Support Vector Machine Model for Automatic Classification of Seismic Events in the region of Funil Reservoir, Minas Gerais, State, Brazil

Vesna Barros[1], Lucas Barros[1], Juraci Carvalho[1].
1- Seismological Observatory of the University of Brasilia, Brasília – DF - Brazil

Abstract

The present study introduces the first steps of a Support Vector Machine (SVM) algorithm using two discrimination features: spectral ratio and waveform complexity on a dataset characterized by lower magnitude events recorded by a local network in Minas Gerais State, Brazil. The selected data set consists of 43 microearthquakes and 39 mining blasts of similar magnitudes and locations. The analysis was carried out using data of vertical short period components of two 3C seismic stations. Our results revealed that despite of some overlapping, these techniques show a good capacity to discriminate artificial by natural events for the studied region. However, for short distances (~3 km), the technique deserves more attention.

Introduction

The study area is located close to the Funil Reservoir power plant, in the the Minas Gerais State, south of the São Francisco Craton, where there are low-magnitude seismic events triggered as well as artificial events generated by several active quarries in the region (Fig. 1) (Barros et al., 2014).

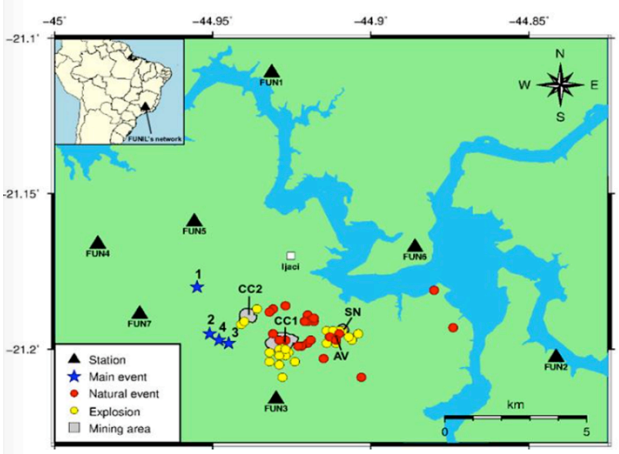


Figure 1: Red circles are natural events and yellow circles are artificial events. CC1, CC2, AV and SN represent the location of quarries close to the Funil Reservoir. The local seismic network is composed by seven 3C stations represented by the black triangles.

The classification of the seismic events is challenging when natural and anthropogenic seismicity overlap in magnitude, space and time. The need of an automatic event classifier for monitoring local and regional seismicity increases since the visual screening phase is very difficult and time-consuming for analysts.

SVMs (Support Vector Machines) are a useful technique for data classification. A classification task usually involves separating data into training and testing sets. Each instance in the training set contains one "target value" (i.e. the class labels) and several "attributes" (i.e. the features). The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes (Hsu, C.-W. et al., 2010).
Training vectors x_i are mapped into a higher dimensional space by the function φ. The goal is to find a linear separating hyperplane with the maximal margin in this higher dimensional space. $C > 0$ is the penalty parameter of the error term, also known as cost factor. Furthermore, $K(x_i,x_j) \equiv \varphi(x_i)T\varphi(x_j)$ is called the kernel function (Cortes and Vapnik, 1995). In this study, the relation between attributes for two independent data sets led us to use two different kernels:

• Linear: $K(x_i,x_j) = x^T_i\, x_j$.

• Radial basis function (RBF): $K(x_i,x_j) = \exp(-||x_i - x_j||2)$, $\gamma>0$.

Here, γ is a kernel parameter. Together with C, they are the two main parameters of a kernel function. Because it is not known beforehand which C and γ are best for a given problem, a parameter search must be done ("Grid-search"). The goal is to identify good (C,γ) so that the classifier can accurately predict unknown data (i.e. testing data). This can be done through cross-validation method and, after the best (C, γ) is found, the whole training set is trained again to generate the final classifier.

