Nonlin. Processes Geophys., 19, 411–419, 2012
www.nonlin-processes-geophys.net/19/411/2012/
doi:10.5194/npg-19-411-2012

Nonlinear Processes in Geophysics

Application of *k*-means and Gaussian mixture model for classification of seismic activities in Istanbul

H. S. Kuyuk[1,2], **E. Yildirim**[3], **E. Dogan**[1], **and G. Horasan**[3]

[1]Department of Civil Engineering, Sakarya University, Turkey
[2]Seismological Laboratory, University of California, Berkeley, USA
[3]Department of Geophysical Engineering, Sakarya University, Turkey

Correspondence to: H. S. Kuyuk (erdarkuyuk@gmail.com)

Received: 23 December 2011 – Revised: 7 July 2012 – Accepted: 9 July 2012 – Published: 3 August 2012

Abstract. Two unsupervised pattern recognition algorithms, *k*-means, and Gaussian mixture model (GMM) analyses have been applied to classify seismic events in the vicinity of Istanbul. Earthquakes, which are occurring at different seismicity rates and extensions of the Thrace-Eskisehir Fault Zone and the North Anatolian Fault (NAF), Turkey, are being contaminated by quarries operated around Istanbul. We have used two time variant parameters, complexity, the ratio of integrated powers of the velocity seismogram, and S/P amplitude ratio as classifiers by using waveforms of 179 events ($1.8 < M < 3.0$). We have compared two algorithms with classical multivariate linear/quadratic discriminant analyses. The total accuracies of the models for GMM, *k*-means, linear discriminant function (LDF), and quadratic discriminant function (QDF) are 96.1 %, 95.0 %, 96.1 %, 96.6 %, respectively. The performances of models are discussed for earthquakes and quarry blasts separately. All methods clustered the seismic events acceptably where QDF slightly gave better improvements compared to others. We have found that unsupervised clustering algorithms, for which no a-prior target information is available, display a similar discriminatory power as supervised methods of discriminant analysis.

1 Introduction

Identification and classification of different seismic events with similar characteristics in a region of interest is one of the most important subjects in seismic hazard studies. Different techniques using various predictants to discriminate earthquakes and man-made explosions such as quarry blasts, nuclear tests, underwater explosions etc. have been derived worldwide, each one appropriate to a particular region. One well-known statistical method, linear discriminant analysis (LDA), is frequently applied to seismic studies to find a linear combination of features that characterizes or separates two or more classes of events (Koch and Fah, 2002; Rodgers and Walter, 2002; Horasan et al., 2009). Alternatively, soft computing techniques such as artificial neural networks, self organizing map, adaptive neuro-fuzzy inference system etc. have been employed recently (Yıldırım et al., 2011; Kuyuk et al., 2010; Campus and Fah, 1997; Falsaperla et al., 1996; Musil and Plesinger, 1996; Muller et al., 1999; Dowla et al., 1990; Tiira, 1999; Jenkins and Sereno, 2001; Ursino et al., 2001; Del Pezzo et al., 2003; Scarpetta et al., 2005).

Cluster algorithms (we use this term for distinguishing earthquakes from non-earthquake events, but not aftershocks) can be characterized into three groups, with respect to the type of feedback to which the learner has access: (1) supervised, in which learning is the machine learning task of inferring a function from supervised training data; (2) semi (or reinforcement)-supervised, in which learning is a class of machine learning techniques that make use of both labeled and unlabeled data for training – typically a small amount of labeled data with a large amount of unlabeled data; and (3) unsupervised, in which learning refers to the problem of trying to find hidden structure in unlabeled data. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution. Classic LDA statistical based approaches fall into the supervised group which means a-priori information is provided to train an algorithm where *k*-means and Gaussian mixture models are considered as unsupervised since no target is given (factor analysis, principal and independent component analysis

etc.). k-means algorithm is used for clustering source zonation in the Aegean region (Weatherill and Burton, 2009) for pattern recognition of the three-dimensional structure of the active part of a fault network using the spatial location of earthquakes (Ouillon et al., 2008) and for separating stresses from earthquake focal mechanism data (Otsubo et al., 2008). Gaussian mixture model is used for seismic facies classification (Han et al., 2011). Han et al. (2011) proposed an application of the expectation–maximization algorithm to automatically identify geological facies from seismic data. Yet, two algorithms analyzed in this study are not deeply investigated in clustering of seismicity or removing contamination of seismicity catalogs by man-made noise in a region.

Therefore, western part of the North Anatolian Fault (NAF), Turkey, which is one of the most active seismic regions all over the world, is considered to investigate. There is high seismic activity in this area especially in Marmara region, where the metropolitan mega-city Istanbul is located in the north (Fig. 1). Kandilli Observatory and Earthquake Research Institute (KOERI) operates, records and processes the seismic activities for seismic hazard assessment constantly. However, these tasks need to be auto-operated, and systematic since there is high seismic activity, and the operation should be immune to personnel changes. Such a systematic approach does not require an expert's continuous attention, and reduces time-consuming tasks such as late/night work.

As a predictant in clustering studies, amplitude peak ratios, power ratios, and spectral amplitude ratios etc., which are derived from time- and frequency-domain analysis of seismograms, were utilized in the literature (Bennett and Murphy, 1986; Wüster, 1993; Gitterman et al., 1998; Wiemer and Baer, 2000). Horasan et al. (2009) first studied in the Marmara region/Turkey using three parameters: amplitude peak ratio, power ratio, and spectral amplitude ratio with linear discriminants analysis. They advised adding origin time of events as a parameter, because quarry blasts are happening in daytime. Then, Yıldırım et al. (2010) demonstrated the use of feedforward neural networks (FFNNs), adaptive neural fuzzy inference systems (ANFIS), and probabilistic neural networks (PNNs) to discriminate between earthquakes and quarry blasts for the region. The input vectors consist of the peak amplitude ratio (S/P ratio) and the complexity value. The success of the developed models on regional test data varies between 97.67 % and 100 %. The same year Kuyuk et al. (2010) extended discriminants by using four parameters (complexity, spectral ratio, S/P wave amplitude peak ratio and origin time of events) and applied an unsupervised learning approach self-organizing map (SOM); however, they have showed that complexity and S/P parameters are more useful than others. Compared to spectral ratios, time domain parameters are more convenient and reliable for the seismicity-related clustering (Pomeroy et al., 1982; Bennett et al., 1989; Baumgard and Young, 1990).

