EARTH SCIENCES
RESEARCH JOURNAL

Earth Sci. Res. J. Vol. 11, No. 2 (December 2007): 131-138

CLUSTERING ON DISSIMILARITY REPRESENTATIONS FOR DETECTING MISLABELLED SEISMIC SIGNALS AT NEVADO DEL RUIZ VOLCANO

Mauricio Orozco-Alzate, and César Germán Castellanos-Domínguez

Universidad Nacional de Colombia Sede Manizales, Grupo de Control y Procesamiento Digital de Señales, Campus La Nubia, km 7 vía al Magdalena, Manizales, Colombia.
Corresponding author: Mauricio Orozco-Alzate, email: morozcoa@unal.edu.co

ABSTRACT

Classification of seismic signals at Colombian volcanoes has been carried out manually by visual inspection. In order to reduce the workload for the seismic analysts and to turn classification reliable and objective, the use of supervised learning algorithms has been explored; particularly classifiers built in dissimilarity spaces. Nonetheless, the performance of such learning methods is subject to the availability of a representative and a priori well classified training sets. To detect mislabeled events, the use of clustering techniques on the dissimilarity representations is proposed. Our experiments, performed on re-analyzed seismic signals, show a significant improvement respect to recognition accuracies for the original data sets.

Key words: Clustering, dissimilarity, mislabeling, seismic signals.

RESUMEN

La clasificación de las señales sísmicas en los volcanes de Colombia ha sido llevada a cabo manualmente mediante inspección visual. Con el fin de reducir la carga de trabajo de los analistas y para tornar la clasificación confiable y objetiva, se ha explorado el uso de algoritmos de aprendizaje supervisado; particularmente, clasificadores construidos en espacios de disimilitud. No obstante, el desempeño de dichos métodos de aprendizaje está sujeto a la disponibilidad de un conjunto de entrenamiento representativo y, a priori, bien clasificado. Para detectar eventos mal clasificados, se propone el uso de técnicas de agrupamiento sobre las representaciones de disimilitud. Los experimentos, realizados sobre las señales sísmicas verificadas, muestran una mejora significativa respecto a las tasas de reconocimiento para los datos originales.

Palabras claves: Agrupamiento, disimilitud, etiquetado incorrecto, señales sísmicas.

Manuscript received September 9 2007.
Accepted for publication November 30 2007.

INTRODUCTION

In many applications of pattern recognition, it is extremely difficult or expensive, or even impossible, to reliably label a training sample with its true category (Jain *et al*., 2000). Particularly, in automatic classification of seismic-volcanic signals, night and rotating shift work schedules, tedious evaluations, and changes of personnel turn the task of recognition by visual inspection susceptible to human errors. Besides, analysts often engage in differences of opinion about interpretations of dubitable signals.

In order to reduce the workload for the seismic analyst and the risks associated to subjective judgments, a number of supervised classification methods have been used (Scarpetta *et al*., 2005; Langer *et al*., 2006; Orozco-Alzate *et al*., 2006a). It is supposed for those supervised classification techniques that a well-labeled data set is available. However, due to the same reasons cited above, it is highly likely that training sets include mislabeled events.

In Langer *et al*. (2006), an automatic classification of seismic events at Soufrière Hill volcano was carried out. In addition, a careful manual revision of the original a-priori classification was achieved by an expert not involved in the previous labeling of the data set. It was found that a considerable number of the events were erroneously attributed to other classes. As a result, a remarkable improvement in classification accuracy was obtained when the revised data set was used.

The *Nevado del Ruiz* Volcano is monitored by the Volcanological and Seismological Observatory at Manizales (VSOM). Because of the considerable amount of data, the labelling task of the recorded seismic signals is distributed among several analysts (e.g. one trainee per volcanic station). A second or third opinion is requested just in case of serious doubt. As a result, classifications performed by different experts are not available and an analysis of concordance for such a-priori labels was not conducted. In this study, a revision of the original labelled Nevado del Ruiz volcano (*Ruiz*) data set is conducted. In contrast to the approach followed by Langer *et al*. (2006), the revision by using clustering techniques was automated.

Several clustering algorithms on a given data set were used due to the lack of a single appropriate clustering algorithm (Jain *et al*., 2000). Therefore, experiments were conducted by using the most popular clustering approaches, which belong to two basic strategies: hierarchical and partitioning methods. In addition, the *Ruiz* data set was arranged to consider two separated problems: the *Ruiz-VT,LP* (two classes) and the *Ruiz-all* (three classes) data sets. Revised data sets were used according to our previous dissimilarity-based classification approach (Orozco *et al*., 2006a, Orozco *et al*., 2006b) and compared against the performances obtained with the original data sets.

