IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Special Issue, 4, Feb 2019 E-ISSN: 2347-2693

A Survey on Earth Quakes Prediction Techniques with Clustering Methods

A. Mary Subashini

Department of Computer Applications, Idhaya College for Women, Kumbakonam, India

Corresponding Author: Subasingh2013@gmail.com

Available online at: www.ijcseonline.org

Abstract**—** 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. Computational simulation and data acquisition in scientific and engineering domains have made tremendous progress over the past two decades. A mix of advanced algorithms, exponentially increasing computing power and accurate sensing and measurement devices have resulted in more data repositories. Advanced technologies in networks have enabled the communication of large volumes of data across the world.This paper aims at further data mining study on scientific data. This paper highlights the data mining techniques applied to mine for surface changes over time (e.g. Earthquake rupture). The data mining techniques help researchers to predict the changes in the intensity of volcanos. This paper uses predictive statistical models that can be applied to areas such as seismic activity , the spreading of fire. The basic problem in this class of systems is unobservable dynamics with respect to earthquakes. The space-time patterns associated with time, location and magnitude of the sudden events from the force threshold are observable. This paper highlights the observable space time earthquake patterns from unobservable dynamics using data mining techniques, pattern recognition and ensemble forecasting. Thus this paper gives insight on how data mining can be applied in finding the consequences of earthquakes and hence alerting thepublic.

Keywords**—** Earthquake, data mining techniques, space-time patterns

I. INTRODUCTION

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 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 satellitemaps. An interesting aspect of many of these applications is that they combine both spatial and temporal aspects in the data and in the phenomena that is being mined. Data sets in these applications come from both observations and simulation. Investigations on earthquake predictions are based on the assumption that all of the regional factors can be filtered out and general information about the earthquake precursory patterns can beextracted.

Feature extraction involves a pre selection process of various statistical properties of data and generation of a set of seismic parameters, which correspond to linearly independent coordinator in the feature space.

The seismic parameters in the form of time series can be analyzed by using various pattern recognition techniques. Statistical or pattern recognition methodology usually performs this extraction process. Thus this paper gives insight of mining the scientific data.

II. DATAMINING-DEFINITIONS

- Data mining is defined as process of extraction of relevant data and hidden facts contained in databases and datawarehouses.
- It refers to find out the new knowledge about an application domain using data on the domain usually stored in the databases. The application domain may be astrophysics, earth science or about solarsystem.

Data mining techniques support to identify nuggets of information and extracting this information in such a way that, this will support in decision making, prediction, forecasting andestimation.

III. DATA MININGGOALS

- Bring together representatives of the data mining community and the domain science community so that they can understand the current capabilities and research objectives of each other communities related to datamining.

- Identify a set of research objectives from the domain science community that would be facilitated by current or anticipated data mining techniques.
- Identify a set of research objectives for the data mining community that could support the research objectives of the domain science community.

IV. DATA MININGMODELS

Data mining is used to find patterns and relationships in data patterns. The relationships in data patterns can be analyzed via 2 types ofmodels.

1. Descriptivemodels
2. Predictivemodels

This paper focuses on predictive models.

In large databases data mining and knowledge discovery comes in two flavors:

a) Event basedmining:

- Known events/knownalgorithms
- Known events/unknownalgorithms
- Unknown events/knownalgorithms
- Unknown events/unknown algorithms

This paper focuses on unknown events and known algorithms.

b) Relationship basedmining:

- SpatialAssociations
- TemporalAssociations
- CoincidenceAssociations

This paper focuses on all relationship-based mining.

c) User requirements for data mining in largescientific databases:

- Crossidentifications
- Crosscorrelation
- Nearest neighbor identification
- Systematic dataexploration

This paper focuses on correlation and Clustering.