However, it is a fact that clustering algorithms that might be efficient in one region could be incompetent in some other regions due to local site and source effects, geological structure of path, etc (Zeiler and Velasco, 2009).

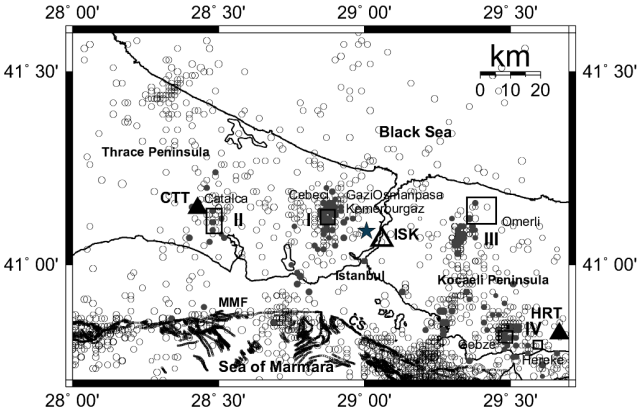

Fig. 1. Map showing the study area (40.70–41.60° N latitude and 28.00–29.70° E longitude) and locations of seismic events used for statistical analysis of the seismicity catalogs (KOERI-NEMC) for 1995–2007 (first 8 months) marked by open black circles and locations of seismic events with duration magnitude between 1.8 and 3.0 (filled black circles) used for waveform of digital data for vertical seismograms recorded at ISK broad-band station (open triangle) and CTT and HRT short-period (filled triangles) stations (KOERI-NEMC, 2001–2004). Boxes show the quarry sites (I: Gaziosmanpaşa/Cebeci and Kemerburgaz, II: Çatalca, III: Ömerli, IV: Hereke) determined from satellite images (Musaoğlu et al., 2004) and field observations. Filled black star shows the location of 20 November 2005 explosion. MMF: Main Marmara Fault; ÇS: Çınarcık segment of the MMF (from Horasan et al., 2009).

The aim of this study is to examine and discuss the performance of three statistical pattern recognition methods, namely Gaussian mixture model, k-means and two algorithms of discriminant functions including QDF, in order to distinguish microearthquakes from quarry blasts in the vicinity of Istanbul. Complexity and S/P ratio are first derived from the seismograms, and they are used as a criterion for the investigation. Comparison of the results has revealed that all methods satisfactorily cluster the seismic events where the QDF slightly gave better improvement compared to others.

2 Data and study area

In this study, the parameters S/P amplitude ratio and complexity are used for classification of seismic activities using statistical analysis. These parameters are obtained by Horasan et al. (2006), which is supported by Bogazici University Research Fund Project. Within the scope of this project, 179 events (Fig. 1) including earthquakes and quarry blasts recorded by KOERI-operated HRT, ISK, and CTT stations by means of computational vertical-velocity seismograms were chosen. These events occurred between 2001 and 2004, and

the duration magnitudes are between 1.8 and 3.0 (KOERI, NEMC).

The study area is located 40.70–41.60° N latitude and 28.00–29.70° E longitude. The rectangles shown in Fig. 1 (I: Gaziosmanpaşa; II: Çatalca, III: Ömerli, IV: Gebze-Hereke) represent quarrying areas as determined by satellite and field observations (Musaoğlu et al., 2004; Horasan et al., 2006).

Horasan et al. (2009) indicated that the seismic events with a magnitude less than three are related to both earthquakes and man-made explosions from the quarries in the study area. No explosions with magnitude 3 or more were observed in the study area.

The number of seismic events in the quarries increased during the daytime interval of 07:00 to 16:00 GMT (09:00 and 18:00 local time), which corresponds to regular blasting hours of the quarries (Horasan et al., 2009). This information is not sufficient to discriminate earthquakes and quarry blasts in the high seismic activity region. Therefore, the waveforms of these events must be investigated. The quarry blast waveform is dominated by the P-wave (the first arrival), whereas the earthquake has a much larger S-wave and surface waves. Therefore, seismogram features may play an important role in the discrimination methods.

Station ISK has a broad-band seismometer, while CTT and HRT have short-period seismometers. The short-period seismometers for CTT and HRT stations were changed with broadband ones after 2007. The sampling frequency of the recorded data was 20 samples per second before October 2001 and 50 samples per second after that.

The first parameter, S/P amplitude ratio, was obtained from the P and S wave peak to peak amplitude measurements on the seismograms using GURALP visualization program Scream 4.3. by Horasan et al. (2009).

The second parameter, complexity, is the ratio of integrated powers of the velocity seismogram $s^2(t)$ in the selected time windows length (t_1 and t_2: first and second time window lengths; t_0: the onset time of P-wave). The complexity (C) can be expressed as follows:

$$C : \int_{t_1}^{t_2} s^2(t)\mathrm{d}t \Big/ \int_{t_0}^{t_1} s^2(t)\mathrm{d}t. \quad (1)$$

The limits of the integrals (t_0, t_1, t_2) of C given in Eq. (1) are $t_0 = 0$ s, $t_1 = 2$ s , and $t_2 = 4$ s. The complexity has higher value for earthquakes, because the S-wave amplitude on the earthquake waveform is greater than the P-wave amplitude (Fig. 2).