DISSIMILARITY REPRESENTATION AND CLASSIFIER

Differences in spectral content allowed a visual discrimination of different types of volcanic earthquakes. Therefore, spectra of seismic records are commonly used for classification and monitoring of seismic activity (Zobin, 2003). In addition, recent studies have claimed that the dissimilarity-based classification approach is a feasible and sometimes advantageous alternative to the feature-based method (Duin *et al*., 1998, Pękalska *et al*., 2001, Pękalska and Duin, 2002, Paclík and Duin, 2003b, Pękalska and Duin, 2005). According to those facts, a dissimilarity representation for the *Ruiz* data set can be derived as follows: (i) the power spectral density (PSD) for each

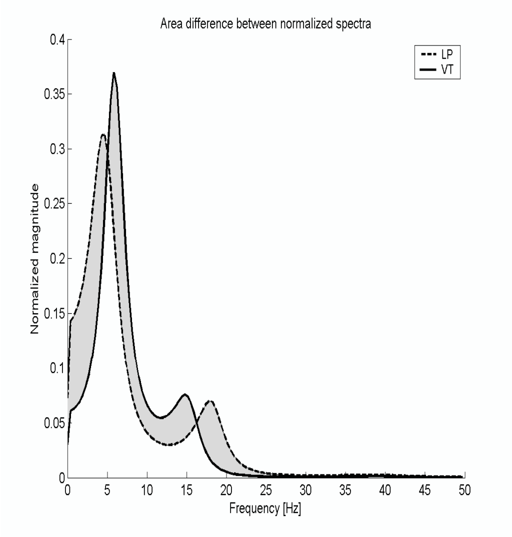

Figure 1. Dissimilarity measure as the difference between normalized spectra.

record is estimated via the Yule-Walker autoregressive method: DC bias must be removed before computing the spectra, (ii) a dissimilarity measure between normalized spectra is calculated as the area difference of the non-overlapping parts (L_1-norm) between spectra, see Fig. **1**.

Figure 1. Dissimilarity measure as the difference between normalized spectra.

A dissimilarity matrix *D(T,T)* was constructed by having those pairwise measures. Each entry d_{ij} of *D* corresponds to the dissimilarity between a pair of seismic records from the training set *T*. Then, a proper classifier can be defined on such a dissimilarity representation, either by using the entire training set *T* or a representation set $R \subseteq T$.

Linear Normal Density Based Classifier

A number of studies have showed that normal density based classifiers perform well in dissimilarity spaces (Pękalska *et al*., 2001, Pękalska and Duin, 2002, Paclík and Duin, 2003b, Paclík and Duin, 2003a, Pękalska *et al*., 2004, Orozco *et al*., 2006a). Particularly, in our previous study with the *Nevado del Ruiz* volcano data set (Orozco *et al*., 2006b), the linear normal density based classifier (BayesNL) outperformed the nearest neighbor rule 1-NN and the quadratic normal density based classifier (BayesNQ). For a two-class problem, the BayesNL classifier is given by

$$f(D(x,R)) = \left[D(x,R) - \frac{1}{2}\left(\mathbf{m}_{(1)} + \mathbf{m}_{(2)}\right)\right]^T \times C^{-1}\left(\mathbf{m}_{(1)} - \mathbf{m}_{(2)}\right) + \log\frac{P_{(1)}}{P_{(2)}} \qquad (1)$$

where *C* is the sample covariance matrix; $\mathbf{m}_{(1)}$, $\mathbf{m}_{(2)}$ are the mean vectors and $P_{(1)}$, $P_{(2)}$ are the class prior probabilities. If *C* is singular, a regularized version must be used. The following regularization is typically used with *λ* equals 0.01 or less (Pękalska *et al*., 2006):

$$C_{reg}^{\lambda} = (1-\lambda)C + \lambda\, diag(C) \quad . \qquad (2)$$

CLUSTERING TECHNIQUES

Unsupervised classification refers to

situations where the objective is to construct decision boundaries based on unlabeled training data (Jain *et al*., 2000). Hierarchical and partitioning methods are the two basic strategies to find clusters. In this study, the following clustering techniques are used: single linkage (SL), average linkage (AL), complete linkage (CL), *k*-means and *k*-centres. SL, AL, and CL are hierarchical, whereas the latter are partitioning methods. A brief description of these approaches is given below.

