

A Review of Application of Data Mining in Earthquake Prediction

G. V. Otari[#], Dr. R. V. Kulkarni^{*}

[#] Kolhapur Institute of Technology's College of Engineering, Kolhapur, Maharashtra, India

^{*} Shahu Institute of Business Education & Research(SIBER), Kolhapur, Maharashtra, India

Abstract- A *natural disaster* is the effect of a natural hazard (e.g., flood, tornado, hurricane, volcanic eruption, earthquake, heatwave, or landslide). Earthquakes, landslides, tsunamis and volcanos are complex physical phenomenon that leads to financial, environmental or human losses.

Prediction of such geological disasters is the need of the day. Also, prediction of these disasters is a complex process that depends on many physical and environmental parameters. Many approaches exist in the literature based on scientific and statistical analysis. Data mining techniques can also be used for prediction of these natural hazards.

This paper presents a review of application of data mining in the prediction of natural geological calamities. 16 journal articles on the subject published between 1989 and 2011 was analyzed. The main data mining techniques used for earthquake prediction are logistic models, neural networks, the Bayesian belief network, and decision trees, all of which provide primary solutions to the problems inherent in the prediction of earthquakes, tsunamis, landslides and other micro seismic activities. This paper also aims to encourage additional research on topics, and concludes with several suggestions for further research.

1. INTRODUCTION:

Earthquake is the sudden movement of the Earth's crust caused by the abrupt release of stress accumulated along geologic fault in the interior. The energy released passes through the Earth as seismic waves (low-frequency sound waves), which cause the movement. Seismic waves continue to travel through the Earth after the fault motion has stopped. Earthquake prediction research has been going on for nearly a century. A successful prediction, specifying the time, location, and magnitude of an earthquake, would save lives and billions of dollars in housing and infrastructure costs. Unfortunately, successful earthquake predictions are extremely rare. There are two basic categories of earthquake predictions: forecasts (months to years in advance) and short-term predictions (hours or days in advance). Forecasts are based a variety of research, including the history of earthquakes in a specific region, the identification of fault characteristics (including length, depth, and segmentation), and the identification of strain accumulation. Data from these studies are used to provide rough estimates of earthquake sizes and recurrence intervals. An example of an earthquake forecast is the identification of seismic gaps, portions of the plate boundaries that have not ruptured in a major earthquake for a long time. These regions are most likely to experience great earthquakes in the future. Short-term earthquake prediction is still a challenge and no method is known to be reliable. Due to the complex and chaotic nature of the

earthquake process it is being considered that short-term prediction may be inherently impossible.

With the advanced technologies in networks voluminous geographic data have been, and continue to be, collected with modern data acquisition techniques such as global positioning systems (GPS), satellite, high-resolution remote sensing, location-aware services and surveys, and internet-based volunteered geographic information. This results in a need of tools & Technologies for effectively analyzing the scientific data sets with the objective of interpreting the underlying physical phenomena.

Data mining consists of evolving set of techniques that can be used to extract valuable information and knowledge from massive volumes of data. The field of data mining has evolved from its roots in databases, statistics, artificial intelligence, information theory and algorithms into a core set of techniques that have been applied to a range of problems. This paper reviews various data mining techniques that can be applied to the areas such as seismic activity, volcanic eruption, liquefaction and many others factors relevant to earthquake. Data mining applications in geology and geophysics have achieved significant success in the areas as weather prediction, mineral prospecting, ecology, modeling etc and finally predicting the earthquakes from satellite maps. The paper is organized as follows: first, the research methodology used in the study is described; second, the method for classifying articles ; third, articles about data mining in Earthquake prediction are analysed; and finally, the conclusions of the study are discussed.

DATA MINING MODELS:

Data mining is used to find patterns and relationships in data patterns. Two types of models are used to analyze the relationships in data patterns.

1. **Descriptive models:** They describe patterns and create meaningful subgroups or clusters.
2. **Predictive models:** using the patterns in known results forecast explicit values.

