(IJIEST) 2020, Vol. No. 6, Jan-Dec **e-ISSN: 2454-9584; p-ISSN: 2454-8111**

DEVELOPING AN INTEGRATED MODEL BASED ON THE PROBABILISTIC NEURAL NETWORK (PNN) IN THE EARLY DETECTION AND PREDICTION OF AN IMPENDING EARTHQUAKE

Romharsh Mittal

ABSTRACT

A wide variety of tragedies happen over the globe; the forecast of such catastrophe is imperative for early safeguard and clearing process. The expectation of a quake could be accomplished used antecedents or seismographic information; however, all such strategies can be completed distinctly by the area specialists (seismologist). Information mining strategies have been utilized in a wide assortment of utilizations and different areas. It permitted the forecast of execution and expected movement, which empowers the inference of wise choices. Forecasting earthquake history information can be accomplished by utilizing information mining ideas. In this paper, an expectation model is proposed for envisioning seismic tremors by applying clustering and affiliation rule mining on quake history information. At first, the data is gathered, and they are bunched, this grouped information is passed to the next stage where successive examples are acquired by applying affiliation rule mining, at long last by utilizing the case, the future quakes are anticipated by performing rule coordinating. This paper focuses on a prescient model using noteworthy quake information and mining strategies, which predicts the fore coming earthquake. This expectation model can be utilized to anticipate different seismic occasions, and they can be used for making a forecast in various fields by using a fitting dataset.

1. INTRODUCTION

Data mining involves analyzing and uncovering new patterns and correlations from a massive amount of data by making use of mathematical and statistical techniques. It can also extract knowledge from noisy, incomplete, and uncertain data. It makes use of technologies like visualization techniques, machine learning, and techniques employing statistics. One of the primary methods of data mining is the prediction that utilizes techniques like regression and obtains pattern from the available data. The following sequence of steps are involved in any data mining process: data management, which includes acquisition of data, transformation, and integration of data, storage, and management of data; data pre-processing, involves removal of noises, filling missing data, eliminating inconsistent data; data mining tasks and algorithm includes the technique employed for the process, and finally, post-processing of data involves refining and also evaluation of results obtained from the previous step. Data mining is incorporated for data like financial data, bioinformatics, meteorological data, retail data, scientific data, social media data. This paper suggests a predictive model using data mining techniques. A wide variety of disasters occur across the globe; the prediction of such failure is the requisite for early precaution and evacuation process. On the contrary, one could observe some increase in the same. Thus, it becomes an urgent necessity in designing a proactive system to ward off the calamities. One such disaster, namely the occurrence of earthquakes, has been brought into focus here. The bumping of plates present in the earth's crust leads to earthquakes; this results as the energy is emitted as seismic waves due to the collision of tectonic plates. Disaster management

requires a mechanism to anticipate the happening of crisis, and these mechanisms help in reducing demolition during disasters by foretelling the extent, location, severity, and ramification of the disaster. A wide variety of studies have shown anticipation of an earthquake using geographical information system, precursors like foreshocks, plate movements, vibration in earth's crust, temperature change, radon gas in groundwater, and seismic activities. The classical model uses precursors to predict an earthquake, which requires the domain expert knowledge to carry out prediction. It is also possible to predict an earthquake from the earlier earthquake data using predictive data mining models. With the advent of data mining techniques, it is possible to predict an earthquake from previous data. Several data mining prediction algorithms are available by applying those algorithms or by making use of two or more mining techniques prediction can be accomplished. Basic approaches for prediction are K-Nearest Neighbour, Naïve Bayes, Bayesian Network, and Regression. This work emphasizes on developing a model for predicting earthquake using mining methodology. The paper is organized in the following manner. Section 2 summarizes the literature survey. Section 3 provides detail about the dataset and acquisition method. Section 4 briefs the methods employed in this work. Section 5 elaborates on the design of the proposed system. Section 6 includes the results and discussion. Finally, Section 7 provides the conclusion. The conventional earthquake prediction model employs seismic data, sensor data, and satellite data, which can only be understood by the expert; thus, the prediction can be carried out only by a seismologist. Many studies are available that support the pertinence of data mining techniques for earthquake prediction. A system was developed for forecasting seismic events by recognizing patterns from seismic data using Spanish temporal data, which are clustered by K means. This system predicts a medium earthquake in Spain region and provides adequate performance. A Neural Network (NN) based system for estimating the concentration of radon in soil developed, which perceives the relationship between environmental parameters and the concentration of radon. It is used for reckoning the variation of radon linked with environmental parameters, which impose a nonlinear impact on radon concentration. A Decision tree-based model to analyze radon data to forewarn seismic event was developed, and this outperforms other regression models; it predicts seismic activity by employing radon data of Slovenia.

