Sample-Time Microarray Data Analysis. In: Proceedings of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, IEEE, Montreal, Canada, May 2010; 183–189.

[73] Lindon J. C., Tranter G. E., Holmes J. L Encyclopedia of Spectroscopy and Spectrometry, Three-Volume Set. Elsevier, 2000.

[74] Little R. J. A, Rubin D. B. Statistical Analysis with Missing Data. Wiley, New York 1987.

[75] Madsen K, Nielsen H. B, Tingleff O. Methods for Non-linear Least Squares Problems. Dept MAthematical Modelling, Technical Universityof Denmark, Lyngby, Denmark, 2004.

[76] Masotti M., Falsaperla S. , Langer H. , Spampinato S. , Campanini R. Automatic classification of volcanic tremor using Support Vector Machine, Conception, verification and application of innovative techniques to study active volcanoes. Istituto Nazionale di Geofisica e Vulcanologia Press (2008).

[77] Mecca V. F. , Ramakrishnan D., Krolik J. L. MIMO Radar Space-Time Adaptive Processing for Multipath Clutter Mitigation. In: Proceedings Fourth IEEE Workshop on Sensor Array and Multichannel Processing, Waltham, MA, 2006, 249–253.

[78] Meng J., Zhang W. Volume measure in 2DPCA-based face recognition. Pattern Recognition Letters 2007; **28**: 1203–1208.

[79] Millán-Giraldo M, Duin R. P. W., Snchez J. S. Dissimilarity-based Classification of Data with Missing Attributes. In: Cognitive Information Processing (CIP), 2nd International Workshop 2010; 293–298.

[80] Mitchell T.M., Hutchinson R., Niculescu R.S., Pereira F., Wang X. Learning to decode cognitive states from brain images Machine Learning (2004); **57**: 145–175.

[81] Molina-Félix L.C, Oliveira-Rezende S., Monard, M.C., Caulkins C.W. Transforming a regression problem into a classification problem using hybrid discretization Computación y Sistemas (2000); **4**(1): 44–52.

[82] Møller J.K.S., Parolari G. , Gabba L., Christensen J., Skibsted L.H. Evaluated surface autofluorescence spectroscopy in order to measure age-related quality index of Parma ham during processing. J. Agr. Food Chem. 2003; **51**: 1224-1230.

[83] Okabe A., Boots B., Sugihara K. Spatial Tessellations: Concept and applications of Voronoi diagrams. John Wiley & Sons, Chichester, 1992.

[84] Oliveira R. L., de Lima B. S. L. P.,Ebeckena N. F. F. The use of multi-way analysis in the classification task of passive sonar contacts. Mecánica Computacional 2010; **29**: 9389–9405.

[85] Orozco-Alzate M, Garca M.E, Duin R.P.W, Castellanos C.G. Dissimilarity-based classification of seismic signals at Nevado del Ruiz Volcano. Earth Sci. Res. J. 2006; **10**(2): 57–65.

[86] Orozco-Alzate M., Skurichina M., Duin R. P. W. Spectral characterization of volcanic earthquakes at Nevado del Ruiz Volcano using spectral band selection/extraction techniques. In: Progress in Pattern Recognition, Image Analysis and Applications.Proceedings of the 13th Iberoamerican Congress on Pattern Recognition CIARP 2008, Volume 5197 of LNCS, Springer, 2008, 708–715.

[87] Paclik P., Duin R.P.W. Classifying spectral data using relational representation. In:Spectral Imaging Workshop, Graz, Austria (2003).

[88] Paclik P, Duin R.P.W. Dissimilarity-based classification of spectra: computational issues. Real Time Imaging 2003; **9**(4): 237–244.

[89] Pekalska E, Duin R.P.W. Classifiers for dissimilarity-based pattern recognition. In: International Conference on Pattern Recognition: Barcelona, Spain, 2000, 12–16.

[90] Pekalska E., Duin R. P. W. On combining dissimilarity representations. In: MCS '01: Proceedings of the Second International Workshop on Multiple Classifier Systems, Volume 2096 of LNCS, London, UK, 2001, 359–368.