There are two flavors of data mining and knowledge discovery in large databases:

1. Event based mining:

- Known events/known algorithms: Use existing physical models (descriptive models and algorithms) to locate known phenomena of interest either spatially or temporally within a large database.
- Known events/unknown algorithms: Use pattern recognition and clustering properties of data to discover new

observational (physical) relationships (algorithms) among known phenomena.

- Unknown events/known algorithms: Use expected physical relationships (predictive models, Algorithms) among observational parameters of physical phenomena to predict the presence of previously unseen events within a large complex database.
- Unknown events/unknown algorithms: Use thresholds or trends to identify transient or otherwise unique events and therefore to discover new physical phenomena.

2. Relationship based mining:

- Spatial Associations: Identify events (e.g. astronomical objects) at the same location. (e.g. same region of the sky)
- Temporal Associations: Identify events occurring during the same or related periods of time.
- Coincidence Associations: Use clustering techniques to identify events that are co-located within a multi-dimensional parameter space.

DATA MINING TECHNIQUES:

The various data mining techniques are

1. Statistics
2. Clustering
3. Visualization
4. Association
5. Classification & Prediction
6. Outlier analysis
7. Trend and evolution analysis

1. Statistics:

◆ Data cleansing i.e. the removal of erroneous or irrelevant data known as outliers.

◆ Data analysis is a measure of association and their relationships between attributes interestingness of rules, classification, prediction etc.

2. Visualization:

◆ Make patterns visible in different views.

3. Clustering(cluster analysis):

◆ Clustering is a process of grouping similar data.

4. Association (correlation and causality):

◆ Mining association rules finds the interesting correlation relationship among large databases.

5. Classification and Prediction:

◆ Finding models (functions) that describe and distinguish classes or concepts for future prediction.

6. Outlier analysis:

◆ Outlier: A data object that is irrelevant to general behavior of the data, it can be considered as an exception but is quite useful in fraud detection in rare events analysis

7. Trend and evolution analysis:

- ◆ Trend and deviation: regression analysis
- ◆ Sequential pattern mining, periodicity analysis
- ◆ Similarity-based analysis

2. RESEARCH METHODOLOGY

As the nature of research in data mining is difficult to confine to specific disciplines, the relevant materials are scattered across various journals. Business intelligence and knowledge discovery are the most common academic discipline for data mining research. Prediction of earthquake can be one of the major aspects of knowledge discovery and also helpful in saving life and economy of most of the countries. Consequently, the following online journal databases were searched to provide a comprehensive bibliography of the academic literature on and Data Mining:

- ACM journals;
- Ingenta Journals;
- Science Direct; and
- IEEE Transaction.

The literature search was based on the descriptor, "Earthquake prediction" and "data mining", which originally produced approximately 500 articles. The full text of each article was reviewed to eliminate those that were not actually related to application of data mining techniques in prediction of earthquake. The selection criteria were as follows:

_ Only those articles that had been published in business intelligence, knowledge discovery or data mining related journals were selected, as these were the most appropriate outlets for data mining research and the focus of this review.

_ Only those articles which clearly described how the mentioned data mining technique(s) could be applied and assisted in earthquake prediction were selected.

_ Conference papers, master's and doctoral dissertations, textbooks and unpublished working papers were excluded, as academics and practitioners alike most often use journals to acquire information and disseminate new findings. Thus, journals represent the highest level of research.

Each article was carefully reviewed and separately classified according to the two categories of prediction dimensions (Neural Networks and data mining). Although this search is not exhaustive, it serves as a comprehensive base for an understanding of data mining application in earthquake prediction.

3. CLASSIFICATION MODEL

Prediction studies can be broadly grouped based on the basic approach, which vary from purely theoretical geophysics, to genetic mutations and biology, to statistical, mathematical, and computational modeling including neural networks, genetic programming and data mining of earthquake parameter data recorded in historical catalogs of seismic regions. The papers reviewed in this article are classified into two groups based on two different approaches: (1) Neural Network based approach; and (2) Data Mining approach.

