Models for Identifying Structures in the Data: A Performance Comparison

Anna Esposito[1], Antonietta M. Esposito[3], Flora Giudicepietro[3], Maria Marinaro[2], and Silvia Scarpetta[2]

[1] Dipartimento di Psicologia, Seconda Università di Napoli, and IIASS, Italy
iiass.annaesp@tin.it
[2] Dipartimento di Fisica, Università di Salerno, INFN, and INFM Salerno, Italy
[3] Istituto Nazionale di Geofisica e Vulcanologia, (Osservatorio Vesuviano), Italy

Abstract. This paper reports on the unsupervised analysis of seismic signals recorded in Italy, respectively on the Vesuvius volcano, located in Naples, and on the Stromboli volcano, located North of Eastern Sicily. The Vesuvius dataset is composed of earthquakes and false events like thunders, man-made quarry and undersea explosions. The Stromboli dataset consists of explosion-quakes, landslides and volcanic microtremor signals. The aim of this paper is to apply on these datasets three projection methods, the linear Principal Component Analysis (PCA), the Self-Organizing Map (SOM), and the Curvilinear Component Analysis (CCA), in order to compare their performance. Since these algorithms are well known to be able to exploit structures and organize data providing a clear framework for understanding and interpreting their relationships, this work examines the category of structural information that they can provide on our specific sets. Moreover, the paper suggests a breakthrough in the application area of the SOM, used here for clustering different seismic signals. The results show that, among the three above techniques, SOM better visualizes the complex set of high-dimensional data discovering their intrinsic structure and eventually appropriately clustering the different signal typologies under examination, discriminating the explosion-quakes from the landslides and microtremor recorded at the Stromboli volcano, and the earthquakes from natural (thunders) and artificial (quarry blasts and undersea explosions) events recorded at the Vesuvius volcano.

Keywords: models for data structure, seismic events, clustering, classification.

1 Introduction

Dimension reduction techniques, used for analyzing and visualizing complex sets of data, can be distinguished into two classes: the linear ones, like Principal Component Analysis (PCA) [4] or the classical Multidimensional Scaling (MDS) [10], and the nonlinear methods, like the Self-Organizing Map (SOM) [6] or nonlinear variants of MDS, as the recently proposed Curvilinear Component Analysis (CCA) [1].

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part III, LNAI 4694, pp. 275–283, 2007.

PCA is able to perform eigenvalue decomposition on the data, detecting linear dependencies between vectors of features which constitute the dataset of interest. However, this linear detection may be a limitation when it is necessary to capture higher order structure in the data. To this intent, the Self-Organizing Map (SOM) is one of the most powerful projection methods since it can transform input data of arbitrary dimension into a low dimensional topology preserving map. However, the obtained fixed topological structure can represent a prior constraint. When no matching takes place between the discovered and the intrinsic structure of the input data, this technique leads to sub-optimal mappings. In this case, in order to obtain a more consistent representation of the input data, it is suggested to use the CCA algorithm that has proved to be successful for several applications [1]. However, this work will show that, for the proposed discrimination task, SOM performs better than CCA, and that CCA does not outperform PCA.

This paper deals with the unsupervised analysis and discrimination of seismic signals associated to the activity of two high risk volcanoes, the Vesuvius and the Stromboli. The Vesuvius is located in Naples, a very populated area in which volcano-tectonic earthquakes and transient signals due to external sources (man-made explosions, thunders, etc) are daily observed by the experts and classified through procedures based on the visual analysis of the spectral and temporal features of the detected signals. The automation of these procedures is strongly desirable in order to identify a more robust description of earthquakes with respect to the signals generated by external sources and to avoid human inconsistencies which can affect the quality of the classification.

Likewise, the Stromboli volcano, one of the Aeolian Islands in the Tyrrhenian Sea, has a permanent eruptive activity, called Strombolian activity, continuously monitored by a broadband network of digital stations. In this case, the seismicity is characterized by explosion-quakes and microtremor. In Dec. 2002 there was a big landslide that generated a small tsunami, creating the necessity to automatically discriminate among these different typologies of events.