[91] Pekalska E, Duin R.P.W. Dissimilarity representations allow for building good classifiers. Pattern Recognition Letters 2002; **23**(8): 943–956.

[92] Pekalska E, Duin R.P.W. Prototype selection for finding efficient representations of dissimilarity data. In: International Conference on Pattern Recognition 2002; textb3: Quebec, Canada, 2002, 37–40.

[93] Duin R.P.W, Pekalska E, Paclik P, Tax D.M.J. The dissimilarity representation, a basis for domain based pattern recognition? Representations in Pattern Recognition, invited talk (refereed) IAPR Workshop, Cambridge 2004: 43–56.

[94] Pekalska E, Duin R.P.W. The Dissimilarity Representation For Pattern Recognition. Foundations and Applications. World Scientific, 2005.

[95] Peng H., Long F., Ding C. Feature selection based on Mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 2005; **27**(8): 1226–1238.

[96] Mortensen Petersen P, Bro R. Real time monitoring and chemical profiling of a cultivation process. Chem. and Intell. Lab. Syst. 2005; **84**(1-2): 106–113.

[97] Porro-Muñoz D, Talavera I, Duin R.P.W, Hernández N. The representation of chemical spectral data for classification. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Proceedings of the 14th Iberoamerican Congress on Pattern Recognition CIARP 2009, vol. 5856, LNCS, Springer: Guadalajara, Mexico, 2009, 513–520.

[98] Porro-Muñoz D., Talavera I., Duin R. P. W. Multi-way data analysis. Tech. Rep. RNPS No.2142, CENATAV (2009).

[99] Porro-Muñoz, D., Duin, R.P.W., Orozco-alzate, M., Talavera, I., Londoño Bonilla, J.M. The dissimilarity representation as a tool for three-way data classification: a 2D measure. In: Structural, Syntactic, and Statistical Pattern Recognition. Proceedings of S+SSPR2010. Volume 6218, LNCS, Springer: Çeşme, Turkey, 2010, 569–578.

[100] Porro-Muñoz D., Duin R.P.W., Talavera I., Orozco-Alzate M. Classification of three-way data by the dissimilarity representation. Signal Processing 2011; **91**(11): 2520–2529.

[101] Porro-Muñoz, D., Talavera, I., Duin, R.P.W., Hernández, N., Orozco-Alzate, M.: Dissimilarity representation on functional spectral data for classification. Journal of Chemometrics **25** (2011) 476–486.

[102] Porro-Muñoz, D., Duin, R.P.W., Talavera, I., Orozco-Alzate, M.: A Study on the Influence of Shape in Classifying Small Spectral Data Sets. In: Similarity Based Pattern Recognition. Proceedings of SIMBAD 2011. Volume **7005** of LNCS., Springer (Octuber 2011): 306–320

[103] Porro-Muñoz, D., Duin, R.P.W., Orozco-Alzate M., Talavera, I.: Continuous Multi-way Shape Measure for Dissimilarity Representation In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Proceedings of the 17th Iberoamerican Congress on Pattern Recognition CIARP 2012. Volume **7441** of LNCS., Springer (September 2012) 430–437.

[104] Porro-Muñoz, D., Duin, R.P.W., Talavera, I., Orozco-Alzate M., Pino-Alea J.A.: Optimizing Dissimilarities for the Classification of Three-way Chemical Spectral Data. Chemometrics and Intelligent Laboratory Systems (under second review after major revision)

[105] Porro-Muñoz, D., Duin, R.P.W., Talavera, I.: Missing values in dissimilarity-based classification of multi-way data. In: Advances in Pattern Recognition and Applications. Proceedings of the 18th Iberoamerican Congress on Pattern Recognition CIARP 2013. Volumes **8258/8259** of LNCS. To be published.

[106] Preda C, Saporta G. Pls regression on stochastic processes. Comput. Statist. & Data Anal. 2005; **48**(1): 149–158.