Spectral features of typical earthquakes and explosions

In seismic analysis, the time-frequency distribution plots, spectrograms, represent a useful tool for discrimination between natural and artificial seismicity. They reveal time and frequency dependent variations in the signal energy distribution and also display relative amplitudes of seismic phases.

Earthquakes are volume sources extended both in time and space and they generate a larger fraction of energy in S waves than in P waves. Their seismic waves have wide frequency content and their energy is evenly distributed over the whole recorded frequency band. Earthquakes also produce rather complex waveforms because of

secondary depth-sensitive seismic phases in their P and S coda (Kortstrom et al., 2015).

In contrast to earthquakes, explosions are compressive point sources from which P wave energy radiates evenly to all azimuth directions. They have smaller energy content as well as lower dominant frequencies than the corresponding P waves. In comparison to an earthquake of similar magnitude, the explosions have narrower frequency content as well as shorter duration of P and S wave coda (Fig. 2).

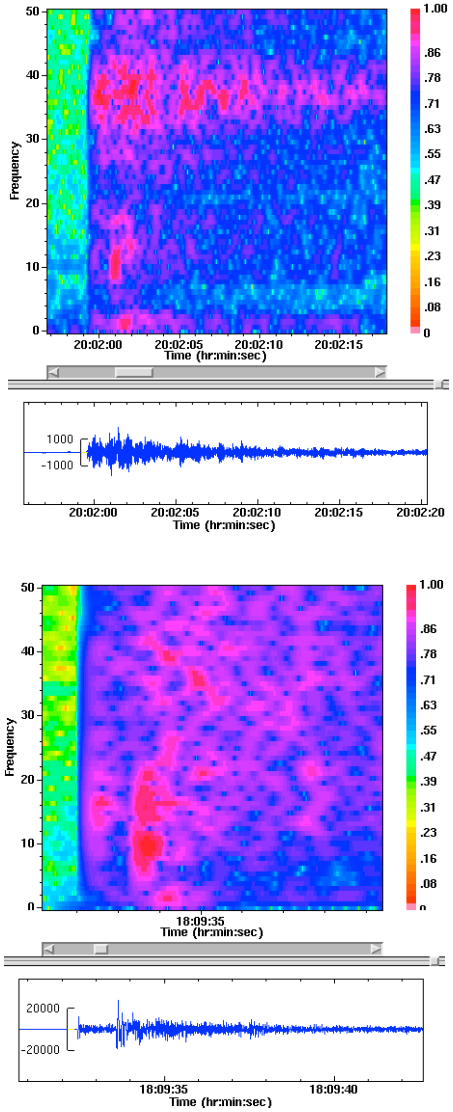


Figure 2: Examples of spectrograms and traces computed for single explosion (at the top) and earthquake (at the bottom). The explosion spectrogram displays bigger P to S ratios, shorter duration of P and S wave coda and more concentrated distribution of signal energy content (around the frequency of 40 Hz) than a shallow earthquake of similar size in Fig. 2b, which displays bigger S to P ratio and the energy is evenly distributed over the whole recorded frequency band.

Event discrimination

Before choosing the features for the classification problem and due to the presence of multiple sources of blasts in the region, cross-correlation function was used in the software SEISAN (Havskov and Ottemöller, 2008) to provide a measure of similarity between different seismic events recorded in each station and, subsequently, to improve our understanding of the nature of each event.

In our study, a high correlation coefficient (>0.6) means a high waveform similarity, which is caused by proximity in hypocenter location and similarity in focal mechanism between two events. Once the cross-correlation was done between the seismic signals, we could identify groups of events that originated from different blasts sites, as well as the group of natural events (earthquakes) originated from the fault line (Fig. 3).