Hierarchical clustering

The most popular hierarchical techniques for clustering are the agglomerative methods. At the beginning, each object is considered as a single cluster; then, the closest two clusters are merged iteratively until a specified number of clusters is reached (Pękalska and Duin, 2005). Let C_k and C_l be two clusters of the cardinalities n_k and n_l respectively, and let ρ_{kl} be a dissimilarity measure between them. Three basic criteria for the agglomerative methods are summarized in Table **1**.

Partitioning clustering

Partitioning methods group the objects into *k* clusters, usually by using representatives or by assuming a specific geometrical structure. Objects are assigned to the clusters, new representatives are estimated and the process is repeated until a stable solution is reached. Two typical partitioning methods are *k*-means and *k*-centres; see Table **2** for a brief description, a detailed one can be found in Pękalska and Duin (2005).

TABLE 2. Clustering methods.

EXPERIMENTAL RESULTS

Volcano-Tectonic (VT) earthquakes, Long-Period (LP) earthquakes and Icequakes (IC) are the seismic signals classes considered in this study. They are contained in the *Ruiz-all* data set. The *Ruiz-VT,LP* data set includes only the first two classes. Signals were digitized at 100.16 Hz sampling frequency by using a 12 bits analog to digital converter.

Table 1. Hierarchical clustering methods.

Method	$_{kl}$	Emphasis/comment
SL	$\min_{p_i \in C_k} \min_{p_j \in C_l} d(p_i, p_j)$	Connectedness. Resulting clusters are elongated and chain-like.
CL	$\max_{p_i \in C_k} \max_{p_j \in C_l} d(p_i, p_j)$	Compactness. It performs well when the objects form naturally distinct clouds.
AL	$\frac{1}{N_k N_l} \sum_{p_i \in C_k} \sum_{p_j \in C_l} d(p_i, p_j)$	Connectedness and Compactness. It performs well for naturally distinct clouds and elongated clusters.

Recording stations are located near to the Olleta crater and the glacier at Nevado del Ruiz volcanic complex.

In order to explore the level of agreement/disagreement between the labels given by the experts and the ones produced by the clustering algorithms, the number of mismatches for the entire data sets is considered. The averaged number of mismatches over 10 runs is reported in Table **3**. Hierarchical methods report the same number of mismatches over the runs, therefore their standard deviations are zero. SL hierarchical criterion for both the *Ruiz-VT,LP* and the *Ruiz-all* problems presents a rate of disagreement considerable high; similarly, mismatches of AL results for the *Ruiz-all* problem reach 45%. In fact, even though the number of cluster is fixed, SL and AL find second and third clusters of a few objects only. As a result, valid data subsets, i.e. randomly generated and including enough objects per class, are not always warranted. In consequence, AL for the *Ruiz-all* problem and SL in both cases are not considered in the subsequent classification experiments.

Orozco *et al*. (2006b) observed an asymptotic behaviour for training set sizes greater than 60 examples per class. In addition, the BayesNL provided the best overall performance, outperforming the 1-NN and the BayesNQ. According to that, the experiments were conducted with the BayesNL using training sets of a fixed size of 60 objects per class.

Clustering was performed on the entire data sets. Then, training and test sets are randomly extracted for each run. The results are shown in Table **4**. For comparison, the results using the original data are also presented. It is clear that performances for re-labelled data sets are much better than those for the original data.

DISCUSSION AND CONCLUSION

A revision of the original labelled seismic events recorded by the VSOM staff provides a significant improvement in the performance of supervised dissimilarity-based classifiers such as the observed for the BayesNL classifier. The use of events labelled by clustering confirmed that labelling errors are frequent and recurrent.

Clustering uses a notion of proximity, judged in a numerical way. In contrast, labels assigned by experts obey to the

Table 2. Clustering methods.

Method	Description
k-means	Representatives are estimated by cluster mean vectors. The dissimilarity is the Euclidean distance of an object to the cluster means.
k-centres	Centre objects are chosen such that the maximum of the distances over all objects to the nearest centre is minimized. Results depend on random initialization.

Table 3. Averaged number of mismatches between the class labels assigned by the VSOM staff and labels assigned by the clustering method over the entire data sets.