3.1. NEURAL NETWORK BASED APPROACH

A. Negarestani (2002) [1] proposed layered neural networks based analysis to estimate the radon concentration in soil related to the environmental parameters. This technique can find any functional relationship between the radon concentration and the environmental parameters. Data was

obtained from a site in Thailand and was analyzed. The analysis indicates that this approach is able to differentiate time variation of radon concentration caused by environmental parameters from those arising by anomaly phenomena in the earth (e.g. earthquake). This method is compared with a linear computational technique based on impulse responses from multivariable time series. It is indicated that the proposed method can give a better estimation of radon variations related to environmental parameters that may have a non-linear effect on the radon concentration in soil, such as rainfall.

Adel M. Hanna (2007) [2] in his paper proposed a general regression neural network model to assess nonlinear liquefaction potential of soil. A total of 620 sets of data including 12 soil and seismic parameters are introduced into the model. The data includes the results of field tests from two major earthquakes that took place in Turkey and Taiwan in 1999. The proposed GRNN model was developed in four phases, mainly: identification phase, collection phase, implementation phase, and verification phase. An iterative procedure was followed to maximize the accuracy of the proposed model. The case records were divided randomly into testing, training, and validation datasets. Generating a model that takes into account of 12 soil and seismic parameters is not feasible by using simplified techniques; however, the proposed GRNN model effectively explored the complex relationship between the introduced soil and seismic input parameters and validated the liquefaction decision obtained by simplified methods. The proposed GRNN model predicted well the occurrence/nonoccurrence of soil liquefaction in these sites. The model provides a viable tool to geotechnical engineers in assessing seismic condition in sites susceptible to liquefaction.

Hung-Ming Lin (March 2009)[3] created an empirical model for assessing failure potential of highway slopes, with a special attention to the failure characteristics of the highway slopes in the Alishan, Taiwan area prior to, and post, the 1999 Chi-Chi, Taiwan earthquake. The basis of the study was a large database of 955 slope records from four highways in the Alishan area. Artificial neural network (ANN) was utilized to "learn" from this database. The developed ANN model was then used to study the effect of the Chi-Chi earthquake on the slope failure characteristics in the Alishan area. Significant changes in the degrees of influence of several factors (variables) are found and possible reasons for such changes were discussed. The developed ANN models were used as a tool to investigate the slope failure characteristics before and after the Chi-Chi earthquake.

P.S. Koutsourelakis (January 2010) [4] proposed in his research a probabilistic framework for assessing structural vulnerability against earthquakes. In this paper, a Bayesian framework is proposed for the derivation of fragility curves which can produce estimates irrespective of the amount of data available. It is particularly flexible when combined with Markov Chain Monte Carlo (MCMC) techniques and can efficiently provide credible intervals for the estimates. Furthermore, a general procedure based on logistic regression is illustrated that can lead in a principled manner to the

derivation of fragility surfaces which express the probability of exceeding a damage level with respect to several measures of the earthquake load and can thus produce more accurate predictions. The methodologies presented are illustrated using data generated from computational simulations for a structure on top of a saturated sand deposit which is susceptible to liquefaction.

L. Dehbozorgi during his research investigated an application of Neuro-Fuzzy classifier [5] for short-term earthquake prediction using saved seismogram data. This method is able to predict earthquakes five minute before, with an acceptable accuracy (82.8571%). The features were obtained from statistical and entropy parameters, Discrete Wavelet Transform (DWT), Fast Fourier Transform (FFT), Chaotic Features (Maximum Lyapunov Exponent), estimated power spectral density (PSD), and the classifier used this extracted features to indicate whether the earthquake were takes place in the next following five minutes or not. Finally, after training of Neuro-Fuzzy classifier effective features were selected with UTA algorithm.

3.2. DATA MINING APROACH

B. Zmazek (June 2003) [6] used different regression methods to predict radon concentration in soil gas on the basis of environmental data, i.e. barometric pressure, soil temperature, air temperature and rainfall. Analyses of the radon data from three stations in the Krško basin, Slovenia, have shown that model trees outperform other regression methods. A model has been built which predicts radon concentration with a correlation of 0.8, provided it is influenced only by the environmental parameters. In periods with seismic activity this correlation is much lower. This decrease in predictive accuracy appears 1–7 days before earthquakes with local magnitude 0.8–3.3.

