Analysis of Seismic data using Machine Learning Algorithms

A A V L Sruthi*, Bhargavi R, Vineesha Reddy Gospati

School of Civil Engineering, SASTRA Deemed University, Thanjavur – 613401. India.

*Corresponding Author: aavlsruthi@gmail.com

Abstract: Earthquakes result in a gigantic loss of lives and properties to people because of its powerful, devastating and deep action. Over the years, a lot of research is going on to forecast the likelihood of occurrence of an earthquake to minimize the loss. In this study, a data mining technique i.e., classification analysis has been applied to estimate the most accurate earthquake model. Previous seismic data were collected and classified by applying k-NN (k-nearest neighbors algorithm) and Random forest algorithms. k-NN is a supervised machine learning algorithm used for bigger datasets (generally for statistical estimation) to determine the accuracy of the model. Random forest algorithm is also a supervised algorithm which is used for both classification and regression. By using this algorithm, multiple decision trees can be created over the datasets as well as predicting and offering a solution. Analysis and visualization of the data has been done and subsequently a comparative analysis of these two algorithms were done and tested to obtain the efficiency in predicting the accuracy of the earthquake model in terms of earthquake magnitude and depth.

Keywords: Seismic analysis; machine learning algorithm; Random forest algorithm; k-NN algorithm; earthquake magnitude; focal depth.

1. Introduction

Due to its devastating and sudden action, earthquake has always resulted in a huge loss of lives and properties to people and great destruction to our environment i.e. it prompts auxiliary calamities, for example, avalanches, flames and tidal waves. In recent times, many minor and major earthquakes occurred on a different scale. In 2004, an earthquake measuring 9.1 magnitude occurred in the Indian Ocean off the coast of Sumatra, causing a sequence of tsunamis killed 227,899 people across 14 nations in the areas with Indonesia being the worst-hit, subsequently India, Sri Lanka and Thailand [1]. Over 8,000 natives were exterminated on April 25, 2015, following an earthquake of magnitude 7.8 that strikes Nepal [2]. In 2017, off the southern coast of Mexico, an earthquake of magnitude 8.1 affecting 1.5 million people by destructing 41,000 homes [3]. Over the period of time with the huge evolution of the technological improvements, a new method benefits the researchers to analyze the seismic data. Currently there are numerous strategies utilized in seismic data analysis resulting in reducing the risk of earthquakes.

Fana et al. [4] has given a new strategy for breaking-down seismic information which can uncover the circulation of quakes, which encourage the researchers to learn the laws of tremor calamities and diminish the dangers fetched by them. A comparative analysis between density-based spatial clustering of applications with noise (DBSCAN) and K-means clustering algorithms has been done to analyse the performance of fitting seismic belts with seismic datasets. The experimental outcomes show that the DBSCAN calculation has the best influence in fitting; furthermore, the results of K-means algorithms for Earthquakes are additionally in accordance with prospect. Further research on time progression can be acquainted into the study to explore the investigation of seismic sharing.
Jalal Kiani et al. [5] investigated various types of classification algorithms to check the sensitivity and size of the employed dataset. Khawaja et al. [6] observed a novel pattern and varieties in execution of seismicity forecast and seismic markers were registered numerically utilizing past seismicity. Seismic activity prediction and modelling of seismic indicators was done using machine learning where three AI methods (Artificial neural networks, Random forest and Support vector machine) were applied to the Cyprus locale with the end goal of Earthquake expectation. The forecasts acquired from the Random Forest method were discovered to be the most precise for earthquake magnitude values of 3.0 and 3.5 above all the forecast periods. Also, the prognoses obtained from the Support Vector Machine exceeded for magnitude values of 4.0 and 4.5. Furthermore, predictions for earthquake magnitude thresholds of greater than the current values (> 4.5) for different topography can be performed.

Antonella et al. [7] aim was to obtain innovative experiences on neighborhood scale examples of seismicity in the zone. Seismic progressions were distinguished by the nearest neighbor technique. This technique allows to feature and explore the inward composition of earthquake arrangements, and to separate the spatial features of seismic activity as per the diverse topological highlights of the clusters arrangement. Clusters of various types produce diverse ideal geographic sites. Rémi et al. [8] observed SVM (support vector machine) can be used to determine the vulnerability curves where pre-processing of the input data is executed, resulting in seismic intensity. Grace et al. [9] explored the application of supervised machine learning categorization methods pertaining to time series resulting from seismic statistics. This examination was to apply AI methods to single station seismic data to group by and large volcanic state as eruptive or non-eruptive. The results indicated that eruptive classification gave high confirmation with visual indicators of eruption, namely emission of ash. Further study can be done by considering more attributes to check the current attributes consistency. The aim of this study is to apply and understand the distinctive classification algorithms in the investigation of seismic data in terms of earthquake magnitude. Since it is a large dataset, k-NN and random forest algorithms were used to analyse the data. The results of the k-NN algorithm and the Random forest algorithm were compared to get better efficiency in obtaining the accurate Earthquake model.