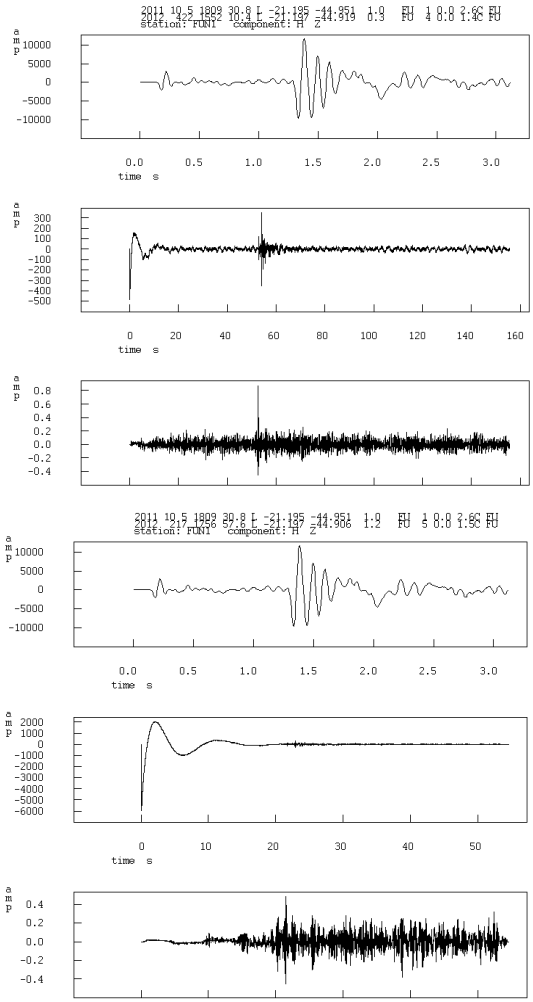


Figure 3: Correlation between events in software SEISAN. The first two traces show the signals being correlated, the third gives the normalized correlation and amplitude is less than 1. At the top, the signals of two explosions are being correlated (maximum correlation coefficient = 0.8). At the bottom, the signals of an explosion and an earthquake show a maximum correlation coefficient of 0.4.

In addition to cross-correlation, all available sources of information in event discrimination were used: i) clear first motion polarities of P waves; ii) reports emitted by the mines with the date and time of the scheduled explosions (not so precise and regular) and iii) time of event occurrences (daytime or nighttime). Normally explosions are compressive (first movement of the P-wave are up) and occur during working hours, whereas earthquakes can happen anytime.

For stations very close to the events, the discrimination of the signals can still be very difficult and confusing, especially when P- and S- phases cannot be separated. Figure 4 shows the similarity between the signals of an artificial and a natural event. The seismic station that recorded these signals (FUN1) was located approximately 10 km of the events.

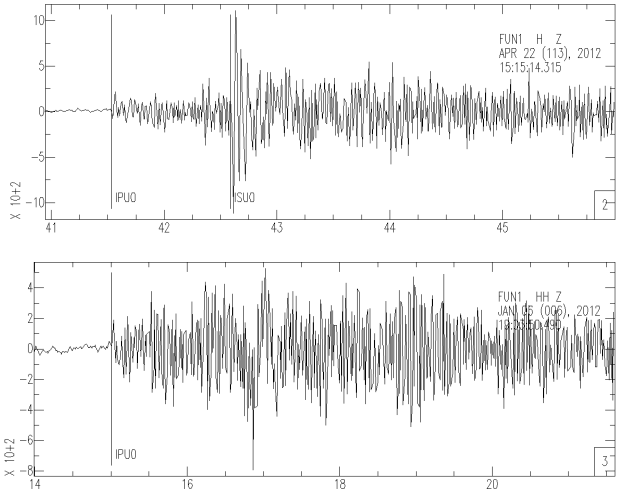


Figure 4: Similarity between events. Seismogram of an earthquake (mc=1.9) at the top and of an explosion (mc=2.1) at the bottom.

Data and methods applied in SVM classification

For this study, we have not set a lower limit to the number of stations that recorded an event because this would have excluded a large portion of the training data set. The selected stations, FUN1 and FUN3, were chosen due to the signal quality of their seismograms. In total, we selected the vertical components of 25 earthquakes and 25 mining blasts that were located by FUN1, and 25 earthquakes and 25 mining blasts located by FUN3. Therefore, two independent data sets consisting of 50 events were used for each station. Together, they compose a total of 82 events (43 earthquakes and 39 mining blasts – some events were recorded by both stations, others just by one of them) with similar locations and magnitude mc ranging from 0.3 to 2.6.