V. DATA MININGTECHNIQUES THE VARIOUS DATA MINING TECHNIQUESARE

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

VI. EARTHQUAKEPREDICTION

i. Ground waterlevels
ii. Chemical changes in Groundwater
iii. Radon Gas in Ground water wells

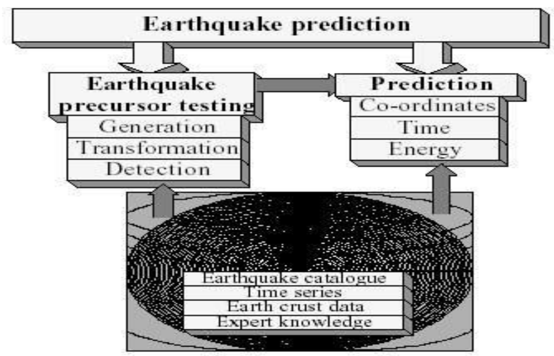

Fig 1: Although radon has relatively a short half

i) Ground WaterLevels

Changing water levels in deep wells are recognized as precursor to earthquakes. The pre-seismic variations at observation wells are asfollows.

1. A gradual lowering of water levels at a period of months or years.
2. An accelerated lowering of water levels in the last few months or weeks preceeding the earthquake.
3. A rebound, where water levels begin to increase rapidly in the last few days or hours before the mainshock.

ii) Chemical Changes in Groundwater

1. The Chemical composition of ground water is affected by seismicevents.
2. Researchers at the university of Tokyo tested the water after the earthquake occured, the result of the study showed that the composition of water changed significantly in the period around earthquake area.
3. They observed that the chloride concentration is almostconstant.
4. Levels of sulphate also showed a similar rise.

iii) Radon Gas in Ground waterwells

An increase level of radon gas in wells is a precursor of earthquakes recognized by research group. Life and is unlikely to seep the surface through rocks from the depths at which seismic is very soluble in water and can routinely be monitored in wells and springs often radon levels at such springs show reaction to seismic events and they are monitored for earthquakepredictions..

1. There is no effective solution to the problem.
2. To solve this problem earthquake catalogs, geo-monitoring time series data about stationary seismo-tectonic properties of geological environment and expert knowledge andhypotheses.

3. To solve this problem earthquake catalogs, geo-monitoring time series data about stationary seismo-tectonic properties of geological environment and expert knowledge and hypotheses about earthquake precursors.

This proposes a multi-resolutional approach, which combines local clustering techniques in the data space with a nonhierarchical clustering in the feature space.

The raw data are represented by n-dimensional vector Xi of measurements Xk. The data space can be searched for patterns and can be visualized by using local or remote pattern recognition and by advanced visualization capabilities. The data space X is transformed to a new abstract space Y of vectors Yj . The coordinates Yl of these vectors represent nonlinear functions of measurements Xk, which are averaged in space and time in given space-time windows. This transformation allows for coarse graining of data (data quantization), amplification of their characteristic features and suppression of the noise and other random components. The new features Yl form a N-dimensional feature space. We use multi-dimensional scaling procedures for visualizing the multi-dimensional events in 3D space. This transformation allows a visual inspection of the N-dimensional feature space. The visual analysis helps greatly in detecting subtle cluster structures which are not recognized by classical clustering techniques, selecting the best pattern detection procedure used for data clustering, classifying the anonymous data and formulating newhypothesis.

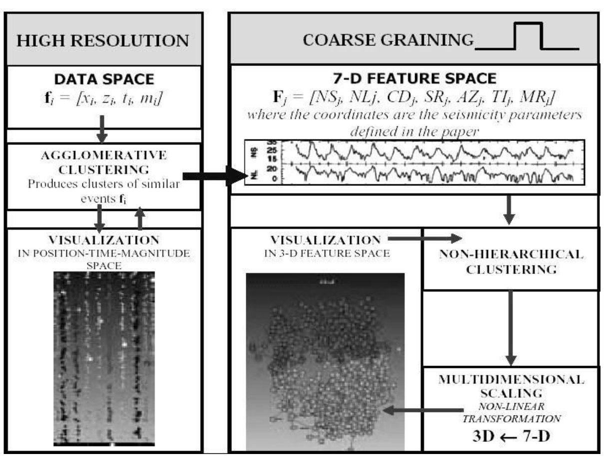

Fig. 2

Clustering schemes Clustering analysis is a mathematical concept whose main role is to extract the most similar separated sets of objects according to a given similarity measure. This concept has been used for many years in pattern recognition. Depending on the data structures and goals of classification, different clustering schemes must beapplied.