Another prediction model for quakes in the coastal region was designed, which perform spatial data mining, to the geophysical data obtained from remote sensing. A regression neural network model to determine nonlinear soil's liquefaction potential was proposed, which assesses seismic conditions in sites liable to liquefaction, and this model can be dynamically updated, increasing the predictability. A model to assess the failure characteristics of the slopes present in the highway before and post-earthquake is developed by using neural networks. And this can act as a tool to determine the slope failure of the slope before and post-seismic events in the specified region. A Probabilistic Neural Network (PNN) is developed to predict the magnitude of the earthquake, which uses eight seismic indicators, here the prediction is considered as a classification problem, and this model is based on the Parzen windows classifier. In spite of the advantages over the conventional model, the drawback of PNN is that the desired output should be expressed as one of the predefined classes for prediction problems. A case study involving the prediction of the magnitude of seismic event for the next day utilizing time-series magnitude data and prediction of next crucial forthcoming earthquake using Seismic Electric Signals (SES) has concluded that earthquake prediction using neural network yields an accurate result when they are trained well using proper and large data. An inductive rule building model for natural disaster management is developed, which is an adaptable, low-cost decision support machine; this is intended to assist relief works and to provide support for decision-makers through an inference tool. The knowledge base used here holds two levels of knowledge, the first level is the historical data about disasters stored in a matrix, and the second level is a set of inference rules. A model for predicting earthquake's magnitude in the Red Sea is developed by utilizing Artificial Neural Network (ANN), data for Northern area of Red Sea are used to train the feed-forward NN, which forecast earthquakes in the Red Sea with better accuracy, and it is appropriate for extracting nonlinear relation as well. Another prediction

system using ANN was developed, which predicts earthquakes in Chile with small temporal uncertainty. The proposed model is a three-layer feed-forward ANN, which uses the b-value of Gutenberg–Richter law as input, and the output is the highest magnitude of the earthquake observed for the next five days. Omor-Utsu law has also been incorporated in this ANN. An innovative mathematical model based on spatial connection theory was developed, which uses past earthquake data of earthquake-prone zones to predict the location of an earthquake. This work suggests that the accuracy of prediction can be further increased by integrating with other models using triangulation.

Earthquake anticipation can be done even by using social media messages. These messages are used for identifying disaster patterns, they are also used for generating alert messages, and these messages can be sent through the social network, thus acting as an alerting system. The disaster signatures on Twitter can also be used for prediction; for the same, a case study has been done by considering three different events: tornado, bombings, and hurricane. Tweets can be used to predict and notify about disasters using real-time social data. Earthquake reporting system based on the tweets was developed, which detects the incident of earthquake utilizing tweets, and an alert message is sent through email. A support vector machine is used to classify positive and negative tweets. Semantic analysis is used to acquire tweets on the intended event, and after that, a spatiotemporal probabilistic model is employed to procure activities from tweets. Thus various mining techniques are used in different models which have been proposed earlier; with this previous knowledge and knowledge about the pros and cons, a predictive model is proposed in this paper for anticipating seismic events.

