FALSE ALARM DETECTION

ABSTRACT

The main purpose of this paper is to detect false alarms whenever there is a sign of any seismic activity; i.e, to classify a seismic activity into a real or false earthquake, and also find out which attributes contribute significantly to this classification. This paper uses concepts of machine learning to classify the seismic activity to the maximum exactness possible. C4.5 and Random Forest algorithms are used to make the classification. Using these models, a f-score of 0.994 and an accuracy of 99.7 has been achieved, while random forest has achieved an f-score of 0.93 approximately. The results of this paper will help reduce unnecessary apprehensions in people and further not cause any delay in work, also it will guide seismologists to make better decisions as the results suggest features that are highly correlated with the trueness of the calamity.

METHODOLOGY

Data Processing

The data for the research had been taken from the USGS website. The data is metadata: seismic waves and other relevant information has been converted into a table in which columns contain the parameters and rows contain the samples. Some of the features include region of occurrence, duration of the activity, intensity, magnitude, error in reading among many other relevant features.

The first dataset on which analysis was performed contained 4870 samples. This was split into training and testing where the train to test size ratio was 9:1. The second dataset comprised data from 2015-2019 that contained eleven-thousand samples where the test train size is 8:2.

Algorithms

C4.5 has been used to predict the class whether it is a false or a real earthquake. First the data of 4870 samples were taken and the train to test ratio was 9:1. Accordingly a C4.5 (A decision tree) was built based on the training data. The constructed tree was used on the test set to and the classes for each sample were found out. The f-score measure was 0.997.

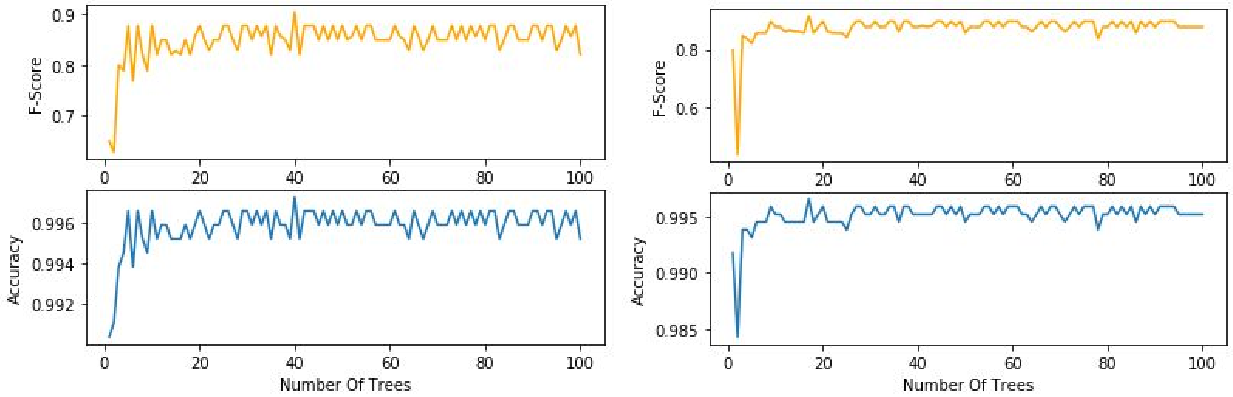
The 2nd data set contained 11,000 samples approximately which comprised all the seismic activity details of California region from 2015-2020 samples and the same process was repeated along with Random Forest except the train to test size ratio was kept to 8:2, here also the f-measure came out to be 0.996 for C4.5 while Random Forest had a f-score value of 0.93.

Feature Correlation

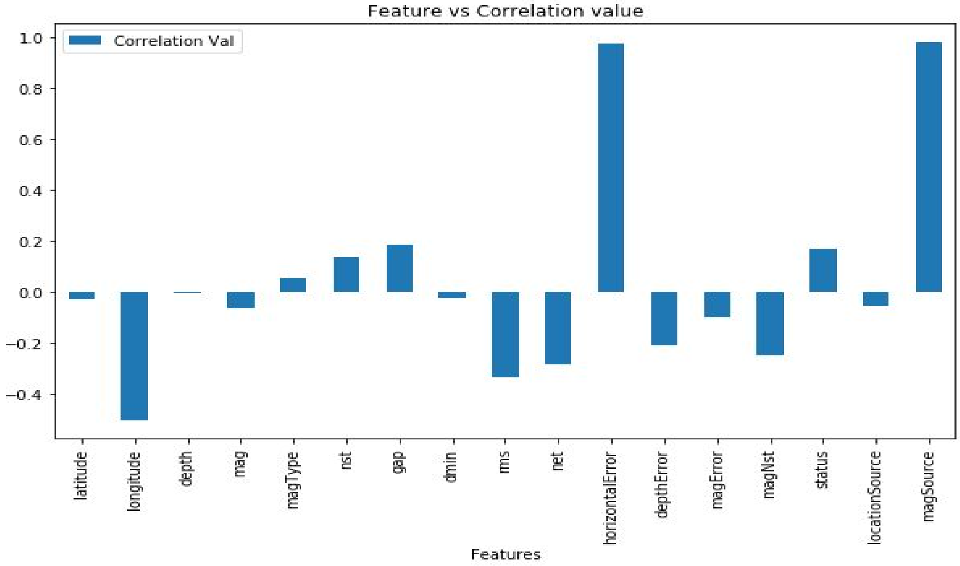
After the classification, np.corr() from python's numpy library was used to get the correlation of each feature to the type of seismic activity, the bar-plot of it is present in the result section.

RESULTS

The plots depict the accuracy and f-score measure of Random Forests and SVM Classifier with increase in the number of trees and folds respectively. We can observe from the plot that after a certain number of trees/folds the values tend to saturate, so it is not further useful to change them. Another result that was observed was the point at which the start to saturate is proportional to the data size.



The correlation plot depicts the features that are significant to the output class: type of earthquake. This may help seismologists discern a seismic activity and take measures accordingly.



CONCLUSION

To put it all together, a comprehensive study was made on the earthquake data to precisely classify a seismic activity into a real or a false earthquake, using the concepts of machine learning. An f-measure above 0.993 was achieved when using C4.5 and above 0.9 when using random forest. The analysis made can help seismologists to send a right alert message to people instead of false alarms. This analysis also discusses the correlation of features to the type of seismic activity helping seismologists to make decisions faster, instead going through the whole process. Finally, and more importantly it can be advantageous to people in California, Italy and other regions that show similar seismic behavior and thus can save lives and property.