An automatic high-performance strategy for discriminating among different seismic signals could not only drastically reduce the probability of false event detections but also decrease the workload of the community involved in the seismological monitoring of the areas. In previous works [2,9] we have already faced this problem using a supervised learning algorithm that was able to implement a very good discrimination on both datasets described above. However, a supervised analysis requires a correctly labeled dataset and this is not always obtainable, above all when there are several and continuous changing events. Thus, the approach here proposed to automatically classify these signals and still overcome the heavy labeling, is based on unsupervised techniques that should be able to visualize the intrinsic data structure and cluster together similar events.

In the following, the Vesuvius and Stromboli datasets are described first. Then, the analysis methods used to preprocess the seismic data are introduced. Section 4 presents the mathematical basics of the three models under study and the obtained results are discussed in Section 5. Section 6 is dedicated to conclusions and remarks.

2 The Vesuvius and the Stromboli Datasets

The **Vesuvius** dataset includes 961 events, recorded by four stations (CPV, NL9, TRZ, and BKE). For the CPV station, situated on the coast of the Gulf of Naples, we have 144 earthquakes and 247 man-made undersea explosions. For the NL9 station, placed in Nola, there are 109 earthquakes and 114 man-made quarry explosions. For the TRZ station, located at the basis of the Vesuvius, 104 events are earthquakes and 103 man-made quarry explosions. Finally, for the BKE station**,** located up on the Vesuvius crater, there are 72 earthquakes and 68 thunders. Each 22s-long signal is described by a vector of 2200 components due to the 100 Hz sampling rate. The labeling made by the experts identified a total of 429 earthquakes, 247 undersea explosions, 114 quarry blasts at the NL9 station, 103 quarry blasts at the TRZ station, and 68 thunders, representing the five classes to discriminate.

The **Stromboli** monitoring network is composed of 13 digital stations, which acquire the data using a sampling rate of 50 Hz and transmit them to the Monitoring Center in Naples (more details are on line at www.ov.ingv.it/stromboli/). The examined dataset contains 1159 records, coming from 5 seismic stations (STR1, STRA, STR8, STR5, STRB), and in particular consists of three classes of signals, with 430 explosion-quakes, 267 landslides and 462 microtremor signals. Each 24s-long record is described by a vector of 1200 components due to the 50 Hz sampling rate.

3 Data Preprocessing

In order to be able to discriminate among seismic, natural and artificial events, it would be suitable to have a signal representation containing both frequency and temporal information. Such a representation is justified by the fact that the experts exploit both these attributes for a visual classification of the seismic signals and is further confirmed by previous works [2,9] based on supervised techniques, in which optimal discrimination performance has been reached.

In this paper, the signal spectral content is obtained using the Linear Predictive Coding (LPC) algorithm [8], while a discrete waveform parameterization gives the amplitude-versus-time information. For both datasets each recording is processed on a short-time basis, dividing it into a certain number of analysis windows, whose length is fixed taking into account all the frequencies of interest in the signal.

The LPC algorithm works modeling each signal sample s_n as a linear combination of its p past values, i.e. formally:

$$\overline{s_n} = \sum_{k=1}^{p} c_k s_{n-k} + G \tag{1}$$

where c_k are the *prediction coefficients*, which efficiently encode the frequency features, G is the *gain* and p indicates the *model order*. The c_k estimation is realized by an optimization procedure which tries to minimize the error between the real value of the signal sample at time t and its LPC estimate.

The correct value for p is problem dependent. However, it must be a good trade off between the compactness and the significance of the data representation. In this paper, the p value for the Stromboli data was settled to p = 6 since this value proved to be effective in a previous work [2]. Likewise, for the Vesuvius dataset, it was settled exploiting the good results obtained in [9], where it was fixed to $p = 6$ in order to allow the two-class discrimination at each station. In our case the task is more complex, having five classes of signals, thus it has been increased to $p = 10$.