Even if the S/P and complexity values might lead us to make wrong classifications, the statistical analysis will eventually show these misclassifications.

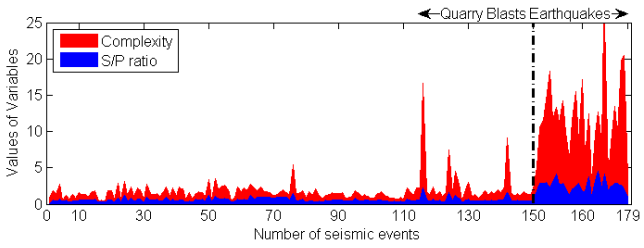

Fig. 2. Distribution of seismic events according to complexity and S/P ratio. There are totally 179 events in the dataset where 150 of them are QBs and 29 of them are EQs. Red color indicates the complexity, which has higher values for earthquakes, and blue color shows the S/P ratios.

3 Methods

Classification techniques can be categorized as supervised, semi-supervised and unsupervised according to their usage of a-priori information (Duda et al., 2001). Mixture modeling and k-means fall into the unsupervised group, which means methods are not trained by feeding them the target identification. On the other hand, classical linear and quadratic discrimination analyses are considered supervised learning.

3.1 Gaussian mixture model

Mixture modeling constructs a model based on a mixture of statistical distribution without requiring that an observed dataset should identify the sub-population to which an individual observation belongs. It supposes that the distribution of the analyzed data is generated from a mixture of simpler statistical distribution, representing the number of clusters within the data. Mixture modeling could be used to obtain cluster classification of point in 2-D. Gaussian mixture model is a probability distribution that is a convex combination of other Gaussian probability distribution (GPF) and is defined as follows:

$$f(x) = 1 \Big/ \sqrt{2\pi\sigma^2} e^{-(r-\bar{x})^2/2\sigma^2} \quad (2)$$

where parameters $\bar{x}$ and σ^2 are mean and the variance. Suppose that the random variable r is a mixture of M Gaussian models. Then, the probability density function (pdf) of r is given as the Gaussian mixture model:

$$P(r|\beta, \Theta) = \sum_{i=1}^{M} \alpha_i P(r|\bar{x}_i, C_i) \quad (3)$$

where $\beta = (\alpha_1, \alpha_2, \cdots, \alpha_M)$ is the mixture of proportions for each $i = 1, 2, \cdots, M$, and portions should satisfy $\sum_{i=1}^{M} \alpha_i = 1$. $\Theta = (\theta_1, \theta_2, \cdots, \theta_M)$ and $\theta_i = (\bar{x}_i, C_i)$ in Eq. (3). $P(r|\bar{x}_i, C_i)$ is a distribution with emphasizing covariance

matrix C_i and mean $\bar{x}_i$. Let $\Phi = (\beta, \Theta)$ donate the mixture model parameters, then the equation becomes

$$P(r|\Phi) = \sum_{i=1}^{M} \alpha_i P(r|\theta_i), \quad (4)$$

We assume in this part that the parameters derived from seismic data fit a Gaussian probability distribution; therefore, the seismic data, which contain both quarry blast and earthquakes, are samples of a mixture model. To identify them, we need to find the Gaussian models and the model from which each datum is sampled. The theoretical framework and explicit solutions for Gaussian mixture models have been originally introduced by Hasselblad (1666, 1969). The commonly used expectation-maximization (EM) algorithm (McLachlan and Peel, 2000) is utilized to derive the parameters of the mixture model distribution (see Appendix A).

3.2 k-means method

Like Gaussian mixture analysis, k-means clustering is a technique of cluster analysis that aims to separate i events from within the g cluster, in which each event belongs to cluster with nearest mean. k-means is a procedure in the form of stochastic hill climbing in the log-likelihood function. Of the various techniques, k-means can be used to simplify the computation and accelerate convergence. k-means uses the distance measure to assign each event to the nearest cluster based on the proximity to its mean.

k-means is an iterative two-step algorithm (MacKay, 2003). In the assignment step, each datum j is assigned to the nearest mean. First guess for the cluster g_J that the point r_J belongs to by $\tilde{g}_J$ is

$$\tilde{g}_J = \arg\min_g \{d(\mathbf{\Theta}_J, r_J)\} \quad (5)$$

where each cluster is parameterized by a vector $\mathbf{\Theta}$ and d is a metric that defines distances between points, which is defined as

$$d(a,b) = \frac{1}{2}\sum_i (a_i - b_i)^2. \quad (6)$$

Alternatively, assignments of data to cluster can be represented by an indicator z_j^g where it is assigned to one, if mean $(\bar{x}_i)$ is the closest mean to r_J; otherwise z_j^g is zero.

$$z_j^g = \begin{cases} 1 \text{ if } & \tilde{g}_J = g \\ o \text{ if } & \tilde{g}_J \neq g \end{cases} \quad (7)$$

In the updated step, model parameters (Θ) are adjusted to match the sample means of the data points that they belong to by Eq. (8):

$$\Theta_g = \frac{\sum_j z_j^g r_j}{\sum_j z_j^g} \quad (8)$$

This repeats until the assignments do not change (Spath, 1985; Seber, 1984).

3.3 Discriminant classification

Each class generates data using a multivariate normal distribution.

- The model has the same covariance matrix for each class; only the means vary for the LDA.
- Both means and covariance of each class vary for the QDA.

Classification is achieved by minimizing the expected classification cost:

$$\hat{y} = \arg\min_{y=1,....K} \sum_{k=1}^{K} \hat{P}(k|x) C(y|k) \quad (9)$$

where $\hat{y}$ is the predicted classification;
K is the number of classes;
$\hat{P}(k|x)$ is the probability of class k for observation x;
$C(y|k)$ is the cost of classifying an observation as y when its true class is k.