[107] Preda C. Regression models for functional data by reproducing kernel Hilbert spaces methods. J. Statist. Plan. Infer. 2007; **137**(3): 829–840.

[108] Ramsay J.O, Silverman B.W. Functional Data Analysis (2nd edn.). Springer Series in Statistics: Springer-Verlag, 1997.

[109] Raudys, S.J., Jain, A.K.: Small sample size effects in statistical pattern recognition: Recommendations for practitioners. IEEE Transactions on Pattern Analysis and Machine Intelligence 1991; **3**(3): 252–264.

[110] Renard N., Bourennane, S. Dimensionality Reduction Based on Tensor Modeling for Classification Methods. IEEE Transactions on Geoscience and Remote Sensing 2009; **47**(4): 1123–1131 .

[111] Rifkin R., Bouvrie J., Schutte K. , Chikkerur S. , Kouh M., Ezzat T., Poggio T. Phonetic Classification Using Hierarchical, Feed-forward, Spectro-temporal Patch-based Architectures, Tech. Rep. MIT-CSAIL-TR-2007-019 CBCL-267, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of technology, Cambridge (2007).

[112] Rong Y. , Vorobyov S. A. , Gershman A. B. , Sidiropoulos N. D. Blind Spatial Signature Estimation via Time-Varying User Power Loading and Parallel Factor Analysis. IEEE Trans. on Signal Processing 2005; **53**(5): 1697–1710.

[113] Rossi F, Delannayc N, Conan-Gueza B, Verleysen M. Representation of functional data in neural networks. Neurocomputing 2005; **64**(2): 183–210.

[114] Rossi F, Francois D, Wertz V, Meurens M, Verleysen M. Fast selection of spectral variables with b-spline compression. Chemometrics and Intelligent Laboratory Systems 2007; **86**(2): 208–218.

[115] Sádecká J., Tóthová J. Fluorescence spectroscopy and chemometrics in the food classification- a Review. Czech J. Food Sci. 2007; **25**: 159–173.

[116] Shinzawa H., Morita S., Ozaki Y., Tsenkova R. New method for spectral data classification: Two-way moving window principal component analysis. Applied Spectroscopy 2006; **60**(8): 884–891.

[117] Sidiropoulos N. D., Dimic G. Z. Blind Multiuser Detection in W-CDMA Systems with Large Delay Spread. IEEE Signal Proc. Letters 2001; **8**(9): 87–89.

[118] Smoliński A., Walczak B., Einax J.W. Exploratory analysis of data sets with missing elements and outliers. Chemosphere (2002); **49**: 233-245.

[119] Schafer J. L., Olsden M. K. Multiple imputation for multivariate missing data problems: A data analyst's perspective. Multivariate Behavioral Research 1998; **33**: 545–571.

[120] Schafer J. L., Graham J. W. Missing data: Our view of the state of the art. Psychological Methods 2002; **7**: 147–177.

[121] Scheuren F. Multiple imputation: How it began and continues. The American Statistician 2005; **59**: 315–319.

[122] Shannon C. E. A Mathematical Theory of Communication. The Bell system technical journal 1948; **27**:379–423,623–656.

[123] Silverman B. Smoothed functional principal components analysis by choice of norm. Ann. Statist. 1996; **24**(1): 1–24.

[124] Skov T., Ballabio D., Bro R. Multiblock Variance Partitioning. A new approach for comparing variation in multiple data blocks Analytica Chimica Acta 2008; **615 (1)**: 18–29.

[125] Skov T., Bro R. Solving fundamental problems in chromatographic analysis. Analytical and Bioanalytical Chemistry 2008; **390**(1): 281–285

[126] Smilde A. K., Bro R., Geladi P. Multi-way Analysis. Applications in the chemical sciences. Wiley, England, 2004.

[127] Sun J. T., Zeng H. J., Liu H., Lu Y., Chen Z. 2005. Cubesvd: A novel approach to personalized web search. In: Proceedings of the 14th International World Wide Web conference; 382–390.