2. Methodology

The dataset employed in this study is the global seismic dataset taken from the website ‘kaggle’. This dataset includes 23,413 documentation, and every evidence incorporates 21 features namely Date, Time, Latitude, Longitude, Type, Depth, Depth Error, Depth Seismic Stations, Magnitude, Magnitude Type, Magnitude Error, Magnitude Seismic stations, Azimuthal Gap, Horizontal Distance, Horizontal Error, Root Mean Square, ID, Source, Location Source, Magnitude Source, Status.

Of these, numerous features are not associated with the content of these two algorithms. So, the data has been cleaned according to the requirements of algorithms. In the k-NN algorithm, the output of the model is basically a categorical value. Now, the output attribute which is magnitude is converted into a categorical values (i.e; Magnitude discrete) namely ‘Medium’ that ranges from 0 to 5.8 and Big that ranges from 5.8 to 10. After cleaning the unnecessary data, Input dataset is [Latitude, Longitude, Depth, Magnitude Type, Root Mean Square, Source, Location Source, Magnitude Source, date_parsed]. In the Random forest algorithm, it is not necessary to convert output data into any categorical values. Therefore, Output dataset is [Depth, Magnitude] and Input dataset is [Latitude, Longitude, Magnitude Type, Source, Location Source, Magnitude Source, date_parsed]. Firstly, to analyse any data, data visualization needs to be performed. Data after cleaning is visualized and the following are the results of visualization.
The occurrence of earthquake from 1965 to 2016 is shown in figure 1, which gives the relation between the location and magnitude. From the graph, it can be observed that the magnitudes of earthquake in most of the regions ranges from 4.5 to 6 and the occurrence of magnitude 8 to 9.5 is very low. This graph can be used to identify the seismic belt. The number of earthquakes in each year is shown in figure 2. Here, the data is grouped by year and the number of earthquakes in each year is calculated. From figure 2, it can be seen that the alignment of graph is changing continuously; number of earthquakes is optimum in the period 2005 to 2015 and minimum in 1965. The average magnitude of each year is shown in figure 3. Here the data is first grouped by year and the mean of magnitude is computed. Though the number of earthquakes is very low, the average magnitude of earthquakes is very high. The number of earthquakes with same magnitude in the data is shown in figure 4, which shows that the occurrence of earthquake magnitude above 8 is almost negligible whereas the number of earthquakes of magnitude 5.5 to 6 is very high.

3. Analysis:
3.1. k-NN algorithm:
The k-nearest neighbors (KNN) algorithm is one of the supervised machine learning algorithms and it is used for larger files. Both classification and regression problems can be solved using this algorithm.
In this study, k-nearest neighbors or KNN Algorithm is applied on classification problem to obtain the accuracy of the model. This algorithm will find the similar features of the testing dataset to the Big and Medium magnitude discretes and based on the most similar features it will put it in either Big or Medium category.

After analyzing dataset, the dependent and independent variables are extracted. By importing train, test, split function from scikit learn library, Inputs and output are divided into training and testing datasets. The number of samples taken in the training dataset are 15563, the testing datasets are 7666 and their proportions are 70% and 30% respectively. Further, the creation of model is done by importing the KNeighborsClassifier from class of Sklearn Neighbors library. After importing the class, the Classifier object of the class is created. The created classifier object is used for fitting the training dataset. Model is created and it can be tested for the accuracy. This can be done with the same classifier object using predict function. After prediction, the accuracy score of testing and training dataset is calculated and the accuracy scores of testing and training datasets are 54.89%, 71.03% respectively. The detailed output of this model can be shown from the classification report (Table 1) and confusion matrix (Table 2).