2. DATA ACQUISITION

The earthquake data is obtained from the United States Geological Survey (USGS) repository, which provides global seismic data. Table 1 shows the count of significant earthquakes that occurred globally from the year 1990- 2013 with a depth of a minimum of 100. From the TableTable, we could observe that an earthquake of magnitude 6.5-7 occurred in a large number of times during 1990-2013. Likewise, some patterns can be obtained from the data, which could be helpful in the prediction process. It also provides various customizations for output like format, the order in which data should be arranged, and data can be obtained in any of these formats (Map, CSV, KML, GeoJSON, and QuakeML). Figure 1 shows the USGS earthquake data archive. Earthquake data for any region with various constraints can be obtained from this data archive in any of the formats mentioned above. Quake data in a particular range of magnitude, depth can be obtained by applying the corresponding filter and also for a specified period.

Figure 1. USGS website - earthquake data archive.

3. METHODOLOGY

The proposed model employs two mining techniques:

Clustering and Association rule mining, where data will be initially clustered, and then the gathered data will be given as input to the next process. Frequent itemsets are identified for every cluster from which association rules are derived.

3.1 Clustering

Classification is used for pre-labeled data instances, whereas clustering allows grouping of the cases for given unlabelled data. Thus clustering involves organizing objects into classes of similar objects or related objects. Clustering targets low and high interclass similarity. Using applying to the cluster, the data set is divided into subclasses or partitions called groups. Clustering can also be used as a pre-processing step, which provides data compression and outlier detection. K-means algorithm is employed here to cluster the input data. This algorithm works using the root mean square distance from the centroid to the data points by repeatedly assigning the data points. The algorithm works as depicted in the following steps.

Randomly chooses K cluster centers as an initial centroid.

μj= random data point, where j=1,...,k

Table 1. Number of earthquakes with minimum depth 100 from 1990-2013

Year Mag	1990- 1992	1993- 1995	1996- 1998	1999- 2001	2002- 2004	2005- 2007	2008- 2010	2011- 2013
6.5-7.0	22	23	24	15	14	14	17	26
7.1-7.5	8	4	12	8	6	10	6	10
7.6-8.0	-	2	2	1	2	3	2	2
8.1-8.5	-	1	-	-	-	-	-	1
>8.5	-	-	-	-	-	-	-	-

Assign every data point to its nearest centroid.

cj= {i: d(xi , μj) ≤ d(xi , μl), l ≠ j , i=1,...,m}

Set each cluster centre to the mean of all assigned data points of that cluster.

$$\mu_j = \frac{1}{|c_j|} \sum_{i \in c_j} x_i, \forall_j$$

Repeat the above two steps until convergence.

Clustering involves the partitioning of data points Xi where i = 1, …, m to K-clusters, each data points are assigned to a cluster such that it minimizes the distance from data points to the group.

$$\text{Distance measure} = \sqrt{\sum_{i=1}^{m} (x_i - \mu_i)^2}$$

3.2 Association Rule Mining

Association rule mining involves disclosing the relations among variables in a transaction. Association mining works on two steps; the first steps are to identify the frequent itemsets, and the second steps produce the rule from the frequent itemsets obtained in the previous step. Every data structures have their traits which may help in reducing the memory requirement and also minimizing the time complexity. By employing tree data structure, complexity in rule generation can be minimized. Apriori-Total from Partial (TFP) algorithm proposed by it is used for this purpose, which offers a remarkable advantage in terms of execution time and storage.

3.2.1 Frequent Itemsets Generation in general candidate itemsets are generated first, and those itemsets which satisfy the minimum support threshold are considered frequent itemsets. There are three classes of itemsets, namely Frequent, Closed, and Maximal; depending on the data size and the requirement, they will be taken into account for association mining. Frequent itemsets is a set of an item which occur a particular number of times in the transaction. For this purpose, support value is used.