An automatic system "CQuake" [7] has been developed by Guido Cervone to carry out spatial and temporal data mining analysis in real time. Satellite remote sensing data are used in providing information about changes in land, ocean, atmosphere and ionosphere. Analysis of data shows that some of the parameters observed from the remote sensing data are associated with impending coastal earthquakes. CQuake performs retrospective analysis of earthquakes, and performs forecast for predefined regions of the world based on the analysis of Surface latent heat flux (SLHF) or any other geophysical parameters.

A.M. Posadas(1993) [8] applied the three point method (TPM) to several seismic series and has provided information about the spatial characteristics (azimuth and dip) of the fault planes activated in the rupture process. A new development of the TPM to determine temporal characteristics, is presented, to obtain the evolution of the fracturing process of an active fault system. For the analysis of the 158 microearthquakes and earthquakes that took place in the seismic series of Antequera in June 1989, the choice of a threshold magnitude ($m_u = 2.5$) has permitted the events related to the most relevant fractures to be distinguished. Only the events between two concentric spheres (here named Spatial Crown) with respect to a given earthquake, have been used in order to avoid taking into account earthquakes that

are too close to each other, together with the very distant events that have little relation to the event analysed. The Spatial Crown has permitted some clear results in the Antequera series, where it is found that the fracturing process began fundamentally with N 80° E planes and evolved to N 65° W planes.

Zhengzheng Xing [9] used two methods for solving the problem of mining sequence classifiers for early prediction. The sequential classification rule (SCR) method mines a set of sequential classification rules as a classifier. A so-called early-prediction utility is defined and used to select features and rules. The generalized sequential decision tree (GSDT) method adopts a divide-and-conquer strategy to generate a classification model. The two methods achieve accuracy comparable to that of the state-of-the-art methods, but typically need to use only very short prefixes of the sequences. The results clearly indicate that early prediction is highly feasible and effective. Since temporal order is important in sequence data, in many critical applications of sequence classification such as medical diagnosis and disaster prediction, *early prediction* is a highly desirable feature of sequence classifiers.

Ifitikhar U. Sikder (January 2009) [10] presented a machine learning approach to characterizing premonitory factors of earthquake. The characteristic asymmetric distribution of seismic events and sampling limitations make it difficult to apply the conventional statistical predictive techniques. The inductive machine learning techniques such as rough set theory and decision tree (C4.5 algorithm) allows developing knowledge representation structure of seismic activity in term of meaningful decision rules involving premonitory descriptors such as space-time distribution of radon concentration and environmental variables. Both the techniques identify significant premonitory variables and rank attributes using information theoretic measures, e.g., entropy and frequency of occurrence in reducts.

Witold Dzwinel [11] developed a novel technique based on cluster analysis of the multi-resolutional structure of earthquake patterns and applied to observed and synthetic seismic catalogs. The observed data represent seismic activities situated around the Japanese islands in the 1997-2003 time interval. The synthetic data were generated by numerical simulations for various cases of a heterogeneous fault governed by 3-D elastic dislocation and power-law creep. At the highest resolution, the local cluster structure was analyzed in the data space of seismic events for the two types of catalogs by using an agglomerative clustering algorithm. It was found that small magnitude events produce local spatio-temporal patches corresponding to neighboring large events. Seismic events, quantized in space and time, generate the multi-dimensional feature space of the earthquake parameters. Using a non-hierarchical clustering algorithm and multidimensional scaling, the multitudinous earthquakes were explored by real-time 3-D visualization and inspection of multivariate clusters. At the resolutions characteristic of the earthquake parameters, all of the ongoing seismicity before and after largest events accumulate to a global structure consisting of a few separate clusters in the

feature space. By combining the clustering results from low and high resolution spaces, precursory events were recognized more precisely.