### Table 1. Classification Report

|                | Precision (a) | Recall (b) | F1-score (c) | Support (d) |
|----------------|---------------|------------|--------------|-------------|
| Big (> 5.8)   | 0.33          | 0.41       | 0.37         | 2438        |
| Medium (< 5.8)| 0.69          | 0.61       | 0.65         | 5228        |
| Accuracy      | 0.55          | 0.55       | 0.55         | 7666        |

### Table 2. Confusion Matrix

|        | Big | Medium |
|--------|-----|--------|
| Big    | 1010| 1428   |
| Medium | 2030| 3198   |

From the classification report, the results of test dataset are shown i.e. 2438 values are predicted to be Big (> 5.8) and 5228 values are predicted to be medium (< 5.8). Comparing the tested values and predicted values from the model it is shown that from confusion matrix only 1010 (33%) values gave the exact result (True Positives) for Big, 3198 (69%) values gave the exact result (True positives) for Medium and the remaining values are predicted to be wrong. From the above results it can be inferred that the model doesn’t fit exactly.

3.2. Analysis & Results of Random forest:

It is one of the Supervised Machine Learning algorithms which use ensemble learning method for classification and regression. It is a bagging technique, used to reduce the variance for those algorithms that have high variance, typically decision trees. After analyzing the dataset, the dependent and independent variables are extracted. By importing train, test, split function from scikit learn library, inputs and output were divided into training and testing datasets. The number of samples taken in the training dataset was 18583, the testing datasets are 4646 and their proportions are 80% and 20% respectively. Further, the creation of model has been done by importing the Random Forest Regressor from class of sklearn ensemble library. After importing the class, the object of the Random forest was created. The created object is then used for fitting the training dataset. Model was created and it can be tested for the accuracy. This can be done with the same object using predict function. After prediction, the accuracy score of testing and training dataset is calculated and the accuracy scores of testing and training datasets are 99.999979081535 %, 99.9999949224969 % respectively.
Since the output data is numerical, it is not appropriate to get the comparison between tested and predicted values. By plotting a scatter plot between tested and predicted values from this model, comparison of values can be done as shown in figure 5. Since the scatter plot obtained is a straight line, it can be inferred that the accuracy of the model is high.

4. Conclusion:
From the collected dataset, the most accurate Earthquake Model in terms of earthquake magnitude is predicted. In this study, two distinctive machine learning algorithms i.e. k-NN and Random forest were employed in order to analyse the seismic data. The comparison was made between the two algorithms to analyze and visualize the data. The results infer that the random forest algorithm gives most accurate prediction when compared to the k-NN algorithm. Earthquakes can occur both on a smaller and larger scale in a very short span creating an unexpected loss to humanity and the environment. Thus it is essential to study, analyze and predict the possibility of occurrence of an earthquake to minimize the loss. As a result, it is convenient to employ classification algorithms in the analysis of seismic data.

5. References:
[1] "Magnitude 9.1 – Off the West Coast of Sumatra". U.S. Geological Survey. Archived from the original on 17 August 2012. Retrieved 26 August 2012.
[2] Chidanand Rajghatta (26 April 2015). "Is this the 'Big Himalayan Quake' we feared?". The Times of India. Retrieved 26 April 2015.
[3] "Death toll rises to 60 in powerful Mexico earthquake". Associated Press. 9 September 2017. Retrieved 9 September 2017.
[4] Fan Z, Xu X 2019 Application and visualization of typical clustering algorithms in seismic data analysis, Procedia Comput. Sci. 151 171.
[5] Jalal Kiani, Charles Camp, Shahram Pezeshk 2019 On the application of machine learning techniques to derive seismic fragility curves, Comput. Struct. 218 108.
[6] Khawaja M Asim, Sayed SR Moustafa, Iftikhar Azim Niaz, Eslam A Elawadi, Talat Iqbal, Francisco Martínez-Álvarez 2020 Seismicity analysis and machine learning models for short-term low magnitude seismic activity predictions in Cyprus, Soil Dyn. Earthq. Eng. 130 105932.
[7] Antonella Peresan, Stefania Gentili 2018 Seismic clusters analysis in Northeastern Italy by the nearest-neighbor approach, *Phys. Earth Planet. In.* 274 87.

[8] Rémi Sainet, Cyril Feau, Jean-Marc Martinez, Josselin Garnier 2020 Efficient methodology for seismic fragility curves estimation by active learning on support vector machines, *Struct. Saf.* 86 101972.

[9] Grace F Manley, David M Pyle, Tamsin A Mather, Mel Rodgers, David A Clifton, Benjamin G Stokell, Glenn Thompson, John Makario Londoño, Diana C Roman 2020 Understanding the timing of eruption end using a machine learning approach to classification of seismic time series, *J. Volcanol. Geoterm.* 401 106917.