4 Results and discussions

There are rich and various ranges of other clustering techniques in the literature, such as exclusive/non-exclusive, complete/partial, hierarchical/partitioned, and fuzzy. In this study, three clustering techniques are analyzed: Gaussian mixture model, k-means, and two discriminant functions. From a practical point of view in clustering, applying a threshold value to a raw dataset appears to be the easiest way (Fig. 2). However, we could not find an effective value that separates the data into two clusters. In Table 1, the descriptive statistical information about calculated parameters indicates that averages of complexity and S/P ratio are quite higher for earthquakes than quarry blasts. Large standard deviations in both parameters made it difficult to distinguish clusters with a threshold value. Correlation coefficients between the discriminants for just earthquakes, quarry blasts and total are 0.39, 0.79 and 0.83, respectively. The correlation between complexity and S/P ratio is twice for quarry blasts compared to earthquakes.

We have employed a probability density function (Gaussian mixture model) of the seismic events as criterion in order to determine clusters. 2-D probability density function contour plots are shown in Fig. 3. Seismic events are drawn from a two-dimensional Gaussian lie in two clouds centered on the means μ_i. Note that axes in the figure are in log scale. The relations between complexity and S/P ratio of the events are also plotted in Fig. 3. The method classifies the event into two categories according to central distance. The ellipses show lines of equal probability density of the Gaussian. Also

Table 1. Fundamental descriptive statistical information about the data and their ranges.

	Earthquakes		Quarry Blasts		Total	
	Complexity	S/P ratio	Complexity	S/P ratio	Complexity	S/P ratio
Mean	8.99	2.58	1.18	0.61	2.44	0.93
Standard deviation	5.03	0.87	1.39	0.29	3.73	0.85
Maximum	22.13	4.47	14.42	2.28	22.13	4.47
Minimum	1.58	1.00	0.20	0.06	0.20	0.06
Correlation Coef.	0.39		0.79		0.83	

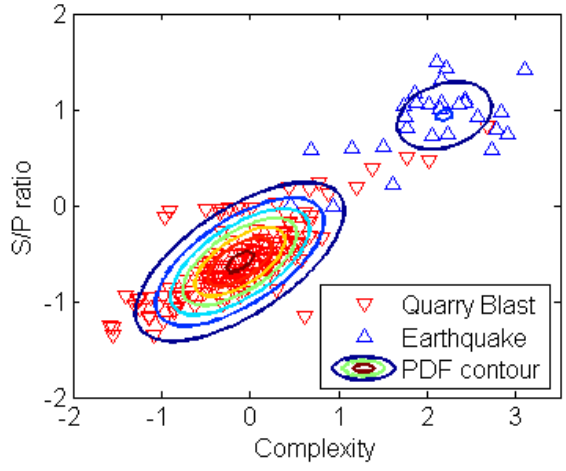

Fig. 3. Seismic events drawn from a two-dimensional Gaussian lie in two clouds centered on the means μ_i. The method classifies the event into two categories according to central distance. The ellipses show lines of equal probability density of the Gaussian. Values are in log scale.

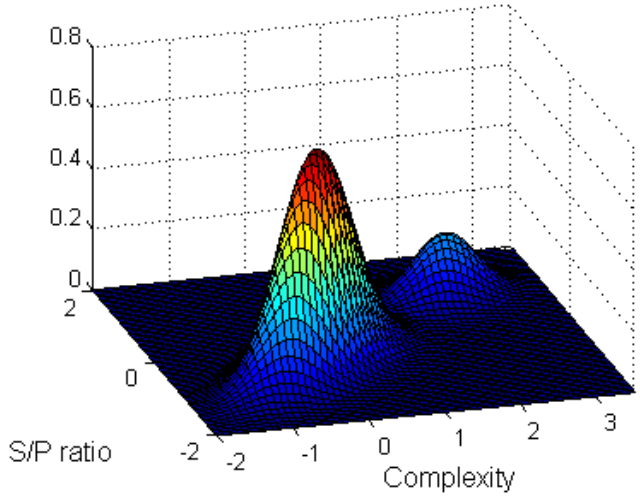

Fig. 4. Contours of probability density function (PDF) of the Gaussian mixture model.

a PDF of events as a function of complexity and S/P ratio is illustrated as 3D plot in Fig. 4. A closer look at this figure gives us an interesting threshold for recognition of events. Neglecting the complexity parameter, a threshold of 0.55 for the S/P gave four misclassifications in total. However, we think this is a database-dependent value that might not be valid for any other dataset/region.

The clustering results after applying of Gaussian mixture models are shown in Fig. 5. Blue and red colors indicate earthquakes and quarry blasts, respectively after classification. The black circles show the misclassified events. Five QBs and two EQs are misclassified by the method. The accuracy of classifier for over all events is 96.1 %. The performance of the algorithms for earthquakes is 82.8 % whereas for quarry blasts it reaches 98.7 % (number of correct classifications/number of the event type).

The k-means clustering fundamentally assumes that the events come from spherical Gaussian distributions, and thus other types of statistical distributions may not cluster correctly using the technique. Therefore, k-means algorithm does not identify the attributes that are more significant in the clustering process as it assumes that all attributes have the same weight. This is probably why it gave the worst result. The clustering results of k-means are shown in Fig. 6. Blue and red colors show earthquakes and quarry blasts, respectively. The black circles indicate the misclassified events. Three QBs and six EQs are misclassified by the method. The accuracy of classifier for over all events is 95.0 %. The performance of the algorithms for earthquakes is 89.7 % whereas for quarry blasts it reaches 96.0 %. The cluster centroids are located at [0.92, 0.56] and [8.69, 2.78], and the within-cluster sums of point-to-centroid distances are 103.89 and 98.119 for QBs and EQs respectively. Compared to previous algorithm, five of the misclassified events are the same where three earthquakes with the highest complexity and S/P ratio and the same two quarry blasts are shown in Fig. 5. Algorithm achieved to identify two quarry blasts with the lowest complexity and S/P ratio of misclassified earthquakes (in Fig. 5) where Gaussian mixture model could not be recognized. However, four new additional misclassified earthquakes are recognized by k-means.