The SVM models are station-specific and depend on the relation between the attributes (features) calculated for FUN1 and FUN3 stations. The first feature used was waveform complexity (C), which can be calculated by comparing the energy carried in different window lengths of the seismogram. It is, by definition, the ratio of the seismogram's integrated powers S2(t) in the selected time windows (t2 – t1 and t1 – t0):

$$C = \int_{t_1}^{t_2} S^2(t)\,dt \Big/ \int_{t_0}^{t_1} S^2(t)\,dt$$

The time limits of the integrals (t0, t1 and t2) of C were determined by a trial-and-error approach to find the best representative C values for both blasts and earthquakes of similar magnitudes. For both stations, FUN1 and FUN3, the duration of the signal is very short, so the selected time window was t0=0 s, t1=1 s and t2=2 s, where t0 is the P-wave onset time.

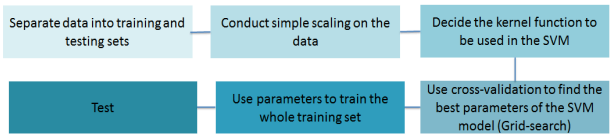
The second feature (SR) is the ratio of the seismogram's integrated spectral amplitudes a(f) in the selected frequency bands (high-frequency band, h1-h2, and low-frequency band, l1 – l2). SR can be written as below:

$$SR = \int_{h_1}^{h_2} a(f)\,df \Big/ \int_{l_1}^{l_2} a(f)\,df$$

The integrals limits (h1, h2, l1 and l2) used in the calculation of SR were determined comparing the spectra of quarry blasts with those of earthquakes. We tested for different frequency ranges to find the spectral frequency band where the spectral ratio has a maximum efficiency.

These parameters are commonly used in the classification of regional and teleseismic events e.g. Horasan et al. (2009) and Yilmaz et al. (2013). In this study, we intent to use the same classification method applied in a local seismograph network.

LIBSVM (Chih-Chung Chang and Chih-Jen Lin, 2010) software package was used in MATLAB to obtain the classifying models. The procedure used in our work followed exactly what is recommended in the LIBSVM practical guide:



From the total dataset of 50 events for each station, 35 were selected for training the classifier and the 15 left were used to test its performance. In order to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges, all attributes were linearly scaled to the range [0, 1].

Results and Discussion

Plotting the results obtained for the calculation of C and SR for FUN1 and FUN3 stations, we obtained the following graphs (Fig. 5):

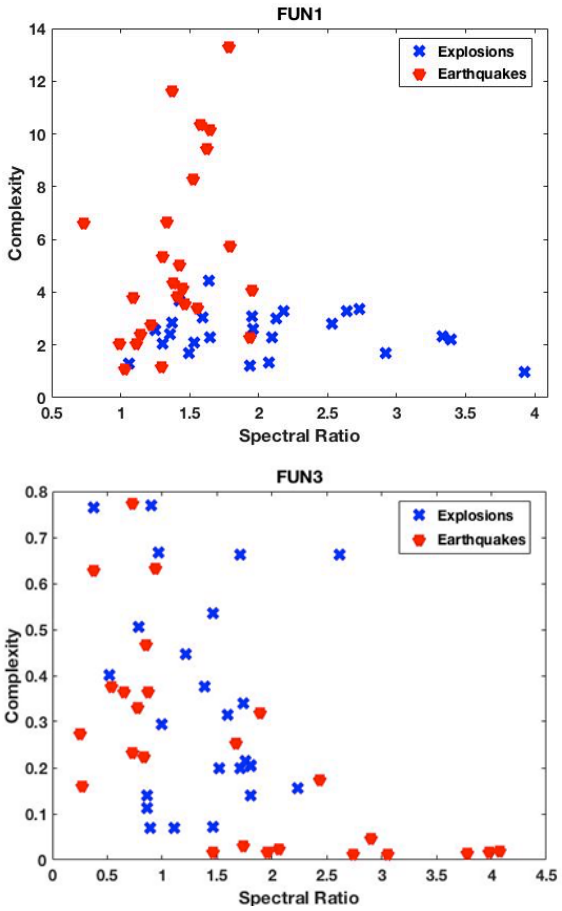


Figure 5: At the top, C x SR obtained for FUN1. At the bottom, for FUN3.

