ISSN: 0975-766X
CODEN: IJPTFI
Research Article

Available Online through
www.ijptonline.com

AN APPROACH TO PREDICT EARHQUAKE DAMAGE USING REGRESSION TECHNIQUE

Ankit Panchariya, J.Jabanjalin Hilda, C.Srimathi, Anix Mary Javitha[1,2,3,4]
[1,2,3] School of Computer Science and Engineeering,VIT University,Vellore
[4]Department of Computer Science , St. Joseph University in Tanzania, Tanzania
Emai: jabanjalin.hilda@vit.ac.in

Received on: 08-02-2017 *Accepted on: 21-03-2017*

Abstract:

Background and Objectives: An earthquake is the sensible shuttering or trembling of the surface of the Earth, originating due the sudden release of energy in the Earth's crust resulting in generated seismic shocks. The quake can be savage enough to flip people around and damage whole cities. The tectonic or seismic activity of a zone refers to the particular frequency or wavelength, type and size of shock felt over a period of time duration. **Materials and Methods:** Measurement of intensity is done by seismometer as referred to Richter magnitude scale. The various Causes are landslides, fires, avalanches, soil liquefaction, Tsunami and mostly human impact. The aim of the effort to predict the damage of human lives caused by earthquake with given user inputs and implementing machine learning specifically regression models using Orange tool. **Result:** The data which is used in history of earthquakes and their respective damages in particular area is based on the parameters like depth, duration, longitude, latitude etc. Various methods like linear regression, Mean regression, Random forest regression are compared to predict the earthquake.

Keywords: Amplitude, Earthquake Prediction, Linearregression, Random Forest

1. Introduction:

The earthquake or seismic waves are generated based on the following condition:a. Strike Slip b. Normal c. Thrust[1]. The various types of earthquakes are

1. Earthquake away from boundaries (in continental lithosphere)
2. Earthquake due to volcanic activities. (Due to movement of magma in volcanos)
3. Tidal forces (Due to gravitation pool of moon and earth)
4. Earthquake Clusters (Group of plates keep resisting against each other)

Body-wave magnitude expression is expressed[2] as follows:

$m_b = \log_{10}(A/T) + Q(D,h)$ *A* - Amplitude of wave motion (in units of microns), *T* is the referring period (in units of seconds), Q (*D*,*h*) is a correction factor that as a function of distance, *D* (degrees), between epicenter and station, focal depth, *h* (in kilometers), of the earthquake. The standard surface-wave formula is

$M_S = \log_{10} (A/T) + 1.66 \log_{10} (D) + 3.30.$

Energy[3] is expressed as $\log_{10}E = 11.8 + 1.5M_S$, *E* is measured in ergs. M_S is calculated from a bandwidth between approximately 18 to 22 s. It is now known that the energy radiated by seismic wave is concentrated at different bandwidth with higher frequencies.

$M_e = 2/3 \log_{10}E - 2.9.$

For every increment in magnitude by the unit of one, the associated seismic energy gets multiplies by about 32 times. The phases which will be carried out are Data collection, Data Preprocessing, Data Mining, Data Visualization, Result and Analysis

Figure 1: Block diagram of Earth Quake prediction process.

2. Related Work:

Deterministic Approach suggested [4] selects a source with predefined magnitude and distance to measure its actual location and then predicts based on population density.

Second approach is in suggested [5] is Probabilistic which determines frequency of earthquake based on plate acceleration and threshold of intensity assuming they can act independently.

Third Approach[6] is Hybrid of both which will include first two and PHSA is used for calculating hazard ,a source with such magnitude-distance pair charters tics is selected as controlling event which uses probability and then deterministic analysis is performed.

3. Methodology: The first step is collection of data and all analysis and implementation is performed in Orange. [3]In preprocessing the operations which are performed are

- Entropy MDL discretization
- Equal frequency discretization
- Equal width discretization

The distance which is used is Euclidian with k=1 for nearest neighbors. And the numbers of outliers are 4 they are excluded from analysis. The second operation which is performed is K-means clustering.

First assign a range of no of clusters will be used. Then calculate Euclidian distance on attributes considering as coordinates. The attributes are YEAR, FOCAL_DEPTH, EQ_PRIMARY, LOCATION_NAME, LATITUDE, LONGITUDE, DEATHS, DEATHS_DESCRIPTION, INJURIES, DAMAGE_DESCRIPTION, HOUSES_DESTROYED, TOTAL_DEATHS, TOTAL_INJURIES

And final clusters based output is predicted values of death.

The regression algorithms used are

- Linear Regression
- Mean Regression
- Random Forest Regression.

Linear Regression is a method for finding relationship between scalar dependent variables x and y. Mean Regression is an approach is used if a variable shows extreme on its first measurement but the closer behavior tends to lean towards second measurement. And again it tends towards first finally ends toward mean[7].

Random Forest Regressionuses ensemble learning method for classification and regression using various types of decisiontrees and uses corrects the habit of overfitting in decision trees. It can be random subspace or bagging even etc[8].

