Classifying Three-way Seismic Volcanic Data by Dissimilarity Representation

Diana Porro-Muñoz *†, Isneri Talavera*, Robert P.W. Duin†, Mauricio Orozco-Alzate‡ and John Makario Londoño-Bonilla§
*Advanced Technologies Application Center, CENATAV, Havana, Cuba
Email: dporro, italavera@cenatav.co.cu
†Pattern Recognition Lab, TU Delft, The Netherlands
Email: r.duin@ieee.org
‡Departamento de Informática y Computación,
Universidad Nacional de Colombia Sede Manizales, Colombia
Email: morozcoa@bt.unal.edu.co
§Observatorio Vulcanológico y Sismológico de Manizales,
Instituto Colombiano de Geología y Minería (INGEOMINAS), Colombia
Email:jmakario@ingeominas.gov.co

***Abstract*—Multi-way data analysis is a multivariate data analysis technique having a wide application in some fields. Nevertheless, the development of classification tools for this type of representation is incipient yet. In this paper we study the dissimilarity representation for the classification of three-way data, as dissimilarities allow the representation of multi-dimensional objects in a natural way. As an example, the classification of seismic volcanic events is used. It is shown that in this application classification based on 2D spectrograms, dissimilarities perform better than on 1D spectral features.**

***Keywords*-volcanic seismic data, three-way representation, dissimilarity representation, classification**

I. Introduction

Multi-way data analysis [1], [2], is the extension of multivariate analysis when the analyzed data is arranged in higher order arrays; several sets of variables measured on different samples can be used. The most common is the three-dimensional array, but it is even possible to generate higher dimensional data i.e. multi-way array. The analysis of such data is often used for extracting hidden structures and exploring the interrelations in the data. It has been shown that this information may not be analyzed accurately by a two-way analysis, because it does not respect the multi-way design of the data. Nowadays, most of the applications of multi-way analysis are for exploratory and regression purposes. Classification has been studied much less. This might be caused by the lack of appropriate classification tools.

In recent studies [3], [4], the advantage of learning from dissimilarities between the objects instead of traditional features has been shown, in what is known as Dissimilarity Representation (DR) [5]. This representation was mainly designed for classification. It is based on the important role that is played by the pairwise dissimilarities between objects. 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 the user defined dissimilarity measure, in which application background information may be expressed. Any traditional classifier that operates in feature spaces can also be used in the dissimilarity space.

The automatic classification of seismic volcanic signals is an essential task nowadays, with the goal of discovering the interaction between volcanic earthquakes and volcanic processes. Traditionally, signals are naturally represented in the time domain. Although this representation has been used for automatic analysis, they are usually represented by a spectrum in terms of energy spread over its frequency components, from their Fourier transform (See Fig. I) [6], [7]. Recent studies have also shown that training the classifiers on the space generated by the dissimilarities between the objects, is a feasible and more reliable alternative for automatic classification of seismic signals than the frequency-based one [3]. Nevertheless, time or frequency representations alone may not be optimal for seismic signal analysis, since spectral energy changes in time. This relation is not considered in any of the previously mentioned representations [6]. Due to this limitation, the use of a time-frequency representation like spectrograms, showing frequency changes in time, may be advantageous. So far, the spectrograms have just been averaged to obtain the spectral representation [8]. The 2D object representations has not intensively been exploited as such in automatic classification systems. Examples are the use of Hidden Markov Models for continuous seismic-event classification [9] or dynamic time warping.

In this paper we study the DR based on the time-frequency information in three-way volcanic seismic data. It is generated from the spectrograms of the signals measured by the Olleta crater station of the Nevado del Ruiz Volcano in Colombia. Two classes of events are analyzed: Volcano-Tectonic (VT) earthquakes and Long-Period (LP) earthquakes. A 2D dissimilarity measure is proposed. Results

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.205

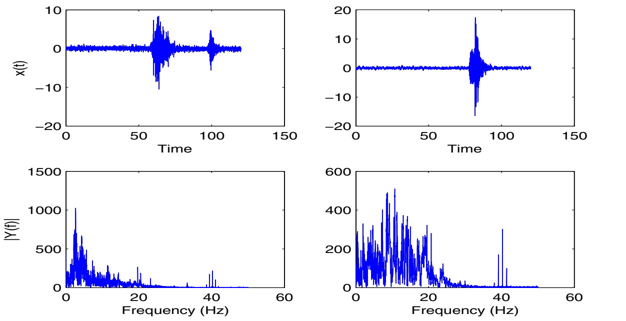


Figure 1. Time signals and frequency spectra for two classes, VT (left), LP (right)

are compared with 1D feature representation using the time integrated spectra.