Table 2. Comparison of the methods.

Methods	# of events	# of QB	Misclassified QB	% of accuracy for QB	# of EQs	Misclassified EQ	% of accuracy for EQ	% of total accuracy
Gaussian mixture model	179	150	5	96.7	29	2	93.1	96.1
k-means	179	150	3	98.0	29	6	79.3	95.0
LDF	179	150	2	98.7	29	5	82.8	96.1
QDF	179	150	4	97.3	29	2	93.1	96.6

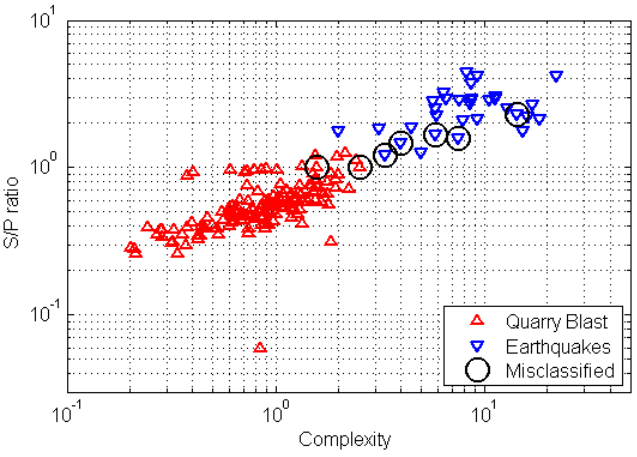

Fig. 5. The results after applying of Gaussian mixture model. Blue and red colors indicate earthquakes and quarry blasts, respectively after classification. The black circles show the misclassified events. Five QBs and two EQs are misclassified by the method. The accuracy of classifier is 96.1 %.

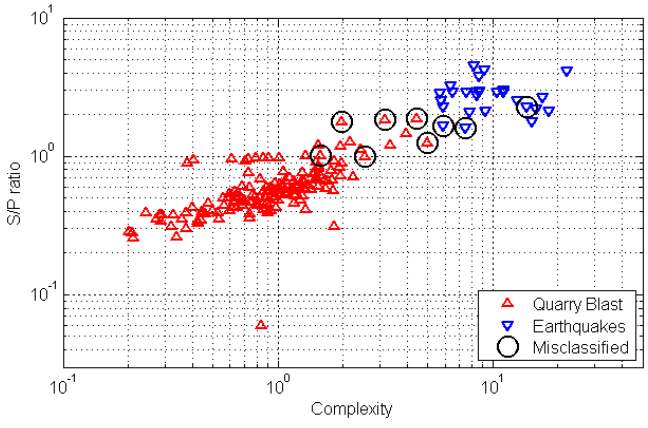

Fig. 6. The results after applying of k-means algorithm. Blue and red colors indicate earthquakes and quarry blasts, respectively after classification. The black circles show the misclassified events. Three QBs and six EQs are misclassified by the method. The accuracy of classifier is 95.0 %.

Results of two cluster analyses, which are based on unsupervised technique, used metric rather than a target classification. Thus, we compare their responses with pre-defined targets that were selected by authors manually. Kuyuk et al. (2011) applied an unsupervised algorithm, called self-organizing map (SOM) as a neural classifier for the same region using the partially similar discriminants. Although they used extra two parameters (spectral ratio and origin time of events) for better classification, their results indicated that these two are fuzzy and misleading classifiers. SOMs reach up to about % 94 accuracy for their problem. Two methods in the present study achieved better success where applications of algorithm are rather simple and faster than SOM. We hope these approaches could be employed routinely in observatory practice in Marmara region. In order to apply the methodologies to other regions, different statistics as potential classifiers could give better results, because selection of discriminants and methods are specific in this study. Another point needs to be considered in the application that the discriminants should be normally distributed to fulfill requirements of the two models. Strong abnormality from this assumption might result in indiscriminate results of clustering without a meaningful interpretation.

We applied a traditional, discriminant analysis that uses input events to estimate the parameters of discriminant functions of the predictor variables. Discriminant functions determine the boundaries, in predictor space, between two classes. The resulting classifier then discriminates among the classes based on the predictors. Decision boundaries of clustering represent the optimal trade-off between performance on the dataset and simplicity of classifier, thereby giving the highest accuracy using different algorithms: (a) linear discriminant functions and (b) quadratic discriminant functions (QDF) are shown in Fig. 7. Blue and red colors indicate earthquakes and quarry blasts, respectively. The overall accuracies are 96.1 % and 96.6 %, respectively (see Fig. 7). Quadratic discriminant analysis resembles LDF, where it is assumed that there are only two classes of points, and that the measurements are normally distributed. Unlike LDF however, in QDF there is no assumption that the covariance of each of the classes is equal. LDF and Gaussian mixture model have the best estimation with two misclassifications for quarry blasts. However, this is not the case for earthquakes. They gave the worst estimation with 82.8 % accuracy, whereas QDF produces two misclassifications. Below, LDF and QDF are given:

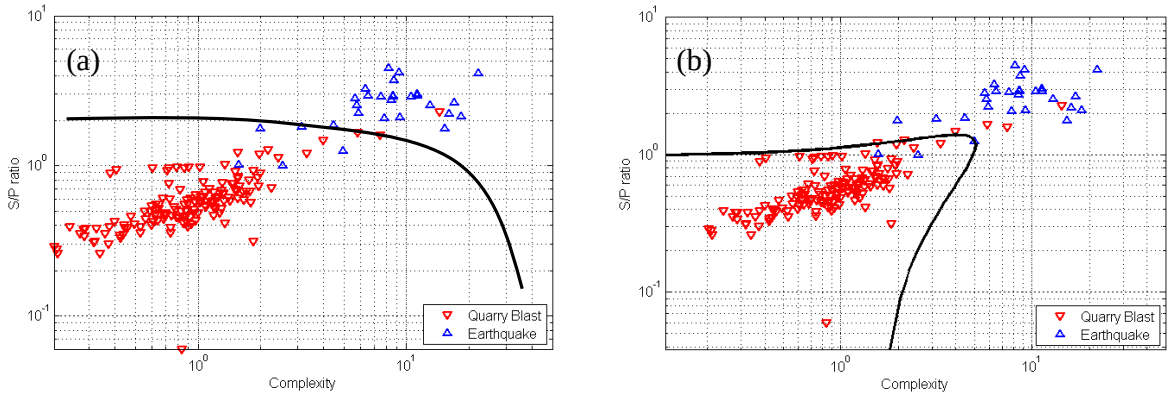

Fig. 7. Classification with discriminant functions: decision boundaries (black lines) shown represent the optimal trade-off between performance on the dataset and simplicity of classifier, thereby giving the highest accuracy. The two algorithms **(a)** linear discriminant functions **(b)** quadratic discriminant functions are presented. Blue and red colors indicate earthquakes and quarry blasts. The overall accuracies are 96.1 % and 96.6 % respectively.

$$f = 16.82 + \begin{bmatrix} -0.56 & -8.77 \end{bmatrix} \begin{bmatrix} C \\ \mathrm{SP} \end{bmatrix} \tag{10}$$

$$f = 4.34 + \begin{bmatrix} -1.61 & 9.73 \end{bmatrix} \begin{bmatrix} C \\ \mathrm{SP} \end{bmatrix} + \sum \begin{bmatrix} C & \mathrm{SP} \end{bmatrix} \begin{bmatrix} -0.63 & 2.40 \\ 2.40 & -14.41 \end{bmatrix} \begin{bmatrix} C \\ \mathrm{SP} \end{bmatrix} \tag{11}$$

and criteria for the function are given below:

$$F = \begin{cases} \mathrm{EQ} & f < 0 \\ \mathrm{QB} & f \geq 0 \end{cases} \tag{12}$$

Horasan et al. (2009) used LDF on S/P ratio with log S and complexity with Sr (spectral ratio) together to discriminate the earthquakes and the quarry blasts in the same region. They obtained similar results using LDF analysis from the amplitude ratio with log S. They investigated the region by dividing the events into the four locations. On the contrary, this study evaluated all events at once and the predictors executed using whole data. This study gives a more general image, and these clustering techniques are more promising for the vicinity of Istanbul.

5 Conclusions

Two unsupervised pattern recognition algorithms, k-means and Gaussian mixture model (GMM) analyses, for which we could not find any similar study of this type of clustering problems, have been applied to classify seismic events in the vicinity of Istanbul. We have found that unsupervised clustering algorithms, for which no a-prior target information is available, display a similar discriminatory power as supervised methods of discriminant analysis. It is fascinating that GMM gave the same total accuracy with LDF, for which LDF uses a-priori target information to train. Furthermore, the two unsupervised techniques are very promising and straight-forward compared to the other unsupervised algorithms, such as the well-known self-organizing maps. On the other hand, k-means and GMM are not a panacea for the classification of different seismic events but are two likely methods that can be easily employed in other earthquake active regions with even more sophisticated datasets. Additional discriminants would cause problems across further dimensions, whereas this study dealt with only 2-D. Applications of all algorithms into practice, once they were set up, are quite simple and computationally inexpensive. Seismologists, certainly with some caution, may use unsupervised algorithms in regions where no initial information is available.

Appendix A

Expectation maximization algorithm

Supposing that a set of measurements (discriminants in this study) $R = (r_1, r_2, \cdots, r_M)$ are samples of Gaussian mixture model of Eq. (4). The log-likelihood function of model parameters Φ with the data R is

$$L(\Phi|R) = \ln[P(R|\Phi)]. \tag{A1}$$

The maximum likelihood estimate of Φ can be easily determined if we know from which Gaussian model each datum is sampled – called the complete-data problem. Due to the nature of unsupervised learning, no a priori information is available on grouping, which makes it an incomplete-data problem and can be solved by iteratively using an EM algorithm. Alternatively, the log-likelihood function can be written as

$$L(\Phi_n|\mathrm{R}) = Q(\Phi_n|\Phi_{n-1}) + \Delta(\Phi_n|\Phi_{n-1}) \quad \text{(A2)}$$

where n stands for the iteration number and Q is the expectation of complete-data likelihood function given by

$$Q(\Phi_n|\Phi_{n-1}) = E_g\{\ln[P(R,G|\Phi_n)]|\mathrm{R},\Phi_{n-1}\} \quad \text{(A3)}$$

and Δ is the difference between the incomplete-data (Eq. A1) and complete-data (Q) log-likelihood. $G = (g_1, g_2, \cdots, g_N)$ where g_j is a group indicator (quarry blast or earthquake) for r_j. E_g symbolizes a conditional expectation value.

If r_j belongs to i-th group, then $g_j = i$.