As it can be seen in figure 5 (left), it is possible to visually distinguish the relation between features in FUN1. However, data points from different classes remain mixed between each other. In general, the waveform complexity of earthquakes was higher than of explosions, which is expected according to similar studies. In this case, the distance from the events to FUN1 was about 10 km. This is a reasonable distance to visually discriminate P- and S-phases in the seismogram (recall Fig. 4), and the waveform complexity calculation works here.

For FUN3, on the other hand, we can't see a clear relation between explosions and earthquakes in the plot. Because the distance from FUN3 to the event was so short (about 3 km), P- and S-phases overlap. In fact, the spectrogram of a shallow earthquake recorded by FUN3 displayed explosion-like characteristics, i.e. large P to S rations. For this reason, some events might have been misidentified during the discrimination analysis.

Even though there might not be a relation between attributes for FUN3, we wanted to explore what could be done with the results. Therefore, a linear SVM was used for FUN1 and a SVM with a RBF kernel for FUN3. According to Chang, C.-C, and Lin, C.-J, 2010, in general, the RBF kernel is a reasonable first choice. This kernel nonlinearly maps samples into a higher dimensional space so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear.

The grid-search done in FUN3 showed that the best pair (C, γ) was C=512, γ=2 with a cross-validation accuracy of 71.73%. For FUN3, various input cost factor values were tried and the one with the less misclassification rate was picked. The optimized value found was 100, with cross-validation accuracy of 66.67%. Figure 6 shows the plots with the hyperplanes for both cases.

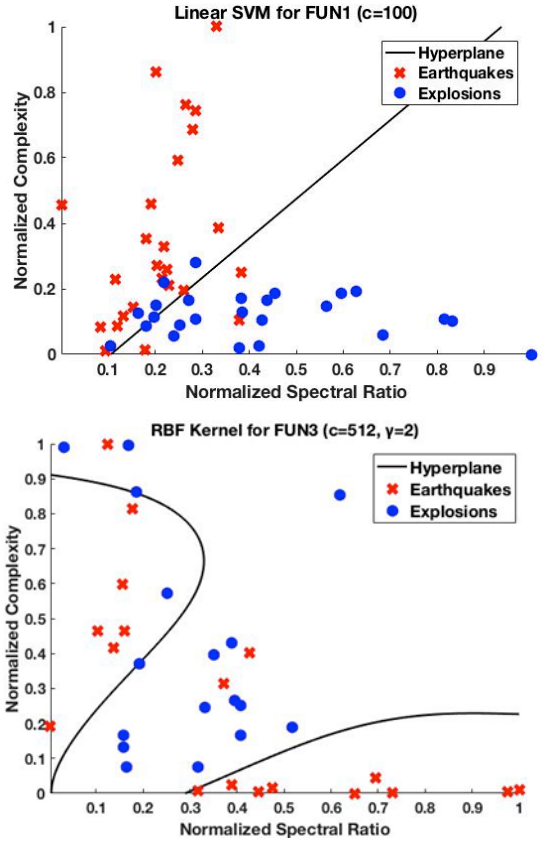


Figure 6: At the top, linear SVM for FUN1. At the bottom, RBF kernel SVM for FUN3.

SVM classification performance

From the total dataset of 50 events for each station, randomly sampling 35 for the training set and 15 for testing test, we reiterated this experiment 100 times. This is a procedure suggested by Kirsopp and Shepperd, 2002, when inferences need to be made based on a small number of dataset. The statistics of the 100 experiments is listed in the table below:

Table 1: SVM classifying recognition rates

Station	Mean (%)	Min Accuracy (%)	Max Accuracy (%)
FUN 1	75.33	53.33	100
FUN 3	70.37	53.33	86.67

The results varied significantly from one training set to another. This gives an indication that the number of sampled data sets might not be enough to gain a particular level of certainty. Therefore, increasing the data set is necessary to get confident inferences from the validation. Moreover, misidentified events in the training data (specially in FUN3) may affect the performance of the SVM classification.