Dave A.Yuenl [12] presented a web client-server service WEB-IS for remote analysis and visualization of seismic data consisting for both small and large earthquakes. A problem-solving environment (PSE) designed for predicting of large magnitude earthquakes can be based on this WEB-IS idea. The clustering schemes, feature generation, feature extraction techniques and rendering algorithms form a computational framework for this environment. Easy and fast access to both the seismic data distributed among distant computing resources and to computational and visualization resources can be realized within a GRID framework. NaradaBrokering (iNtegrated Asynchronous Real-time Adaptive Distributed Architecture) was used as a flexible middleware for providing a high throughput in remote visualization of geophysical data. The WEB-IS functionality was tested for both synthetic and actual earthquake catalogs.

Aydin [13] proposed a prediction algorithm using time series data mining based on fuzzy logic is proposed. Earthquake prediction has been done from a synthetic earthquake time series by using investigating method at first step ago. Time series has been transformed to phase space by using nonlinear time series analysis and then fuzzy logic has been used to prediction optimal values of important parameters characterizing the time series events. Truth of prediction algorithm based fuzzy logic has been proved by application results.

Seismic data is generated in nature by the changes or movement of the earth crust. This data has evolutionary patterns. Since this data is based on time, a model can be formed to predict the future pattern. Sajjad Mohsin[14] in his work used both deterministic and un-deterministic optimized algorithms to determine the future values. The results of different applied techniques show the possibility of future earthquakes in Pakistan region.

Clustered events are usually deemed as feature when several spatial point processes are overlaid in a region. They can be perceived either as a precursor that may induce a major event to come or as offspring triggered by a major event. Hence, the detection of clustered events from point processes may help to predict a forthcoming major event or to study the process caused by a major event. Tao Pei [15] employed the support domain of feature (SDF), the region over which any feature event has the equivalent likelihood to occur, to approximate the “territory” of feature events. A method is developed to delineate the SDF from a region containing spatial point processes. The method consists of three major steps. The first is to construct a discrimination function for separating feature points from noise points. The second is to divide the entire area into a regular mesh of points and then compute a fuzzy membership value for each grid point belonging to the SDF. The final step is to trace the boundary of the SDF.

W. Dzwinel [16] in his paper presented a novel technique based on a multiresolutional clustering and nonlinear multidimensional scaling of earthquake patterns to investigate

observed and synthetic seismic catalogs. The synthetic data were generated by numerical simulations for various cases of a heterogeneous fault governed by 3-D elastic dislocation and power-law creep. At the highest resolution, he analyzed the local cluster structures in the data space of seismic events for the two types of catalogs by using an agglomerative clustering algorithm. He demonstrated that small magnitude events produce local spatiotemporal patches delineating neighboring large events. Seismic events, quantized in space and time, generate the multidimensional feature space characterized by the earthquake parameters. Using a non-hierarchical clustering algorithm and nonlinear multi-dimensional scaling, multitudinous earthquakes are explored by real-time 3-D visualization and inspection of the multivariate clusters. At the spatial resolutions characteristic of the earthquake parameters, all of the ongoing seismicity both before and after the largest events accumulates to a global structure consisting of a few separate clusters in the feature space. It has been shown that by combining the results of clustering in both low and high resolution spaces, the precursory events can be recognized more precisely and unravel vital information that cannot be discerned at a single resolution.

4. CONCLUSION:

A critical part of any new research venture is the construction of a good classification framework and the organization of a reference collection of relevant literature. The research area of earthquake prediction is no exception. Although the importance of data mining techniques in the prediction of earthquake has been recognized, a comprehensive classification framework or a systematic review of their application is lacking.

In this study, the authors conduct an extensive review of academic articles and provide a comprehensive bibliography and classification framework for the applications of data mining to earthquake prediction. Author's intention is to inform both academics and practitioners of the areas in which specific data mining techniques can be applied to prediction of earthquake. Although the study cannot claim to be exhaustive, authors believe that it will prove a useful resource for anyone interested in earthquake prediction research, and will help simulate further interest in the field.