Due to difficulties in obtaining maximum likelihood estimate of the Φ by directly maximizing of Eq. (A1), EM algorithm improves $L(\Phi|\mathrm{R})$ by increasing Q.Q and can also be written as

$$Q(\Phi_n|\Phi_{n-1}) = \sum_{j=1}^{N}\sum_{i=1}^{M} h_j^i(\Phi_{n-1}) \ln\left[\alpha_i P(r_j|\gamma_j^i = 1, \Phi_n)\right] \quad \text{(A4)}$$

where $\gamma_j^i = 0$, if $g_j \neq i$ and is equal to 1, if $g_j = i h_j^i(\Phi_n)$is the expectation of γ_j^i with given data. M is number of given model and Φ_n at previous iteration defined as

$$\begin{aligned} h_j^i(\Phi_n) &= E_g\left\{\gamma_j^i|R,\Phi_n\right\} \\ &= P(\gamma_j^i = 1|r_j,\Phi_n) \\ &- \frac{(\alpha_i)_n P\left(r_j|(\bar{x}_j)_n,(Z_j)_n\right)}{\sum_{t=1^M}(\alpha_t)_n P\left(r_j|(\bar{x}_t)_n,(Z_t)_n\right)} \end{aligned} \quad \text{(A5)}$$

where

$$\begin{aligned} &P\left(r_j|(\bar{x}_i)_n,(Z_i)_n\right) \\ &= \frac{1}{\sqrt{2\pi\det(Z_i)_n}} \exp\left[-\frac{1}{2}(r_j-(\bar{x}_i)_n)^T (Z_i)_n^{-1}(r_j-(\bar{x}_i)_n)\right]. \end{aligned} \quad \text{(A6)}$$

At expectation step, the membership matrix is calculated for the $(n+1)$-th iteration by the above equation. $h_j^i(\Phi_n)$ defines the probability of the j-th datum belonging to i-th event. In this step, entries $(h_j^i(\Phi_n))$ of the expectation membership matrix $\mathbf{H}_{NxM} = H(\Phi_n)$ are constructed for a given Φ_n.

In the next step, maximization step, the mixture model parameters are estimated by maximizing the Q defined in Eq. (A3) using the membership matrix, which is estimated from the previous step as

$$\Phi_{n+1} = \arg\max Q(\Phi_n|\Phi_{n-1}). \quad \text{(A7)}$$

Solutions for the Gaussian mixture model parameters are

$$(\bar{x}_i)_{n+1} = \frac{\sum_{j=1^N}(h_j^i)_n r_j}{\sum_{j=1^N}(h_j^i)_n}, \quad \text{(A8)}$$

$$(C_i)_{n+1} = \frac{\sum_{j=1^N}(h_j^i)_n (r_j-(\bar{x}_i)_{n+1})(r_j-(\bar{x}_i)_{n+1})^T}{\sum_{j=1^N}(h_j^i)_n}, \quad \text{(A9)}$$

$$(\alpha_i)_{n+1} = \frac{1}{N}\sum_{j=1^N}(h_j^i)_n. \quad \text{(A10)}$$

These results are used as new entries for next E-step until the pre-defined convergence is reached (Han et al., 2011).

Acknowledgements. H. S. Kuyuk thanks Earthquake Research Institute, University of Tokyo for hosting as a visiting scholar while part of manuscript is written in the institute. We also thank referees for their constructive critics.

Edited by: L. Telesca
Reviewed by: Q. Wang and two anonymous referees

References

Baumgardt, D. R. and Young, G. B.: Regional seismic waveform discriminates and case-based event identification using regional arrays, B. Seismol. Soc. Am., 80, 1874–1892, 1990.

Bennett, T. J. and Murphy, J. R.: Analysis of seismic discrimination capabilities using regional data from western United States events, B. Seismol. Soc. Am., 76, 1069–1086, 1986.

Bennett, T. J., Barker, B. W., McLaughlin, K. L., and Murphy, J. R.: Regional discrimination of qarry blasts, earthquakes and underground nuclear explosions, Final Report, GL-TR-89-0114, S-Cubed, La Jolla, California, 1989.

Campus, P. and Fah, D..: Seismic monitoring of explosions: a method to extract information on the isotropic component of the seismic source, J. Seismol., 1, 205–218, doi:10.1023/A:1009781722363, 1997.

Del Pezzo, E., Esposito, A., Giudicepietro, F., Marinaro, M., Martini, M., and Scarpetta, S.: Discrimination of earthquakes and underwater explosions using neural Networks, B. Seismol. Soc. Am., 93, 215–223, 2003.

Dowla, F. U., Taylor, S. R., and Anderson, R. W.: Seismic discrimination with artificial neural networks: preliminary results with regional spectral data, B. Seismol. Soc. Am., 80, 1346–1373, 1990.

Duda, R. O., Hart, P. E., and Stork, D. G.: Pattern classification, ISBN SSN:0-471-05669-3, 2nd Edn., Wiley, New York, 2001.

Falsaperla, S., Graziani, S., Nunnari, G., and Spampinato, S.: Automatic classification of volcanic earthquakes by using multilayered neural Networks, Nat. Hazards, 13, 205–228, 1996.

Gitterman, Y., Pinky, V., and Shapira, A.: Spectral classification methods in monitoring small local events by the Israel seismic network, J. Seismol., 2, 237–256, 1998.

Han, M., Zhao, Y., Li, G., and Reynolds, A. C.: Application of EM algorithms for seismic facices classification, Comput. Geosci., 15, 421–429, doi:10.1007/s10596-010-9212-4, 2011.

Hasselblad, V.: Estimation of Parameters for a Mixture of Normal Distributions, Technometrics, 8, 431–444, 1966.

Hasselblad, V.: Estimation of Finite Mixtures of Distributions from the Exponential Family, J. Am. Stat. Assoc., 64, 1459–1471, 1969.