[128] Tomasi G., van den Berg F., Andersson C. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. Journal of Chemometrics 2004; **18**: 231–241.

[129] Tomasi G., Bro R. PARAFAC and missing values. Chemometrics and Intelligent Laboratory Systems 2005; **75**: 163–180.

[130] Torgo L., Gama J. Regression using classification algorithms. Intelligent Data Analysis (1997); **1**(4): 275–292.

[131] Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Botstein D., Altman R. B. Missing value estimation methods for DNA microarrays. Bioinformatics 2001; **17**(6): 520–525.

[132] Tucker, L. R. The extension of factor analysis to three-dimensional matrices. In: Contributions to Mathematical Psychology. Holt, Rinehart and Winston, New York (1964); 110–182.

[133] Vapnik V. Statistical Learning Theory. John Wiley & Sons, Inc., 1998.

[134] Varmuza K, Karlovits M, Demuth W. Spectral similarity versus structural similarity: infrared spectroscopy. Anal. Chimica Acta 2003; **490**(1-2): 313–324.

[135] Varmuza K., He P., Kai-Tai F. Boosting applied to classification of mass spectral data. Journal of Data Science 2003; **1**: 391–404.

[136] Vega-Pons S., Ruiz-Shulcloper J. A Survey of Clustering Ensemble Algorithms. International Journal of Pattern Recognition and Artificial Intelligence 2011; **25**(3): 1–36.

[137] Vichi M. One-Mode classification of a three-way data matrix. Journal of Classification 1999; **16**: 27–44.

[138] Rossi F, Villa N. Support vector machine for functional data classification. ESANN'05.

[139] Rossi F, Villa N. Support vector machine for functional data classification. Neurocomputing 2006; **69**(7-9): 730–742.

[140] Wahba G. Spline Models for Observational Data. SIAM [Society for Industrial and Applied Mathematics]:Philadelphia, USA, 1990.

[141] Walczak B., Massart D.L. Tutorial Dealing with missing data: Part I. Chemometrics and Intelligent Laboratory Systems (2001); **58**: 15-27.

[142] Walczak B., Massart D.L. Tutorial Dealing with missing data: Part II. Chemometrics and Intelligent Laboratory Systems (2001); **58**: 29-42.

[143] Watson D. F., "Contouring: A guide to the analysis and display of spacial data", Pergamon Press, 1992.

[144] Ramsay J.O. Fdafuns toolbox. ftp://ego.psych.mcgill.ca/pub/ramsay/FDAfuns/ [2007].

[145] Andersson Claus A., Bro R. The N-way toolbox for MATLAB Chemometrics & Intelligent Laboratory Systems 2000; **52**(1):1–4. http://www.prtools.org/download.html/ [2000].

[146] Oil dataset. http://cac2008.teledetection.fr/shootout [2007].

[147] PARAFAC2 model for MATLAB 5.2. http://www.models.kvl.dk/go?filename=parafac2.m [2011].

[148] Pirouette software. Infometrix. http://www.infometrix.com.

[149] Duin R.P.W, Juszczak P, de Ridder D, Paclik P, Pekalska E, Tax D.M.J. PRTools: a Matlab toolbox for pattern recognition. http://www.prtools.org/download.html/ [2010].

[150] Tablets dataset, http://www.models.kvl.dk/research/data. [2008].

[151] Thodberg H.H. Tecator dataset. Danish Meat Research Institute. http:lib.stat.cmu.edu/datasets/tecator/ [2007].

[152] Wine dataset, http://www.models.kvl.dk/datasets.html [2008].

[153] Wicks M. C., Rangaswamy M., Adve R., Hale T. B. Space-Time Adaptive Processing: A knowledge-based perspective for airborne radar. Signal Processing Magazine, IEEE 2006; **23**(1): 51–65.

[154] Wold S, Sjostrom M. SIMCA: A method for analyzing chemical data in terms of similarity and analogy. Chemometrics Theory and Application 1977; **52**: 243–282.