REFERENCES

- [1] A. Negarestani, S. Setayeshi, M. Ghannadi-Maragheh, B. Akashe, "Layered neural networks based analysis of radon concentration and environmental parameters in earthquake prediction", *Journal of Environmental Radioactivity*, Volume 62, Issue 3, 2002, Pages 225-233.
- [2] Adel M. Hanna, Derin Ural, Gokhan Saygili, "Neural network model for liquefaction potential in soil deposits using Turkey and Taiwan earthquake data", *Soil Dynamics and Earthquake Engineering*, Volume 27, Issue 6, June 2007.
- [3] Hung-Ming Lin, Shun-Kung Chang, Jian-Hong Wu, C. Hsein Juang, "Neural network-based model for assessing failure potential of highway slopes in the Alishan, Taiwan Area: Pre- and post-earthquake investigation", *Engineering Geology*, Volume 104, Issues 3-4, 23 March 2009.
- [4] P.S. Koutsourelakis, "Assessing structural vulnerability against earthquakes using multi-dimensional fragility surfaces: A Bayesian framework", *Probabilistic Engineering Mechanics*, Volume 25, Issue 1, January 2010.
- [5] Dehbozorgi, L.; Farokhi, F., "Effective feature selection for short-term earthquake prediction using Neuro-Fuzzy classifier", *Centran Tehran Branch, Sci. Assoc. of Electr. & Electron. Eng., Islamic Azad Univ., Tehran, Iran*.
- [6] B. Zmazek, L. Todorovski, S. Džeroski, J. Vaupotič, I. Kobal, "Application of decision trees to the analysis of soil radon data for earthquake prediction", *Applied Radiation and Isotopes*, Volume 58, Issue 6, June 2003.
- [7] Guido Cervone, Menas Kafatos, Domenico Napoletani, Ramesh P. Singh, "An early warning system for coastal earthquakes", *Advances in Space Research*, Volume 37, Issue 4, 2006.
- [8] A.M. Posadas, F. Vidal, J. Morales, J.A. Peña, J. Ibañez, F. Luzon, "Spatial and temporal analysis of a seismic series using a new version of the three point method: application to the 1989 Antequera (Spain) earthquakes", *Physics of the Earth and Planetary Interiors*, Volume 80, Issues 3-4, November 1993.
- [9] Zhengzheng Xing, Jian Pei, Guozhu Dong, Philip S. Yu, "Mining Sequence Classifiers for Early Prediction".
- [10] Iftikhar U. Sikder, Toshinori Munakata, "Application of rough set and decision tree for characterization of premonitory factors of low seismic activity", *Expert Systems with Applications*, Volume 36, Issue 1, January 2009.
- [11] Witold Dzwinel¹, David A.Yuen², Krzysztof Boryczko^{1,2}, Yehuda Ben-Zion³, Shoichi Yoshioka⁴, Takeo Ito, "Cluster Analysis, Data-Mining, Multi-dimensional Visualization of Earthquakes over Space, Time and Feature Space".
- [12] Dave A.Yuen¹, Benjamin J Kadlec¹, Evan F Bollig¹, Witold Dzwinel², Zachary A.Garbow¹, Cesar da Silva³, "Clustering and Visualization of Earthquake data in a Grid Environment".
- [13] I. Aydin, M. Karakose, and E. Akin, "The Prediction Algorithm Based on Fuzzy Logic Using Time Series Data Mining Method".
- [14] Sajjad Mohsin, and Faisal Azam, "Computational seismic algorithmic comparison for earthquake prediction", *INTERNATIONAL JOURNAL OF GEOLOGY* Issue 3, Volume 5, 2011.
- [15] Tao Peia, A-Xing Zhua, Chenghu Zhou, Baolin Li, Chengzhi Qin, "Delineation of support domain of feature in the presence of noise", *Computers & Geosciences* 33 (2007).
- [16] W. Dzwinel, D. A. Yuen, K. Boryczko, Y. Ben-Zion, S. Yoshioka, and T. Ito, "Nonlinear multidimensional scaling and visualization of earthquake clusters over space, time and feature space", *Nonlinear Processes in Geophysics* (2005) 12: 117-128, SRef-ID: 1607-7946/npg/2005-12-117, European Geosciences Union© 2005 Author(s).