Conclusions

The study achieved satisfactory results based on the small data set used to train the SVM classifier, and gave us an insight of how this discrimination technique works in the studied region. However, further study need to be done on the calculation of features. In general, the SVM classifier showed a reasonable capacity to discriminate artificial by natural events, but it depends on the recording station location. For short distances (≈3 km), the technique deserves more attention.

Future Work

- Analyze stations with a lower limit distance from the events;
- Increase dataset. Making inferences with small number of training sets can lead to random results;
- More discrimination parameters i.e. S-to-P amplitude ratios.

Acknowledgments

The authors wish to thank the Seismological Observatory of the University of Brasilia for the support during the period of study. Special thanks to Prof. Dr. Monica Von Huelsen, head of the Observatory. The Geotool software (CTBTO; http://www.ctbto.org) was used for the preparation of spectrogram plots in Fig.2.

References

Barros, L. V., Carvalho, J. M., Ferreira, V. M., Albuquerque, D. F., Huelsen, M. G. V., Caixeta, D. F., Fontenele, D. P. Determination of source seismic parameters of micro-earthquakes with epicenter in the south of Minas Gerais State - Brazil. In: VI Simpósio da Sociedade Brasileira de Geofísica, Porto Alegre, 2014.

Barros, L. V., Albuquerque, D. F.,Von Huelsen, M.G., Mourão, V., Fontenele, D. P.,Silvas, F. S.,Soares, J.,Soares, W . Seismicity of Ijaci, south of Minas Gerais State, near the FUNIL UHE reservoir and mineral extraction areas. In: 13th International Congress of the Brazilian Geophysical Society, 2013, Rio de Janeiro. Annuals of 13th Int. Cong of the SBGf, 2013.

Borleanu, F., Grecu, B., Popa, M., and Radulian, M. Use of Various Discrimination Techniques to Separate Small Magnitude Events Occurred in the Northern Part of Romania.The 1940 Vrancea Earthquake. Issues, Insights and Lessons Learnt, Springer Natural Hazards, 135-150, 2016.

Chang, C.-C, and Lin, C.-J. LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Cortes, C. and Vapnik, V. Support-vector network. Machine Learning, 20:273–297, 1995.

Havskov, J. & Ottemöller, L. (Eds.). SEISAN: The Earthquake Analysis Software, Version 8.1. Inst. of Solid Earth Physics, University of Bergen, Norway, 2008. 227 p.

Hsu, C.-W., Chang, C.-C., and Lin, C.-J. A Practical Guide to Support Vector Classification, 2010.

Kekovalı K., Kalafat D. and Deniz P. Spectral discrimination between mining blasts and natural earthquakes: Application to the vicinity of Tunçbilek mining area, Western Turkey. International Journal of the Physical Sciences Vol. 7(35), pp. 5339-5352, 13 September, 2012.

Kirsopp, C. and Shepperd, M. Making inferences with small numbers of training sets. IEE Proc. Softw., Vol. 149, No. 5, October-2002

Kuyuk, H.S., Yildirim, E., Dogan, E., Horasan, G., 2011. 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 (1), 93–100.

Sayed Dahy, A. and Gaber Hassib, H. Discriminating Nuclear Explosions from Earthquakes at Teleseismic Distances. European Journal of Applied Sciences 1 (4): 47-52, 2009 ISSN 2079-2077

Yilmaz, S., Bayrak, Y., Çinar, H., 2013. Discrimination of earthquakes and quarry blasts in the eastern Black Sea region of Turkey. J. Seismol. 17 (2), 721–734.

Yin-ju, B., Han-ming, H., Ting-ting, W. A Research on the SVM Classification of Earthquake and Explosion Based upon Seismic Wave Features, 2012.