[155] Wu W., Guob Q., Massart D.L., Boucon C., de Jong S. Structure preserving feature selection in PARAFAC using a genetic algorithm and Procrustes analysis. Chemometrics and Intelligent Laboratory Systems 2003; **65**: 83–95.

[156] Yang J., Yang J. Y. From image vector to matrix: A straightforward image projection technique-IMPCA vs. PCA. Pattern Recognition 2002; **35**: 1997–1999.

[157] Yang J., Zhang D., Frangi A., Yang J. Two-dimensional PCA: A new approach to appearance-based face representation and recognition. IEEE Trans. Pattern Anal. Machine Intell. 2004; **26**(1): 131–137.

[158] Yuhas R.H., Goet 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 1992: 147–149.

[159] Zerzucha P., Daszykowski M., Walczak B. Dissimilarity partial least squares applied to non-linear modeling problems. Chemometrics and Intelligent Laboratory Systems 2012; **110**: 156–162

[160] Zhang A. Advanced analysis of gene expression microarray data. World Scientific Publishing Co., Singapore 2009; 214–238.

[161] Zuo W., Zhang D., Wang K. An assembled matrix distance metric for 2DPCA-based image recognition. Pattern Recognition Letters 2006; **27**: 210–216.

Summary

For many pattern recognition applications, objects are represented by high-dimensional feature vectors, as the result of measurements that are taken from them. Such is the case of spectral data, which is commonly represented by sampling, as a set of individual observations, ignoring the continuous nature of the original data. However, for a considerable number of applications like chemometrics, psychometrics, neuroinformatics, a more appropriate structure to represent data would be a three-way or multi-way array. This type of data should be analyzed with multi-way methods, and this is how researchers from the related fields, moved from multivariate analysis to multi-way analysis.

The work reported in this thesis is concerned with the representation and classification of spectral, or in general, continuous multi-way data, such that their continuous nature (structure) is used in their analysis. Although multi-way analysis is not a new research area, classification algorithms for data with this structure and the idea of taking into account the nature of data in their analyses are rather poorly developed. This thesis shows the importance of taking into account the structure/nature of data, in this case continuous multi-way e.g. spectral data, for their discrimination.

The first part of our research is dedicated to the study of alternative representations for 1D spectral data, such that their functional nature can be taken into account in their analysis. A representation based on using dissimilarities of an object to other objects as features is studied. In the second part of this study, we show that this dissimilarity representation is also very suitable for multi-way data classification. Particularly, it is useful to incorporate knowledge on the original features/application into the dissimilarities between objects.

The dissimilarity representation is also a powerful approach to cope with the small sample size data sets, i.e. few objects in relation to the amount of features, which is very common in spectral data sets. It is further shown that for spectral data, the dissimilarity representation gives flexibility to concentrate on characteristics of spectra, such as the peaks in the signal. By selecting the most discriminative peaks on each feature direction of the array, noisy and redundant information can be discarded. This procedure reduces the cost of the computation of dissimilarities and may improve the accuracy of classifiers.

The last part of the thesis studies how to deal with missing values in the context of dissimilarity-based classifiers, as this problem can occur frequently in multi-way data applications. We analyze how the data can be reconstructed before computing dissimilarities. Alternatively, we propose a modification of the dissimilarity measure that takes into account the shape information (CMS), in which dissimilarities are based on the available data only. Both approaches have shown to work well for relatively small amounts of missing values (up to 10%). For some imputation strategies, however, classifiers performed relatively well even for large amounts of missing data (70%).

This thesis contributes to the classification of multi-way data. It is shown that including information about the problem at hand in the representation of multi-way data improves the performance of classifiers. As such, it is a new basis for further applications and research in other topics related to multi-way data.