The time domain information, added to the spectral content of each event, is computed as the properly normalized difference between the maximum and the minimum signal amplitude in a 1s-long analysis window. As a final step, the resulting feature vectors for both datasets were *logarithmically* normalized since this operation improves the clustering for both CCA and SOM strategy.

4 PCA, SOM and CCA Description

The three clustering techniques here presented make different assumptions about the representational structure used to define clusters and on the similarity measures which describe the relationships between objects and clusters.

PCA finds the axes of maximum variance of the input data and represents them by a linear projection onto the subspace spanned by the principal axes [4].

CCA [1] instead performs a nonlinear dimensionality reduction and representation in two steps: (1) a vector quantization (VQ) of the input data into k quantized n-dimensional prototypes and (2) a nonlinear projection of these quantized vectors onto a p-dimensional output space. The nonlinear mapping is obtained by minimizing the cost function:

$$E = 1/2 \sum_i \sum_j (X_{ij} - Y_{ij})^2 F(Y_{ij}, \lambda) \tag{2}$$

where $X_{ij}=d(x_i,x_j)$ and $Y_{ij}=d(y_i,y_j)$ are the Euclidean distances between the quantized and the output vectors, respectively, and $F(Y_{ij},\lambda)=exp(-Y_{ij}/\lambda)$ is a weighting function that favors the preservation of the data topology depending on the λ value.

The Kohonen Self-Organizing Map (SOM) performs a non-linear mapping of an n-dimensional input space onto a two-dimensional regular grid of processing units known as *neurons*. A prototype vector is associated to each node. The fitting of the prototype of each node is carried out by a sequential regression process that minimizes the differences between each input vector and the corresponding winning node's prototype (see [6] for mathematical details). However, contrarily to the CCA, the SOM clustering is not critically dependent on its parameters. In our tests, the SOM parameters have been settled in agreement with the prescriptions reported in [5]. The SOM algorithm realizes two important actions: a clustering of the input data into nodes and a local spatial ordering of the map, i.e. the prototypes are ordered on the grid so that similar inputs fall in topographically close nodes. This ordering facilitates the understanding of data structures. Moreover, displaying on the map the Euclidean distances between prototype vectors of neighboring nodes through grey levels, the SOM gives a good representation of the cluster structure, graphically depicting the data density too.

5 Results

The three models above described were applied on the two datasets under examination using a bi-dimensional output representation. Figure 1 displays the PCA clustering for the Vesuvius and the Stromboli. The legends are made exploiting the labeling performed by the experts. Thus, for the Vesuvius data (Fig. 1A), the stars indicate volcanic earthquakes recorded by all the four stations; the empty circles and diamonds are quarry explosions at the NL9 and TRZ stations respectively; the empty down-triangles are thunders and the empty squares undersea explosions. For the Stromboli dataset (Fig. 1B), the empty squares indicate landslides while the empty circles and the up-triangles are explosion-quakes and microtremor respectively. Observing Figure 1 we note that the PCA projection mixes the different signals all together and does not discriminate among them. This because PCA is not able to capture the peculiar characteristics of our data, probably not related to the maximum variance directions.