Support = Ratio of Support count of itemsets to total no. of transactions

Support count = Number of times a particular itemset occurs in a transaction

3.2.2 Association Rule Generation

From the frequent itemsets obtained from the previous step, a set of candidate rules are generated, and those rules that satisfy the minimum confidence are the strong association rules. Several rule generation algorithms are available; the algorithm proposed by in is the most standard algorithm. If M→N is the rule obtained, then the confidence is calculated as follows, Confidence = Support count of (MUN) / Support count of (M)

Another measure, namely lift, is also used in association mining, the lift is the ratio of rule's confidence to the support of itemsets in rule's consequent. Lift = Confidence of rule / Support of rule consequent

3.2.3 Algorithm: Apriori-TFP

This algorithm takes binary valued data as input; it makes use of two tree structures, which provides better storage and consumes less time for execution. Initially, a Partial tree (P-tree) is generated, which stores partial counts for itemsets, if necessary pre-processing can be accomplished in this tree structure, which eliminates all duplicate records. Then it produces frequent itemsets from the given dataset, and it is stored in an enumeration tree data structure called Total tree (T-tree). It generates association rules from the various items present in the T-tree. Pseudocode for Apriori TFP algorithm

Step 1: Generation of Partial Tree from input data 1.1. Before the commencement of the Partial tree, creation initializes array structure, where the size of array = No: of attributes and the nodes in the top row have support value initialized to 0.

1.2. Records are added to corresponding nodes in the tree; increment the support value for the addition of each record.

Step 2: Store the data in P- Tree to a table forming Partial tree table The content of the partial tree obtained in the previous step is stored in a table using arrays.

Step 3: Generation of Total support tree from Partial tree using P-Tree TableTable: T tree is obtained from partial TableTable by parsing it: T tree is generated by passing through the partial tree table, for each pass each level of T tree is formulated. Starting from top-level T tree is constructed by scanning P tree in a level-wise manner. The generation of T-tree is more like in an Apriori way.

1. It starts with T tree whose top level is empty.

2. Scan P -tree level-wise by passing through partial tree table commencing from Level=1.

3. On encountering the records in TableTable, the corresponding node's support in T-tree is updated.

4. At the end of each pass, ignore the top-level nodes of P-tree and delete it from TableTable of the partial tree (i.e.) at the end of each pass node that doesn't provide support are pruned.

5. Produce the next level of T –tree.

6. If the encountered node has a branch in T tree, then move along the office and update the encountered node's elements in T tree if they are present in the chapter.

7. Continue this process for the subsequent levels as well.

This process continues until no nodes exist (i.e.) when zero levels are left to be generated. Thus the total support tree gets updated starting from top level and the proceeds to the following standards.

Step 4: Creation of Association rules by T-tree Association rules are produced by processing the Total support tree obtained in the previous step.

4. PROPOSED MODEL

The proposed system employs clustering and association rule mining for anticipating earthquakes from the last earthquake event data. Figure 2 depicts the architecture of the proposed method. Some of the components involved are:

Input Data: Previous earthquake records collected from various repositories.

Data Translator: The value of attributes latitude, longitude, magnitude and depth are generalized into groups like Southern Hemisphere, Northern Hemisphere,

Quadrant 1, 2, 3, 4; low, medium, large size, and wisdom, respectively.

Rule Generator: Frequent itemsets are identified, and then from the obtained frequent itemsets, a rule generation algorithm is applied, which produces association rules based on the frequent itemsets.

User Interface: The input parameters are obtained from the user through this, and it is also used to provide the predicted output result to the user. Various processes involved in the proposed model have been explained in the following paragraphs.