Clustering method	*Ruiz-VT,LP*	*Ruiz-all*
SL	482	1108
CL	367	495
AL	164	861
k-centres	158.2 (25.2049)	507.5 (52.2797)
k-means	135.6 (0.5164)	506.6 (0.5164)
Total	1063	1891

Table 4. Classification error (in % and averaged over 25 runs) with its standard deviation (in %) for the RNLC applied to the revised data sets.

Clustering method	*Ruiz-VT,LP*	*Ruiz-all*
SL	—	—
CL	2.3494 (0.4045)	—
AL	4.9646 (1.1176)	3.77 (0.76)
k-means	2.7524 (0.7405)	5.6722 (0.6109)
k-centres	2.8810 (0.7441)	4.6792 (0.8413)
Total	13.0075 (1.0354)	20.02 (0.81)

visual resemblance between the event and a canonical waveform which analysts have learnt by reference or experience. Obviously, such a method is highly subjective and supposes that differences are easily detected by visual inspection but in many cases this is not true.

Since the final rule used (stand-alone) was calculated from all the data, clustering methods were used on the entire data sets instead of applying them to the training sets only. AL and CL offer the smallest errors for the *Ruiz-VT,LP* and *Ruiz-all* problems respectively. Even tough, the best clustering is hierarchical in both cases; differences are not enough to claim that hierarchical methods should be preferred over the partitioning ones. Nonetheless, a general conclusion can be drawn from our study: the use of a clustering method to confirm labels assigned by experts is highly beneficial for constructing reliable and accurate supervised classifiers of seismic events.

ACKNOWLEDGEMENTS

We thank the VSOM staff for providing the raw data set.

REFERENCES

Duin, R. P. W., de Ridder, D., and Tax, D. M. J. (1998). Featureless pattern classification. *Kybernetika*, **34**, no. 4, 399–404.

Jain, A. K., Duin, R. P. W., and Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Trans. Pattern Anal. Machine Intell.*, **22**, no 1, 4–37.

Langer, H., Falsaperla, S., Powell, T., and Thompson, G. (2006). Automatic classification and a-posteriori analysis of seismic event identification at Soufrière hills volcano, Montserrat. *Journal of Volcanology and Geothermal Research*, **153**, 1–10.

Orozco, M., García, M. E., Duin, R. P. W., and Castellanos, C. G. (2006a). Dissimilarity-based classification of seismic volcanic signals at Nevado del Ruiz volcano. *2nd Latin-American Congress of Seismology*, Bogotá, Colombia, August, CD-ROM.

Orozco, M., García, M. E., Duin, R. P. W., and Castellanos, C. G. (2006b). Dissimilarity-based classification of seismic volcanic signals at Nevado del Ruiz volcano. *Earth Sciences Research Journal*, **10**, no. 2, 57–65.

Paclík, P. and Duin, R. P. W. (2003a). Classifying spectral data using relational representation. *Proceedings of the Spectral Imaging Workshop*, Graz, Austria, April, 31-34.

Paclík, P. and Duin, R. P. W. (2003b). Dissimilarity-based classification of spectra: computational issues. *Real Time Imaging*, **9**, no. 4, 237–244.

Pękalska, E. and Duin, R. P. W. (2002). Dissimilarity representations allow for building good classifiers. *Pattern Recognition Lett.*, **23**, no. 8, 943–956.

Pękalska, E. and Duin, R. P. W. (2005). *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications*. World Scientific, Singapore, 636pp.

Pękalska, E., Duin, R. P. W., Günter, S., and Bunke, H. (2004). On not making dissimilarities Euclidean. *Proceedings of Structural and Statistical Pattern Recognition*, Lisbon, Portugal, August, 1143–1151.

Pękalska, E., Duin, R. P. W., and Paclík, P. (2006). Prototype selection for dissimilarity-based classifiers. *Pattern Recognition*, **39**, no. 2, 189–208.

Pękalska, E., Paclík, P., and Duin, R. P. W. (2001). A generalized kernel approach to dissimilarity based classification. *J. Mach. Learn. Res.*, **2**, no. 2, 175–211.

Scarpetta, S., Giudicepietro, F., Ezin, E. C., Petrosino, S., Pezzo, E. D., Martini, M., and Marinaro, M. (2005). Automatic classification of seismic signals at Mt. Vesuvius volcano, Italy, using neural networks. *Bulletin of the Seismological Society of America*, **95**, no. 1, 185–196.

Zobin, V. (2003). *Introduction to Volcanic Seismology*. Elsevier, Amsterdam, The Netherlands, 302pp.