II. Three-way volcanic data

The seismic signals to be analyzed belong to the ice-capped Nevado del Ruiz volcano in the Colombian Andes. This volcano is currently studied by the Volcanological and Seismological Observatory at Manizales. Signals from the Olleta crater station (reference station) were selected for the experiments. Signals were digitized at 100.16 Hz sampling frequency by using a 12 bits analogue-to-digital converter. Automatic detection/segmentation stages are based on the short-term average to long-term average ratio (STA-LTA) algorithm, with a captured sample offset of 2048. It is a classical algorithm used as standard in seismic detection [10]. The a-priori classification of the signals is done by visual inspection. The dataset is composed of 12032 points signals of two types of volcanic activities: 235 of LP events, and 235 of VT earthquakes.

The differences in spectral content of these signals allow the discrimination between the events. That is why spectral-based classification is often used for this type of data. However, due to the frequency content changes in time, this should also be taken into account in the analysis.

An intuitive way to represent this time-frequency relationship for all the signals would be what is known as a three-way array $\mathbf{Y}(\mathbf{l},\mathbf{m},\mathbf{n}) \in \mathbb{R}^{I_1 \times I_2 \times I_3}$. In the different research areas, unlike three-way array configurations can be found. The most common design is defined as “profile data” [1], and it has the form $(objects \times variable1 \times variable2)$. This is the kind of design we propose to use for the seismic volcanic data. In this seismic volcanic three-way array configuration $(signals \times time \times frequency)$, the signals are organized in the vertical (first dimension) axis. The second dimension corresponds to the time (horizontal) axis.The third dimension corresponds then to the frequency (depth) axis. To obtain the time-frequency representation of each signal we used spectrograms.

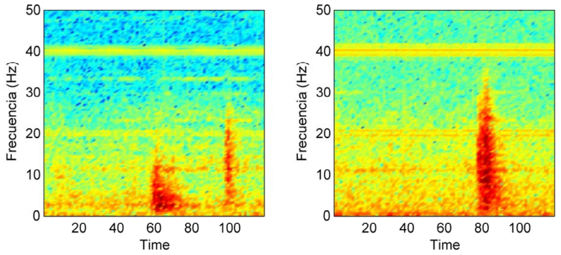


Figure 2. Spectrograms of two events: VT class (left), LP class (right).

While a spectrum (1D) represents the signal in terms of energy spread over its frequency components (in a given interval in time), a spectrogram also displays the changes that occur over time. This technique allows to extract a matrix of frequency spectra corresponding to a sequence of windowed Fourier transforms of the original data trace. Spectral amplitudes are then displayed as a function of time in a 2D way.

III. Dissimilarity Representation for three-way data

Using the Dissimilarity Representation (DR) [5] classifiers are trained in the space of the proximities between objects, instead of the traditional feature. Thus, instead of the feature matrix $\mathbf{X}(\mathbf{l},\mathbf{m})$, where $\mathbf{l}$ runs over the objects (signals) and $\mathbf{m}$ over the measured variables e.g. frequencies, the set of objects is represented by the matrix $\mathbf{D}(\mathbf{X},\mathbf{R})$. This matrix contains the dissimilarity values $d(x_l, r_p)$ between objects $x \in \mathbf{X}$ and the objects of the representation set $\mathbf{R}(r_1, ..., r_p)$. 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. Classifiers are built in this space using a training set.

The elements of $\mathbf{R}$ are called prototypes, and have preferably to be selected by a prototype selection method [5]. 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 completely different sets. As dissimilarities are computed to the representation set $\mathbf{R}$, a dimensionality reduction is reached if a good, small set can be found, resulting in less computationally expensive classifiers. Any traditional classifier that operates in feature space can be used in the dissimilarity space.

For the three-way data $\mathbf{Y}(\mathbf{l},\mathbf{m},\mathbf{n})$ we are dealing with, the theory of the DR is the same. The issue we have to address here is how to obtain the dissimilarities from this three-way representation, in which each object is defined by a 2D matrix of measurements. As a first approach, we propose to take each object (matrix) $y \in \mathbf{Y}$ of the three-way data, and compute the dissimilarities between them by a 2D dissimilarity measures.

