Seperate Earthqukaes from Nuclear explosions

Sabber Ahamed

October 28, 2017

1 Domain Background

The Comprehensive Nuclear-Test-Ban Treaty (CTBT) is a multinational treaty aim to ban all nuclear activities including explosions, for both civilian and military purposes. More than 180 countries have agreed and signed the agreement. But, forty-four countries that have the existing nuclear capabilities must sign and approve the deal for it to have the force of law. Several recognized monitoring systems are used to control the potential violations of the treaty. Seismic wave observation is one of them and is considered the most reliable method for verifying an underground explosion. Seismic stations typically detect as many as 800,000 earthquakes of worldwide. Compare to this number, explosions especially the number of nuclear explosions are very-low. Seismologists use several observations such as polarization, p-wave, S-wave and surface wave, depth of the events, etc. to distinguish explosions signal from earthquakes. All the process are complex and requires the integration of physical and statistical techniques and human interpretations. Not always all the reviews are right. For example, recent quake from North Korea with magnitude 3.0 was first identified as nuclear explosions both by China and South Korea, but the detailed study of the seismic waves later confirmed that the quake was indeed a natural earthquake [4]. Machine learning approach has been adopted to minimizing error in discriminating mine seismic events from mine blasts [1]. In this project, my primary goal is to develop an improved machine learning model that can separate natural earthquake from nuclear explosions using raw seismogram.

2 Statement of work

I plan to use supervised learning techniques toward the final solution. The primary goal is to develop a machine learning model that can distinguish natural earthquakes from nuclear explosions. Since there are only two types of data (earthquakes and nuclear explosions), the model would be binary classification problem. The performance will be evaluated by accuracy, recall, F-1 and ROC-AUC scores based on the predicted and actual values of the test dataset.

3 Datasets and Inputs

I will use the facilities of Incorporated Research Institutions for Seismology (IRIS) Data Services, and specifically, the IRIS Data Management Center, to access waveforms, related metadata, and derived products used in this study. IRIS Data Services are funded through the Seismological Facilities for the Advancement of Geoscience and EarthScope (SAGE) Proposal of the National Science Foundation under Cooperative Agreement EAR-1261681. The events I will be using for this study are listed in table-1.

Each seismogram is 21 minutes long and starts 1 minute before P-wave arrival and ends 20 minutes after P-wave arrival. Characteristics signal features that will be extracted using librosa [2]- a python package

Table 1: Seismic events used for this study

Event Origin	Date	Magnitude	No. of seismograms	Event type
Banda Sea	Oct. 24, 2017	6.7	3382	Natural earthquake
Sumatra, Indonesia	Dec. 26, 2004	9.0	1193	Natural earthquake
Myanmar India border region	Mar. 12, 2010	5.5	3021	Natural earthquake
Southeast of Ryukyu islands	Aug. 15, 1017	4.9	2644	Natural earthquake
Near coast of chiapas mexico	Sep. 08, 2017	8.1	2722	Natural earthquake
Total			12962	
India	May. 11, 1998	5.2	459	Nuclear explosions
Pakistan	May. 28, 1998	4.8	470	Nuclear explosions
Pakistan	May. 30, 1998	4.6	398	Nuclear explosions
North Korea	Feb. 12, 2013	5.1	3763	Nuclear explosions
North Korea	Jan. 06, 2016	5.1	2640	Nuclear explosions
North Korea	Sep. 03, 2017	6.3	2804	Nuclear explosions
Total			10534	

Table 2: Features for the model

Feature Name	Description	No. of features
MFCCs	Mel-frequency cepstral coefficients.	40
Chorma-stft	Chromagram representing the 12 distinct chromas.	12
Amplitude	Maximum and mean amplitude of a waveform	2
Spectral centroid	Measures the 'center of mass' of a waveform.	1
Statistical parameters	Moment, variation, skew, variance, auto-corelation, kurtosis.	6
Total Number of features		61

for music and audio analysis are listed the table-2. To improve the performance of the model, the number of features may vary in the final solution. 25% data will be used for testing the model. Since the amount of data in both classes are not equal, I will also observe if the model performance improves with class weight defined as 'balanced' in scikit-sklearn algorithms.

4 Solution Statement

The next step would make a support vector machine model that outperforms its current scores. I would first find the best parameters using GridSearchCV of scikit-sklearn [3]. I believe the best parameters will improve the model performance. Whatever the case I would also look the learning curve of the model to check if the model requires more training data to outperform.I will add/remove features to see if the model scores improve.

5 Benchmark Model

Table-3 listed five supervised algorithms used to create the benchmark. In these benchmarks, I used sklearn [3] default parameters. It is pretty obvious that the support vector machine (SVM) performs better than all other algorithms. Therefore, I will use SVM as the final model while its scores will be used as the benchmark for further improvement.

Table 3: Model performance in five algorithoms

Algorithom	Accuracy	ROC-AUC	Avg. F1-score	Avg. recall
Random Forest	0.64	0.64	0.64	0.65
Xgboost	0.61	0.60	0.61	0.61
LightGBM	0.48	0.41	0.42	0.48
Gaussian Naive Bayes	0.50	0.63	0.41	0.50
Support Vector machine	0.77	0.76	0.78	0.78

6 Evaluation Metrics

In this project, I will use the following evaluation metrics: (1) Accuracy (2) Recall (3) F-1 Score (4) Receiver Operating Characteristic Curve (ROC AUC) and (4) confusion matrix. Other than recall and F1-score, the confusion matrix is interesting since it provides information on the performance of the model on a set of test data for which the actual values are known. The results will be evaluated based on the predicted and the actual values of the testing dataset.

7 Project Design

First, I will find the best SVM parameters which will improve the model performance. Then I would try to do some feature engineering to find features that are responsible for the better accuracy. I will also look into learning curve to check if the model needs more training data to outperform the current scores. I would iterate the above steps until finding the best model.

References

[1] Longjun Dong, Johan Wesseloo, Yves Potvin, and Xibing Li. Discrimination of mine seismic events and blasts using the fisher classifier, naive bayesian classifier and logistic regression. *Rock Mechanics and Rock Engineering*, 49(1):183–211, 2016.

[2] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pages 18–25, 2015.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[4] The Telegraph. Earthquake in north korea was natural, south korean agency says. *The Telegraph*, 2017.