Horasan, G., Bozepe-Güney, A., Küsmezer, A., Bekler, F., and Öğütçü, Z.: İstanbul ve civarındaki deprem ve patlatma verilerinin birbirinden ayırt edilmesi ve kataoglanması (Discrimination and cataloging of quarry blasts and earthquakes in the vicinity of Istanbul), Report 05T202, Boğaziçi University Research Foundation Bebek-İstanbul, 76 pp., 2006.

Horasan, G., Boztepe-Güney, A., Küsmezer, A., Bekler, F., Ögütçü, Z., and Musaoglu, N.: Contamination of seismicity catalogs by quarry blasts: an example from Istanbul and its vicinity, northwestern Turkey, J. Asian Earth Sci., 34, 90–99, doi:1016/j.jseaes.2008.0.012, 2009.

Jenkins, R. D. and Sereno, T. S.: Calibration of regional S/P amplitude-ratio discriminants, Pure Appl. Geophys., 158, 1279–1300, doi:10.1007/PL00001223, 2001.

Koch, K. and Fah, D.: Identification of earthquakes and explosions using amplitude ratios: the Vogtland area revisited, Pure Appl. Geophys., 159, 735–757, 2002.

Kuyuk, H. S., Yıldırım, E., Dogan, E., and Horasan, G.: Self Organizing Map Approach for Discrimination of Seismic Event and Quarry Blasts in The Vicinity Of Istanbul, 14th European Conference on Earthquake Engineering, Ohrid, Republic of Macedonia, 30 August–3rd September, 2010.

Kuyuk, H. S., Yildirim, E., Dogan, E., and Horasan, G.: An unsupervised learning algorithm: application to the discrimination of seismic events and quarry blasts in the vicinity of Istanbul, Nat. Hazards Earth Syst. Sci., 11, 93–100, doi:10.5194/nhess-11-93-2011, 2011.

MacKay, D. J. C.: Information Theory, Inference, and Learning Algorithms, Cambridge University Press, 2003.

McLachlan, G. and Peel, D.: Finite Mixture Models, Hoboken, NJ, John Wiley & Sons, Inc., 2000.

Muller, S., Garda, P., Muller, J. D., and Cansi, Y.: Seismic events discrimination by neuro-fuzzy catalogue features, Phys. Chem. Earth, 24, 201–206, 1999.

Musaoğlu, N., Coşkun, M. Z., Göksel, Ç., Kaya, Ş., Bektaş, F., Saroğlu, E., Üstün, B., Ipbüker, C., Erden, T., and Karaman, H.: Istanbul Anadolu yakası hazine arazilerinin uydu verileri ve coğrafya bilgi sistemleri (CBS) ile incelenmesi (Investigation of state owned lands in Anatolian side of Istanbul by satellite data and Geographic Information System), TÜBITAK, Project No: 102I022 (IÇTAG-I 433), 69 pp., 2004.

Musil, M. and Pleginger, A.: Discrimination between Local Microearthquakes and Quarry Blasts by Multi-Layer Perceptrons and Kohonen Maps, B. Seismol. Soc. Am., 86, 1077–1090, 1996.

Pomeroy, P. W., Best, J. W., and McEvilly, Th. V.: Test ban treaty verification with regional data – a review, B. Seismol. Soc. Am., 72, 89–129, 1982.

Otsubo, M., Yamaji, A., and Kubo, A..: Determination of stresses from heterogeneous focal mechanism data: An adaptation of the multiple inverse method, Tectonophysics, 457 150–160, 2008.

Ouillon, G., Ducorbier, C., and Sornett, D.: Automatic reconstruction of fault networks from seismicity catalogs: Three-dimensional optimal anisotropic dynamic clustering, J. Geophys. Res., 113, B01306, doi:10.1029/2007JB005032, 2008.

Rodgers, A. J. and Walter, W. R.: Seismic discrimination of the May 11, 1998 Indian nuclear test with short-period regional data from Station NIL (Nilore, Pakistan), Pure Appl. Geophys., 159, 679–700, 2002.

Scarpetta, S., Giudicepietro, F., Ezin, E. C., Petrosino, S., Del Pezzo, E., Martini, M., and Marinaro, M.: Automatic Classification of seismic signals at Mt. Vesuvius Volcano, Italy using Neural Networks, B. Seismol. Soc. Am., 95, 185–196, 2005.

Seber, G. A. F.: Multivariate Observations, Hoboken, NJ, John Wiley & Sons, Inc., 1984.

Spath, H.: Cluster Dissection and Analysis: Theory, FORTRAN Programs, Examples, translated by: Goldschmidt, J., New York, Halsted Press, 1985.

Tiira, T.: Detecting teleseismic events using artificial neural networks, Comput. Geosci., 25, 929–939, 1999.

Ursino, A., Langer, H., Scarfı, L., Di Grazia, G., and Gresta, S.: Discrimination of quarry blasts from tectonic microearthquakes in the Hyblean Plateau (southeastern Sicily), Ann. Geofis., 44, 703–722, 2001.

Weatherill, G. and Burton, P. W.: Delineation of shallow seismic source zones using K-means cluster analysis, with application to the Aegean region, Geophys. J. Int., 176, 565–588, 2009.

Wiemer, S. and Baer, M.: Mapping and removing quarry blast events from seismicity catalogs, short notes, B. Seismol. Soc. Am., 90, 525–530, 2000.

Wüster, J.: Discrimination of chemical explosions and earthquakes in central Europe- a case study, B. Seismol. Soc. Am., 83, 1184–1212, 1993.

Yıldırım, E., Gulbag, A., Horasan, G., and Dogan, E.: Discrimination of quarry blasts and earthquakes in the vicinity of Istanbul using soft computing techniques, Comput. Geosci., 37, 1209–1217, 2011.

Zeiler, C. and Velasco, A. A.: Developing Local to Near-Regional Explosion and Earthquake Discriminants, B. Seismol. Soc. Am., 99, 24–35; doi:10.1785/0120080045, 2009.