A. 2D dissimilarity measure

An important issue (and opportunity) of the DR is the selection of a suitable dissimilarity measure for the problem at hand. In the time-frequency three-way representation, we need to analyze shape changes in the spectral (frequency) direction and connectivity in the time direction. Hence, based on the results obtained with the shape measure for simple spectra [4], we propose to make use of the derivatives into the AMD measure, introduced previously for the 2DPCA algorithm [11]. In such a way, we can take the ordering into account as well as the shape of the spectra, resulting in a 2DShape dissimilarity measure:

1) $D_1(y_1, y_2) = \left(\sum_{j=1}^{m} \left(\sum_{i=1}^{l} (y^{\sigma}_{1ij} - y^{\sigma}_{2ij})^2 \right)^{p/2} \right)^{1/p}$ on the spectral direction with $y^{\sigma}_{i} = \frac{\mathrm{d}}{\mathrm{d}_i} G(i, \sigma) * y_i$

2) $D_2(y_1, y_2) = \left(\sum_{i=1}^{l} \left(\sum_{j=1}^{m} (y^{\sigma}_{1ij} - y^{\sigma}_{2ij})^2 \right)^{p/2} \right)^{1/p}$ on the time direction with $y^{\sigma}_{j} = \frac{\mathrm{d}}{\mathrm{d}_j} G(j, \sigma) * y_j$

3) Combine both dissimilarities $D = \frac{1}{\omega_1} D_1 + \frac{1}{\omega_2} D2$

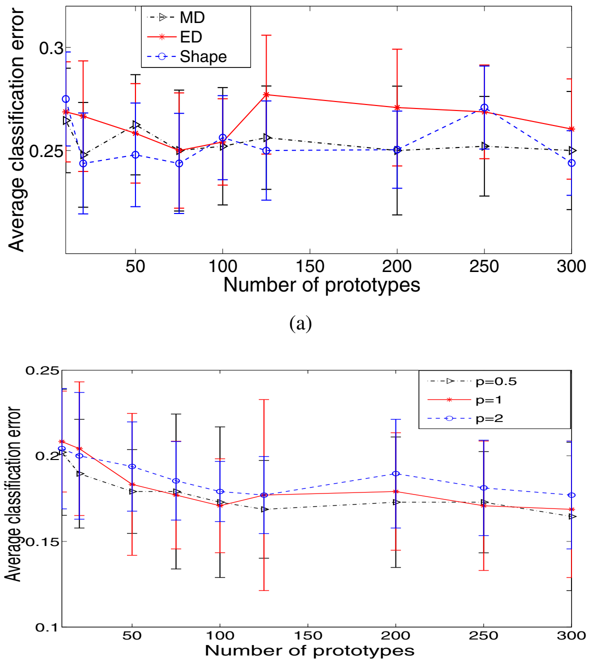
The weight p is used to emphasize either small or large differences between the elements, in dependence of the problem at hand. If $p \leq 1$, all the differences will be reduced, thus the larger ones will not interfere much in the measure. On the other hand, if $p \geq 1$, the larger differences will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

In the combination step, we included a weight for scaling. In this study we defined $\omega_k = var(D_k)$, to scale each dissimilarity matrix by its variance.

IV. Experimental Results

To show the advantages of time-frequency (spectral) based classification over the spectral-based classification, we make a comparison of the classification results on the dissimilarity space derived from both representations. For the experiments, a dataset with 235 objects per class (VT and LP) is considered. For the 1D (spectral) representation we have computed the spectrum by using a 12032-point Fast Fourier Transform (FFT). Thus, the hole signal is analyzed in both 1D and 2D representations. For the 2D (spectrogram) representation, trying to make a trade-off between time and frequency resolution, a 256 short time Fourier transform was calculated using time-windows of 256 points with 50% of overlap. The parameterization values lead to $470 \times 129 \times 93$ three-way data. Before computing both representations, the raw signals were normalized to zero-mean and unit-variance.

A Fisher Linear classifier was computed on the dissimilarity space. Experiments were repeated 10 times. Training and test objects were randomly chosen from the total data set, in a 10-fold cross-validation process. Different sizes of the representation set $[10, 20, 50, 75, 100, 125, 200, 250, 300]$ were randomly selected. For the generation of the dissimilarity space, we computed the Manhattan (MD), Euclidean (ED) and Shape measures on the spectral representation. These measures have shown to perform well for spectral data [3], [4]. In a 5-times 10-fold cross-validation from a range of values $[1 - 50]$, the best results were achieved with $\sigma = 15$. The proposed measure in Sec. III-A was used for the three-way data, with values of $p = [0.5, 1, 2]$. For the spectral direction we selected $\sigma = 3$ and for the time direction $\sigma = 2$. The dissimilarity matrices were computed on the whole spectrograms. Each seismic event has a different start time, and so the spectrograms. In the following figures, the Average Classification Errors (ACE) for the DR on both spectral and three-way data are shown, using different sizes of the representation set.