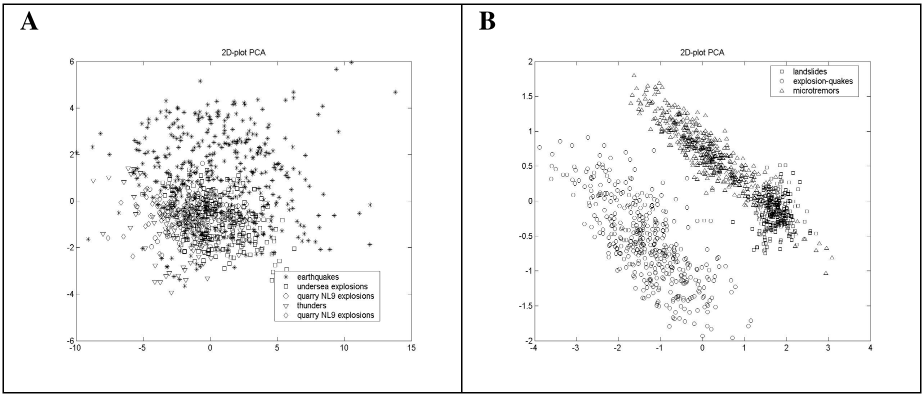


Fig. 1. PCA projection: for the Vesuvius set (A) the stars indicate earthquakes, the empty circles and diamonds quarry blasts at the NL9 and TRZ stations respectively, the empty down-triangles are thunders and the empty squares undersea explosions. For the Stromboli set (B) the empty squares indicate landslides while the empty circles and the empty up-triangles are explosion-quakes and microtremor respectively.

Figure 2 shows the bi-dimensional CCA representation for the Vesuvius data. In particular, figure 2A shows the *dydx* plot obtained with appropriate values for the η and λ parameters, and figure 2B displays the resulting CCA projection.

Figure 3 displays the CCA results for the Stromboli volcano. Figure 3A visualizes the *dydx* plot with suitable values for the η and λ parameters, and Figure 3B shows the bi-dimensional CCA projection on these data. As we can observe, the CCA does not allow to discriminate among the classes of signals under examination probably because the principal curvilinear components are not discriminative of our typologies of signals, thus the resulting clustering shows several overlaps among them (see in particular Figures 2B and 3B). Finally, the SOM results on the Vesuvius (Figure 4)

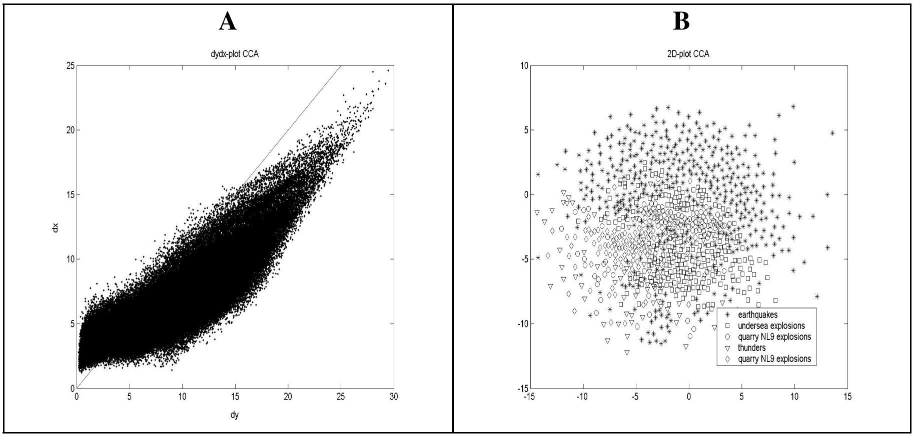


Fig. 2. The CCA results on the Vesuvius dataset. The *dydx* plot (A), obtained using appropriate values for η and λ parameters, and the two-dimensional CCA projection (B) are shown.