4. Analysis: The linear regression is performed with two variations:

4.1 Linear Regression

DEATHS_DESCRIPTION	INJURIES	DAMAGE_DESCRIPTION	HOUSES_DESTROYED	TOTAL_DEATHS	TOTAL_INJURIES	Linear Regression
4	0	4	0	19000	0	-0.095
1	0	4	0	0	0	1.111
0	0	0	0	0	0	0.553
4	0	3	0	1530	0	2.035
0	0	0	0	0	0	-1.151
3	0	3	0	0	0	1.720
0	0	0	0	0	0	1.453
2	0	2	0	0	0	0.412
1	0	1	0	0	0	1.944
3	0	1	0	0	0	-1.036
1	200	2	0	26	200	201.065
1	0	2	0	47	0	1.502
1	40	2	0	15	40	41.645
1	0	1	0	0	0	1.942
0	0	2	0	0	0	-1.371
1	100	3	0	20	100	104.145
1	30	3	0	6	60	35.246
4	1800	4	7500	2000	1800	1799.001
0	0	1	0	0	0	0.587
4	30000	4	0	11000	30000	30000.047
0	25	1	0	0	25	24.860
2	1000	4	8546	56	1000	1000.053
1	200	1	1	23	200	200.872
2	394	4	21100	100	394	393.696
0	0	1	0	0	0	-0.366
4	166836	4	339000	20005	166836	166836.000
1	0	1	40	2	0	-0.045
0	0	1	0	0	0	-0.331
0	0	1	0	0	0	-0.322
0	45	2	0	0	45	44.641
0	0	1	0	0	0	-0.127
0	0	1	0	0	0	-1.402
0	0	2	0	0	0	0.020

Figure 2: Outlier removed for predicting injuries.

The results are the heighted values and are predicted for deaths more deviated from real values.

4.2 Lasso Linear Regression(Less efficient)

This technique is moredeviated (more negative values) from actual results.

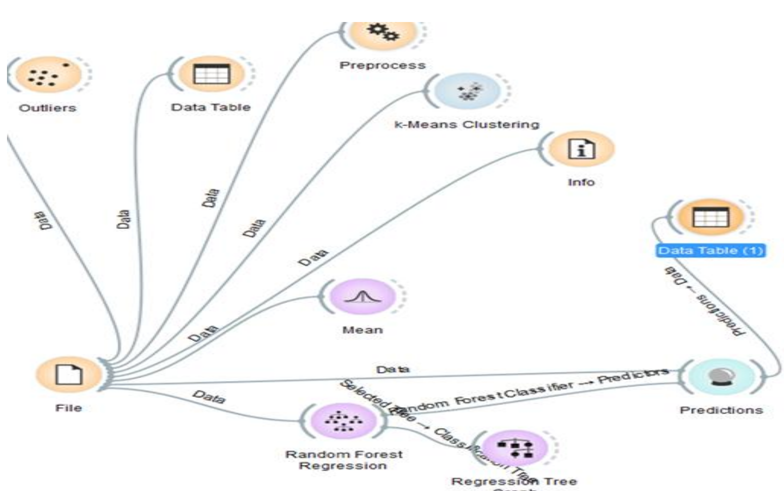
INJURIES	DAMAGE_DESCRIPTION	HOUSES_DESTROYED	TOTAL_DEATHS	TOTAL_INJURIES	Linear Regression
0	4	0	19000	0	-0.095
0	4	0	0	0	1.111
0	0	0	0	0	0.553
0	3	0	1530	0	2.035
0	0	0	0	0	-1.151
0	3	0	0	0	1.720
0	0	0	0	0	1.453
0	2	0	0	0	0.412
0	1	0	0	0	1.944
0	1	0	0	0	-1.836
200	2	0	26	200	201.065
0	2	0	47	0	1.502
40	2	0	15	40	41.645
0	1	0	0	0	1.942
0	2	0	0	0	-1.371
100	3	0	20	100	104.145
30	3	0	6	60	35.246
1800	4	7500	2000	1800	1799.001
0	1	0	0	0	0.587
30000	4	0	11000	30000	30000.047
25	1	0	0	25	24.880
1000	4	8546	56	1000	1000.053
200	1	1	23	200	200.872
394	4	21100	100	394	393.696

Figure 3: Lasso regression for predicting injuries.

4.3 Mean Regression

FOCAL_DEPTH	EQ_PRIMARY	LATITUDE	LONGITUDE	DEATHS	DEATHS_DESCRIPTION	INJURIES	DAMAGE_DESCRIPTION	HOUSES_DESTROYED	TOTAL_INJURIES	TOTAL_DEATHS	Mean
25	7.8	33.000	76.000	19000	4	0	4	0	0	19000	967.985
55	7.6	12.500	92.500	0	1	0	4	0	0	0	967.985
60	7.9	28.500	94.000	0	0	0	0	0	0	0	967.985
33	8.6	28.500	96.500	1530	4	0	3	0	0	1530	967.985
0	7.3	6.500	94.000	0	0	0	0	0	0	0	967.985
0	6.1	23.000	70.000	156	3	0	3	0	0	0	967.985
141	7.6	24.900	93.340	0	0	0	0	0	0	0	967.985
25	5.3	33.900	74.800	80	2	0	2	0	0	0	967.985
53	5.6	28.700	78.900	15	1	0	1	0	0	0	967.985

Figure 4: Mean Regression for predicting injuries.