(a)

(b)

Figure 3. ACE on the 1D (left) and 2D (right) representations for different numbers of prototypes

V. Discussion and Conclusions

We studied the use of the Dissimilarity Representation for classifying three-way volcanic data. This way, the relationship between the different dimensions is analyzed i.e. change of frequency content in time. Besides, information about the data that is missing in the actual representation e.g. shape and connectivity, can be taken into account in the dissimilarity measure. A 2D dissimilarity measure was proposed for this purpose. It takes into account the spectral shape and continuity in time direction. The relations between the objects are analyzed in the dissimilarity space.

It can be observed in Fig. III-A that, the ACE on the dissimilarity space generated from the spectral data is around 25% and 30%. The error values for the Manhattan measure are slightly better than those of the Euclidean and Shape measures. Nevertheless, if the standard deviation is taken into account, the values for the three measures are very similar. The results with the Shape measure (derivative-based) are not as expected (based in previous works). Hence, these results could suggest that there is not more information to be captured from this representation. It is also possible that these measures are not robust enough for this problem, which somehow contradicts the previous studies [3], [4]. Further studies may be done to find a more proper measure for this type of data. However, when we analyze the error of the DR from the three-way data $(15 - 20\%)$ we see a significant improvement. This ratifies the fact that the time-frequency relation is more discriminative than the spectra. Besides, the proposed 2D measure is capable of capturing this information.

If we analyze the ACE on the DR from the spectral data, we can see that for 20 or more prototypes it is approximately stable. The explanation we give to this phenomenon, is that there is not more discriminating information to be found in more prototypes. On the other hand, if we analyze the ACE on the DR from the three-way data, we can see that the behavior is different. While increasing the number of prototypes, the ACE decreases. The more prototypes we add, the more information we have to discriminate between the classes. Nevertheless, due to what is called the peaking phenomenon, when the number of prototypes starts reaching the size of the training set, the errors will increase.

The good performance of classifiers by the proposed approach, compared with the traditional one, shows that our proposal can be a good solution for this kind of problems. The analysis of the prototype selection helps to corroborate it. The inclusion of more information in the three-way data, increases the chances of a better discrimination in this kind of problems. However, a study of the influence of different overlaps and more precise techniques to obtain the time-frequency representation, could improve the three-way analysis results.

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.

References

[1] P. M. Kroonenberg, *Applied Multiway Data Analysis*, ser. John Wiley & Sons, 2008.

[2] D. Porro-Muñoz, I. Talavera, and R. P. W. Duin, "Multi-way data analysis," http://www.cenatav.co.cu/doc/RTecnicos/RT 20SerieAzul014web.pdf, CENATAV, Tech. Rep., 2009.

[3] M. Orozco-Alzate, M. E. García, R. P. W. Duin, and C. G. Castellanos, "Dissimilarity-based classification of seismic signals at Nevado del Ruiz volcano," *Earth Sci. Res. J.*, vol. 10, no. 2, pp. 57–65, 2006.

[4] P. Paclik and R. P. W. Duin, "Dissimilarity-based classification of spectra: computational issues," *Real Time Imaging*, vol. 9, no. 4, pp. 237–244, 2003.

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

[6] M. Benbrahim, A. Daoudi, K. Benjelloun, and A. Ibenbrahim, "Discrimination of seismic signals using artificial neural networks," *World Academy of Science, Engineering and Technology*, vol. 4, pp. 4–7, 2005.

[7] H. Langer, S. Falsaperla, T. Powell, and G. Thompson, "Automatic classification and a-posteriori analysis of seismic event identification at Soufrière Hills volcano, Montserrat," *Journal of volcanology and geothermal research*, vol. 153, pp. 1–10, 2006.

[8] M. Masotti, S. Falsaperla, H. Langer, S. Spampinato, and R. Campanini, *Conception, verification and application of innovative techniques to study active volcanoes.* Istituto Nazionale di Geofisica e Vulcanologia Press, Italy, 2008.

[9] M. C. Benítez, J. Ramírez, J. C. Segura, J. M. Ibáñez, J. Almendros, A. García-Yeguas, and G. Cortés, "Continuous HMM-based seismic-event classification at Deception Island, Antarctica," *IEEE Transactions on Geoscience and remote sensing*, vol. 45, no. 1, pp. 138–146, 2007.

[10] C. A. Vargas-Jimenez and S. Rincon-Botero, "Portable digital seismological ac station over mobile telephone network and internet," *Computers & Geosciences*, vol. 29, pp. 685–694, 2003.

[11] W. Zuo, D. Zhang, and K. Wang, "An assembled matrix distance metric for 2DPCA-based image recognition," *Pattern Recognition Letters*, vol. 27, pp. 210–216, 2006.