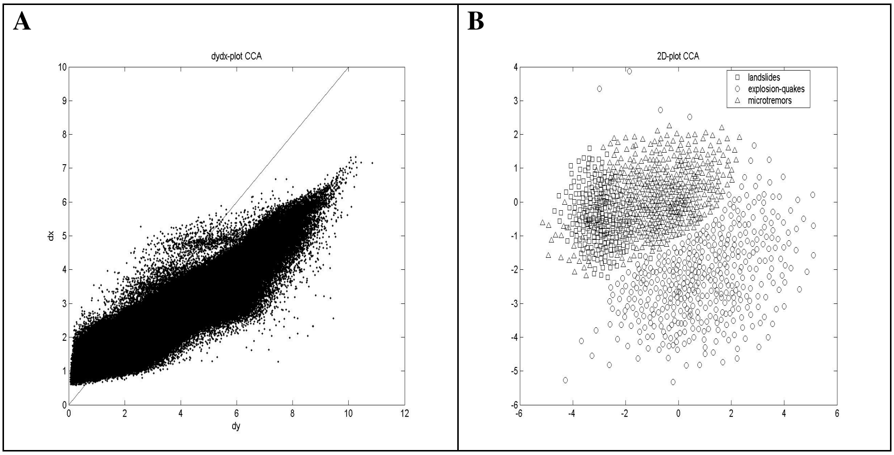
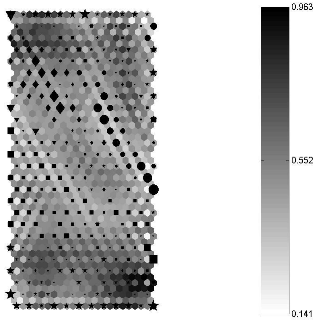


Fig. 3. The CCA results for the Stromboli volcano. The *dydx* plot (A), obtained using specific values for η and λ parameters, and the bi-dimensional CCA visualization (B) are displayed.

and Stromboli (Figure 5) datasets are presented. Each node in both maps is a prototype vector whose size represents the number of feature vectors associated to that prototype.

The distances among the prototypes are visualized on the map using a grey level scale, so that large distances between two prototypes correspond to dark grey color levels on the grid, indicating that the two prototypes and the associated feature vectors are very different. The classes of events are shown on the map using different symbols. Thus, in Figure 4 the stars indicate earthquakes, the circles and diamonds specify quarry blasts at the NL9 and TRZ stations respectively, the down-triangles represent thunders and the squares are undersea explosions. Overlapped symbols indicate that different types of signals belong to the same node. In Figure 4, it is possible to see that each class of signals is clustered on a particular zone of the map and the overlaps between classes are less than those obtained with the PCA or the CCA algorithms. In Figure 5, the squares indicate landslides, the circles are explosion-quakes and the up-triangles represent microtremor signals. From the Figure, it is possible to distinguish a dark gray boundary between explosion-quakes and the other two classes, which instead appear closer to each other. This means that explosion-quakes are well separated from landslides and microtremor, while the less marked distances between these last two types of events suggest that they share similar features.



✱ earthquakes
■ undersea explosions
● quarry NL9 explosions
▼ thunders
◆ quarry TRZ explosions

Fig. 4. The SOM map (with 26x12=312 nodes) for the Vesuvius dataset. The stars indicate earthquakes, the circles and diamonds represent the NL9 and TRZ quarry blasts respectively, the down-triangles specify thunders and the squares are undersea explosions.

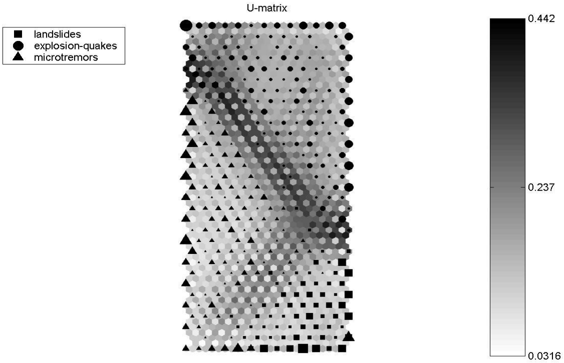


Fig. 5. The SOM map (with 31x13=403 nodes) for the Stromboli dataset. The up-triangles indicate microtremor signals, the circles are explosion-quakes and the squares specify landslides.

Thus, the results reported in Figures 4 and 5 show that the clusters visualized by the SOM better correspond to the classes of signals identified by the experts.