Samenvatting

Voor veel toepassingen patroonherkenning worden objecten worden ingedeeld in klassen vertegenwoordigd door verschillende getallen die een vectorconstruct, als gevolg van metingen die worden ontnomen. Dit is het geval van spectrale data, die gewoonlijk wordt vertegenwoordigd door bemonstering, zoals een reeks afzonderlijke waarnemingen negeren het continue karakter van de oorspronkelijke data. Zou echter een groot aantal toepassingen zoals chemometrie, psychometrie, neuro een geschiktere structuur om gegevens vormen een drieweg of multi array. Dit soort gegevens moeten worden geanalyseerd met multi-way methoden, en dit is hoe onderzoekers van de gerelateerde gebieden, verplaatst van multivariate analyse naar multi-way analyse.

Het werk dat in dit proefschrift betreft de vertegenwoordiging en classificatie van spectrale, of in het algemeen continue multi-way data, zodat de continuteit (structuur) wordt gebruikt in de analyse. Hoewel multi-way analyse niet een nieuw onderzoeksgebied, zijn classificatie algoritmen voor data met deze structuur en het idee van het, rekening houdend met de aard van de gegevens in hun analyse in plaats slecht ontwikkeld. Dit proefschrift toont het belang om rekening te houden met de structuur / aard van de gegevens, in dit geval continue multi-way, e.g. spectrale, gegevens voor hun discriminatie.

Het eerste deel van ons onderzoek is gewijd aan de studie van alternatieve representaties voor 1D spectrale gegevens, zodanig dat de functionele aard kunnen worden gehouden in hun analyse. Een afbeelding gebaseerd op het gebruik van een object verschillen van andere objecten kenmerken bestudeerd. In het tweede deel van deze studie tonen we dat dit verschil voorstelling ook zeer geschikt voor multi-weg classificatie. In het bijzonder is het nuttig om kennis over de originele functies in de verschillen tussen de objecten.

Het verschil weergave blijkt ook een krachtige methode om aan de kleine steekproef datasets dus weinig training objecten ten opzichte van de hoeveelheid functies die worden gemeten, wat heel gebruikelijk is in spectrale data sets. Verder wordt aangetoond dat spectrale data, de ongelijkheid representatie flexibiliteit geeft concentreren op eigenschappen van de spectrale gegevens, zoals de pieken in het signaal. Door de onderscheidende pieken op elke functie richting van de array, kan luidruchtig en redundante informatie worden weggegooid. Deze procedure vermindert de kosten van de berekening van verschillen en kan de nauwkeurigheid van classificatoren.

Het laatste deel van het proefschrift bestudeert hoe ontbrekende waarden toerekenen in de context van ongelijkheid gebaseerde classifiers, want dit probleem kan vaak voorkomen in multi-way data-applicaties. We analyseren hoe de gegevens worden gereconstrueerd voor computers verschillen, en subsidiair, stellen we een aanpassing van de ongelijkheid maatregel die rekening houdt met de vorminformatie (CMS), waarvan de ongelijkheid is gebaseerd op de beschikbare gegevens alleen. Beide benaderingen hebben aangetoond goed te werken voor relatief kleine bedragen van ontbrekende waarden (tot 10%). Voor sommige toerekening strategien echter classifiers presteerde relatief goed, zelfs voor grote hoeveelheden van ontbrekende gegevens (70%).

Dit proefschrift draagt bij aan de indeling van multi-way data door te laten zien dat het opnemen van informatie over het probleem bij de hand in de representatie van de multi-way data, verbetert classifier optredens. Als zodanig is het een nieuwe basis voor verdere toepassingen en onderzoek in andere onderwerpen met betrekking tot multi-way data.

Acknowledgments

Although I always thought about continuing working on my professional growth, doing a PhD was kind of a remote idea. Working at CENATAV and many people I met in that stage of my life, made the PhD become more than just an idea.

I would like to thank my supervisors: Bob and Isneri, for sharing with me all the wonderful knowledge they possess, making me always look for answers, and for their guide, encouragement and support. I really respect and admire you very much. I have to say that you are quite different, but still you managed to work with me both at the same time, respecting each other‘s work and opinions.