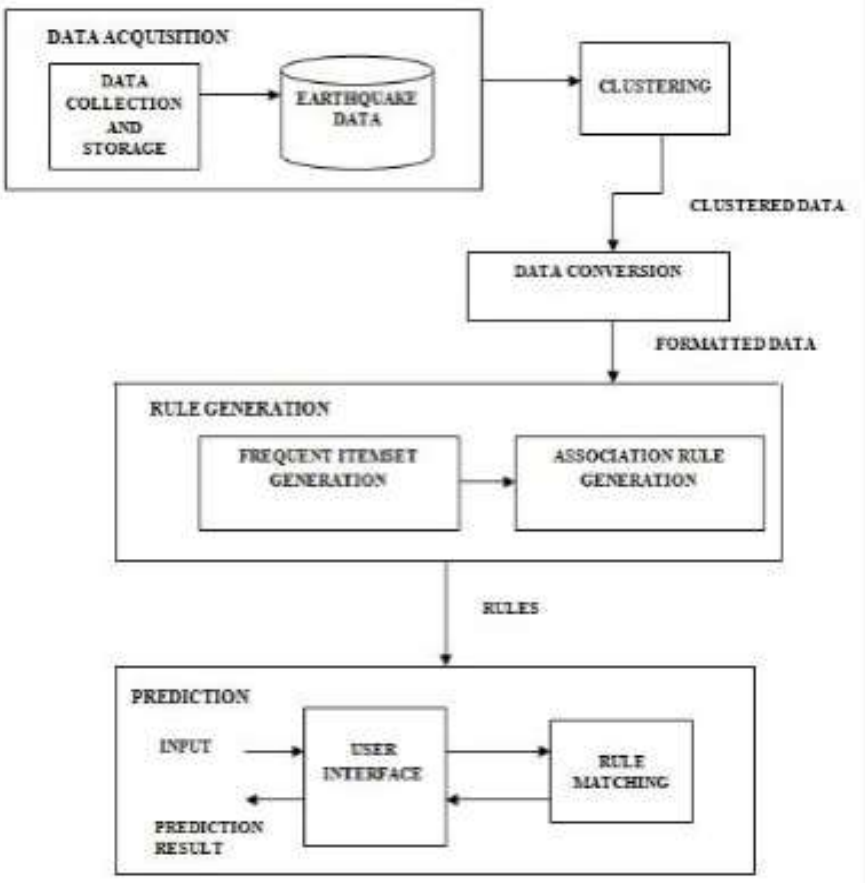

Figure 2. Proposed system architecture.

4.1.1 Data Acquisition and Upload

Initially, the earthquake scientific data are collected from different sources, and they are accumulated. The data obtained from various resources are integrated, and they are loaded into the system so that they can use this data for further processing. For experimental purposes, data from the USGS catalog is used. This data includes earthquake attributes like location (latitude and longitude), time, magnitude, depth, earthquake type, status, and place.

4.1.2 Clustering the Data and Removing Outliers

Here, the collected data will be given as input to the clustering process, where the data will get clustered. Clustering is done by using the K means algorithm. Among various features, the only magnitude is considered for performing clustering. Thus data is grouped based on the size, and from the resulting clusters, the one with minimum extent is viewed as an outlier, and they are removed. After clustering, the clustered data is translated into 1's and 0's by de-normalizing the attributes into a group of values, which is depicted in Figure 3.