4.4 Random forest based Regression

Table 1: Random forest based Regression.

Total_Deaths	Values obtained through Rand Forest
0	363.378
0	577.579
0	129.589
0	82.636
1	82.416
0	94.463
1	130.821
0	243.245
0	86.599

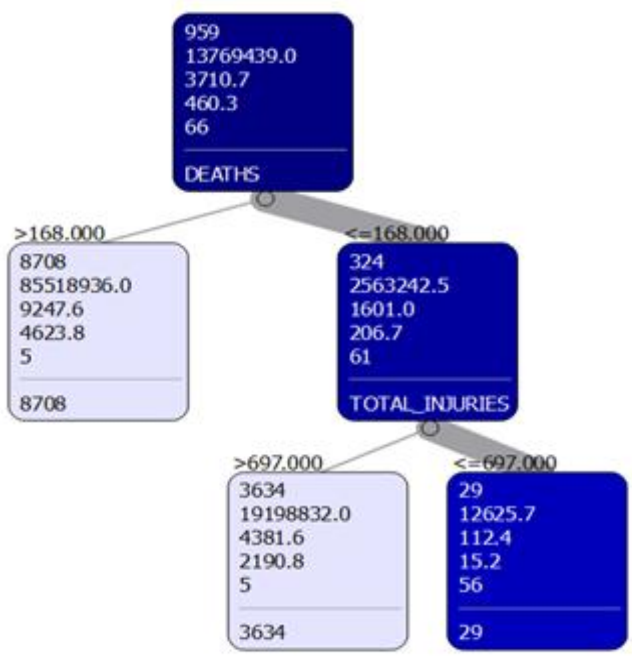


Figure 6: Tree Structure.

Table 2: Probability of death occurrence

Total no.of death	Total instances	Probability
24	67	0.358

Table 3:Comparison of death predicted values with the various regression techniques.

Actual Values	Random forest based Regression	Linear Regression	Lasso (lambda=0.10, tolerance=0.0000010) Based Regression
19000	6325.492	-0.095	18724.9
0	1888.274	1.11	-82.581
0	2842.529	0.553	189.036
1530	1991.254	2.03	1674
0	530.925	-1.151	56
0	1329.107	1.72	46

Table 4: Probability obtained from various regression techniques.

Random forest	Linear Regression	Lasso Model
0.328	0.298	0.283

Although the maximum nearest value is predicted by lasso(But only in minor cases) based regression which may have some negative values but all predicted values are more near to actual values.Here accuracy is calculated after the actual earthquake happened not overall .And all negative predicted values are avoided because damage prediction must be a considerable amount.

5. Conclusion

Various methods like linear regression, Mean regression, Random forest regression are compared to predict the earthquake.The highest accuracy achieved is using random forest and is based on DST (Decision trees) accuracy besidesit is the best type for measuring the intensity of earthquake .Out of the 30 clusters, the parameter that influence much is based on location (longitude and latitude) and focal depth. Some of the entries are zero because the region is either northeast or Andaman Island. The other factors that influences greatly are population density, soil strength, and Development index.

6. References

1. Trifunac MD. Threshold magnitudes which cause the ground motion exceeding the values expected during the next 50 years in a metropolitan area. Geofizika. 1989 Dec 1;6(1):1-2.
2. Trifunac MD. A microzonation method based on uniform risk spectra. Soil Dynamics and Earthquake Engineering. 1990 Jan 1;9(1):34-43.

3. Trifunac MD. A note on the differences in magnitudes estimated from strong motion data and from Wood. InAnderson Seismometer, Soil Dynam. Earthq. Engng 1991.
4. Trifunac MD. A note on the differences in magnitudes estimated from strong motion data and from Wood-Anderson seismometer. Soil Dynamics and Earthquake Engineering. 1991 Nov 30;10(8):423-8.
5. Bouchon M. The complete synthesis of seismic crustal phases at regional distances. Journal of Geophysical Research: Solid Earth. 1982 Mar 10;87(B3):1735-41.
6. Nehybka V, Tilsarova R. Seismic activity in West Bohemia from 2001-2006. ACTA GEODYNAMICA ET GEOMATERIALIA. 2007 Jan 1;4(4):51.
7. Trifunac MD, Herak D. Relationship of MLSM and magnitudes determined by regional seismological stations in South-Eastern and Central Europe. Soil Dynamics and Earthquake Engineering. 1992 Jan 1;11(4):229-41.
8. Trifunac MD. Preliminary analysis of the peaks of strong earthquake ground motion—dependence of peaks on earthquake magnitude, epicentral distance, and recording site conditions. Bulletin of the Seismological Society of America. 1976 Feb 1;66(1):189-219.