6 Conclusions and Remarks

In the previous section, three unsupervised projection techniques have been applied to two different datasets composed respectively by five and three classes of seismic events and encoded through features vectors containing both spectral and time domain information. Our aim was to try to identify among them the one that better represents on a bi-dimensional plane the data structure, such that the resulting clustering can be helpful for the automatic labeling of the events under study. The unsupervised techniques considered were the PCA, the CCA and the SOM.

These techniques work without assumption about the data distribution and no external information, like class labels, is provided to obtain the final output. The analysis is unsupervised, and the possible class labels have been used only afterwards to aid in the results' interpretation, without affecting the structures discovered by the methods.

It has been shown that, among the above techniques, the SOM algorithm, exploiting information on the local topology of the vector prototypes, gives the best performance being able to group the 5 classes of events for the Vesuvius dataset, and the 3 classes of signals for the Stromboli volcano in separated clusters with minor overlaps than those obtained either with the PCA and/or the CCA algorithm. The poor performance of the PCA algorithm can be due to the difficulty of this linear algorithm to capture the peculiar characteristics of our dataset which may not be related to the maximum variance directions. Moreover, the poorer performance of the CCA algorithm seems to be attributed to its critical dependence on the choice of the parameter λ and on its decreasing time-speed. This could be overcame introducing the CCA with geodetic (curvilinear) distance, also called Curvilinear Distance Analysis (CDA) [7] that has proved, in many cases, to perform better than the CCA and to be not critically dependent from the choice of the λ value. A further work could be to check the above hypothesis.

References

1. Demartines, P., Herault, J.: Curvilinear Component Analysis: A Self-Organizing Neural Network for Nonlinear Mapping of Data Sets. IEEE Transactions on Neural Networks, 8(1), 148–154 (1997)
2. Esposito, A.M., Giudicepietro, F., Scarpetta, S., D'Auria, L., Marinaro, M., Martini, M.: Automatic Discrimination among Landslide, Explosion-Quake and Microtremor Seismic Signals at Stromboli Volcano using Neural Networks. Bulletin of Seismological Society of America (BSSA), 96(4A)
3. Esposito, A.M., Scarpetta, S., Giudicepietro, F., Masiello, S., Pugliese, L., Esposito, A.: Nonlinear Exploratory Data Analysis Applied to Seismic Signals. In: Apolloni, B., Marinaro, M., Nicosia, G., Tagliaferri, R. (eds.) WIRN 2005 and NAIS 2005. LNCS, vol. 3931, pp. 70–77. Springer, Heidelberg (2006)
4. Jollife, I.T.: Principal Component Analysis. Springer, New York (1986)

5. Kohonen, T., Hynninen, J., Kangas, J., Laaksonen, J.: SOM_PAK: The Self-Organizing Map Program Package, Report A31. Helsinki University, Finland (1996) Also available at http://www.cis.hut.fi/research/som_lvq_pak.shtml
6. Kohonen, T.: Self-Organizing Maps, Series in Information Sciences, 2nd edn. vol. 30. Springer, Heidelberg (1997)
7. Lee, J.A., Lendasse, A., Verleysen, M.: Nonlinear Projection with Curvilinear Distances: Isomap versus Curvilinear Distance Analysis. Neurocomputing, 57, 49–76 (2004)
8. Makhoul, J.: Linear Prediction: a Tutorial Review. In: Makhoul, J. (ed.) Proceeding of IEEE, pp. 561–580. IEEE, Los Alamitos (1975)
9. Scarpetta, S., Giudicepietro, F., Ezin, E.C., Petrosino, S., Del Pezzo, E., Martini, M., Marinaro, M.: Automatic Classification of Seismic Signals at Mt. Vesuvius Volcano, Italy, Using Neural Networks, Bulletin of Seismological Society of America (BSSA), Vol. 95, pp. 185–196 (2005)
10. Wish, M., Carroll, J.D.: Multidimensional Scaling and its Applications. In: Krishnaiah, P.R., Kanal, L.N. (eds.) Handbook of Statistics, vol. 2, pp. 317–345. North-Holland, Amsterdam (1982)