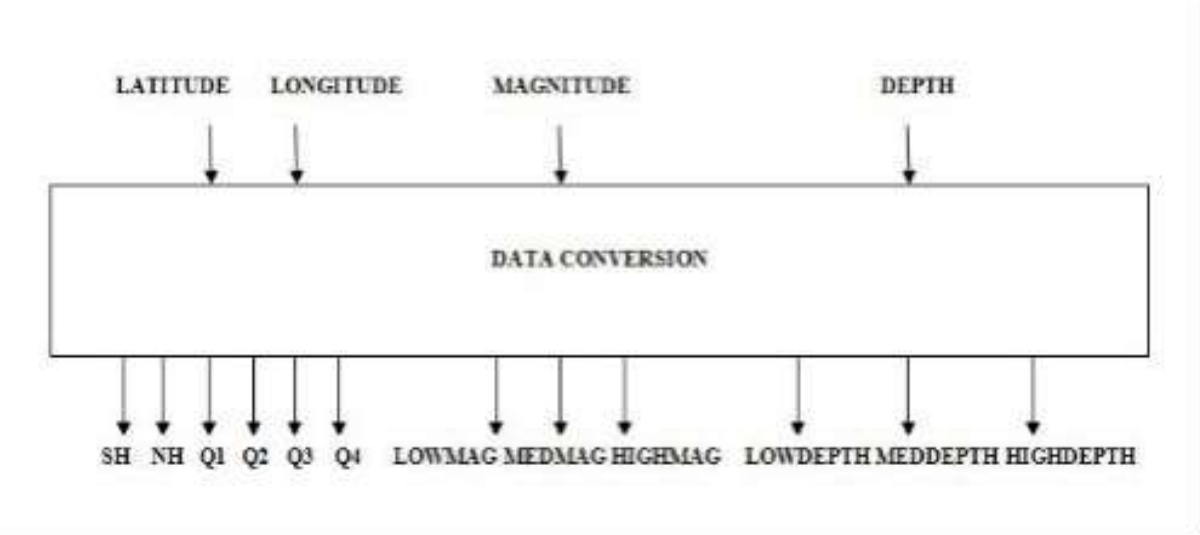

Figure 3. Conversion pattern for attributes.

4.1.3 Association Rule Generation The formatted clustered data obtained from the previous step is given as input to this phase. Here the frequent itemsets are identified for every cluster, itemsets that meet the threshold alone are considered, and then from the obtained frequent itemsets, the association rules are generated. For this process, the Apriori TFP algorithm is used, which finds frequent itemsets by making use of two tree structures. Association rules are created using the rule generation algorithm, where provisions that don't fulfill the minimum confidence are not taken into consideration; only those that are above the threshold are considered. These rules are used for the prediction process.

4.2 Rules Comparison and Prediction

The year and region for which the prediction has to be done are obtained from the user as input, and rule matching is done where the matched rule's consequent provides the prediction result. This obtained consequent, when paired with test data, produces the output, which is the result got. The prediction result is given as output to the end-user.

5. RESULTS AND DISCUSSIONS

Initially, data are grouped into different clusters based on magnitude, significant earthquake data for 1990- 2014 have been collected for experimental purpose. After clustering, outlier removal is performed, here group with minimum value is eliminated, and those clusters retained after the outlier removal only are passed for further

processing. Thus, events with shallow magnitude are removed. Then each cluster data is translated into the binary format by de-normalizing the attributes; Table 2 showcases the de-normalized qualities. For each cluster, frequent itemsets are obtained, and based on this itemset; corresponding association rules are generated. Minimum support of 0.6 is considered for determining the everyday item, and for regulations, the confidence threshold is set to 0.7. Table 3 shows sample rules; the obtained regulations are of the form <antecedent> → <consequent>. Laws that have confidence value above the minimum threshold alone are considered. To perform prediction, the desired region and the year for which forecast has to be carried out is obtained from the user. Based on the input received, rule matching is carried out, which checked the rules against the desired input's previous year data, and the rule consequent yielded the prediction result. The output obtained is the prediction result for the intended year and region. For experimental purpose 8678 seismic events in the year, 2014 was considered among which 1397 events occurred in the first quadrant, the proposed model predicted 1292 events for the input year 2014 and region quadrant 1, which provided the prediction accuracy of about 90%.

Table 2. De-normalized attributes

Original Attribute	De-normalized values
Latitude, Longitude	Southern hemisphere, Northern hemisphere, Quadrant 1, 2, 3, 4
Magnitude	Low_mag, Med_mag, High_mag
Depth	Low_depth, Med_depth, High_depth

Table 3. Sample rules and their associated confidence

Rules	Confidence
nh,depmed → sh,q3,deplow,maglow	0.7778
sh,q3,deplow → sh,q3,deplow,maglow	0.8418
nh,q1,maglow → deplow, maglow	0.9022

6. CONCLUSION

Anticipating earthquakes is still a challenging task without using a diagnostic precursor as they result from sudden energy release from the crust, and the prediction of such natural disasters by pinpointing the exact location and time is not feasible. However, based on the scientific data obtained from USGS, it is possible to find a pattern from these seismic events, which will be helpful in the prediction of the earthquake in the foreseeable future. The target of our research is to find out the pattern obtained from earthquake seismic data. This is accomplished by applying the clustering and association rule mining technique on the global seismic data. Apriori-TFP used for finding frequent patterns efficiently produces rules. The rules obtained from this system are used to predict an earthquake. This long term prediction of the shock may help the government to take necessary precautions, and also it allows seismologists for more precise predictions and the scientists for other researches.