Isneri, I really appreciate how you got me into the research world slowly, guiding me since my first steps and motivating me with all these real world applications. I thank you for your endless support. You showed me that it was possible, to be perseverant all the time, that things always come out.

Bob, if the idea of doing a PhD was something remote, what to say of the possibility of doing it with you and the people from the PRLab!! I still cannot believe it sometimes. When you first came to Cuba for the Pattern Recognition course, I was practically starting in this world, and you made such an impression on me that made me want to go for more. Thank you very much for having me. I learned a lot from you, in every meeting, every talk we had always giving me strength, pointing out interesting questions, problems, weaknesses. This was not only about research, but of the philosophy of life.

I will also be endlessly grateful to my beloved husband Sandro and family. Sandro (mi Miji), you were always one step ahead and I considered you like a guide. I have to thank you for all your love and dedication, helping me in the difficult moments, studying together many nights, pushing me when I needed it. My mom, dad and little sister (you will always be my little sister), I have no words to describe how much I love you and thank you for being always there for me. I have always wanted you to feel proud of me, and that has been a strong moving force for me. To my closest family, my grandparents (second parents and first teachers), I love you so so much, you are the best in the world!!. To my uncles, aunts, cousins and my other family, Teresa, Segundo, Alain, Elizabeth, Fernand, Daniel. To my dear and unconditional friends, One, Rainer y Malena, thank you all for being there for me. I dedicate this work to you all!!.

To my colleagues at CENATAV: Noslen, Heydi, Dina, Yoanna, Yenisel, Ricardo, I cannot mention all names here but you all know that I really appreciate you. Thank you also to the colleagues from PRLab, which have always received me with open arms: David, Marco, Alessandro, Cuong, Yan, Veronika, I have really enjoyed scientific and not-scientific meetings with you, there was always something new to learn. To Marcel, for his good advices. To Mauricio, thank you for your help, it was very nice to cooperate with you. You are all special.

Curriculum Vitae

Diana Porro Muñoz was born in Camagüey, Cuba, on 15 November 1984. From 2002 she studied Computer Sciences Engineering at the Higher Polytechnic Institute José Antonio Echeverría (CUJAE), La Habana, Cuba, obtaining the Eng. degree (first-level university degree after 5 years, equivalent to Master in many universities) in 2007. During this period she was enrolled in different projects, mainly related to software engineering. From 2006-2007 she also trained as assistant professor of Programming, Databases and Algebra for Computer Sciences Engineering, Biomedical Engineering and Technological School of Informatics, receiving a degree of Professor of Computer Sciences.

From 2005-2007 she did her student practice at the Advanced Technologies Application Center (CENATAV), in Havana, Cuba, where she started working in the area of pattern recognition and chemometrics. Her diploma thesis was a project of this institute and it consisted on the development of an automatic system with methods for the analysis and classification of chemical data. Since she graduated until 2012, she worked at CENATAV as a researcher and software developer. She was also lecturer and teaching instructor of post-graduate chemometrics courses for industry and for undergraduate students of Chemistry at Universidad de la Habana and Chemical Engineering at CUJAE. Diana has been member of tribunals of undergraduate Computer Science Engineering thesis defenses, as well as supervisor of two Computer Sciences Engineering diploma thesis. Her current research interests include Chemometrics, Pattern Recognition, Multi-way data Analysis and Signal Processing.

In 2008 she started her Ph.D. research with the Pattern Recognition group at the Information and Communication Theory Department, Faculty of Electrical Engineering, Mathematics and Computer Science of the same university, at Delft Technical University. The Ph.D research focuses on pattern recognition and chemometrics, more explicitly on the classification of multi-way spectral data by the dissimilarity representation, supervised by Prof. Robert P. W. Duin and Isneri Talavera from CENATAV.

Diana is currently a member of: the Cuban Society of Mathematics and Computer Science (SCMC), the Cuban Association of Pattern Recognition (ACRP) and the International Association of Pattern Recognition (IAPR). She has also been member of the organizing committee of national and international conferences.