In our new approach we use two different classes of clustering algorithms for different resolutions. In data space we use agglomerative schemes, such as modified Mutual Nearest Neighbour algorithm (MNN). This type of clustering extracts the localized clusters in the high resolution data space. In the feature space we are searching for global clusters of time events comprising similar events from the whole time interval.

$$D_{W1}(C_p , C_q) = (n_p n_q / n_p + n_q) \| \bar{x}\, p - \bar{x}\, q \|^2$$

The non-hierarchical clustering algorithms are used mainly for extracting compact clusters by using global knowledge about the data structure. We use improved mean based schemes, such as a suite of moving schemes, which uses the k-means procedure and four strategies of its tuning by moving the data vectors between clusters to obtain a more precise location of the minimum of the goal function.

wherezj is the position of the center of mass of the cluster j , while xi are the feature vectors closest to zj . To find a global minimum of function J (), we repeat the clustering procedures at different initial conditions. Each new initial configuration is constructed in a special way from the previous results by using the methods. The cluster structure with the lowest J (w, n) minimum isselected.

VII. HIERARCHICAL CLUSTERINGMETHOD

A hierarchical clustering method produces a classification in which small clusters of very similar molecules are nested within larger clusters of less closely-related molecules. Hierarchical agglomerative methods generate a classification in a bottomup manner, by a series of agglomerations in which small clusters, initially containing individual molecules, are fused together to form progressively larger clusters.

Hierarchical agglomerative methods are often characterized by the shape of the clusters they tend to find, as exemplified by the following range: single-link tends to find long, straggly, chained clusters; Ward and group average - tend to find globular clusters; complete-link - tends to find extremely compact clusters. Hierarchical divisive methods generate a classification in a topdown manner, by progressively sub-dividing the single cluster which represents an entire dataset. Monothetic (divisions based on just a single descriptor) hierarchical divisive methods are generally much faster in operation than the corresponding polythetic (divisions based on all descriptors) hierarchical divisive and hierarchical agglomerative methods, but tend to give poor results. One problem with these methods is how to choose which clusters or partitions to extract from the hierarchy because display of the complete hierarchy is not really appropriate for data sets of more than a few hundred compounds.

VIII. NON-HIERARCHICAL CLUSTERINGMETHODS

A non-hierarchical method generates a classification by partitioning a dataset, giving a set of (generally) non-overlapping groups having no hierarchical relationships between them. A systematic evaluation of all possible partitions is quite infeasible, and many different heuristics have described to allow the identification of good, but possibly suboptimal, partitions. Three of the main categories of non-hierarchical method are single- pass, relocation and nearest neighbour. Single-pass method (e.g. Leader) produce clusters that are dependent upon the order in which the compounds are processed, and so will not be considered further. Relocation methods, such as k-means, assign compounds to a user-defined number of seed clusters and then iteratively reassign compounds to produce the better clusters result. Such methods are prone to reaching local optimum rather than a global optimum, and it is generally not possible to determine when or where the global optimum solution has been reached. Nearest neighbour methods, such as the Jarvis-Patrick method, assign compounds to the same cluster as some numberoftheirnearestneighbours.User-defined parameters determine how many nearest neighbours need to be considered, and the necessary level of similarity between nearest neighbour lists. Other non- hierarchical methods are generally inappropriate for use on large, highdimensional datasets such as those used in chemical applications.

$$S(P) = (1 / n)\ S\ \{\ s(x_i) \mid x_i\ Î\ È\ \{C_p \mid p\ Î\ P\}\}$$

with n being the number of objects concerned by the current partition P. When C includes a bipartition {C¢ , C²} the functions a and bbecome:

$$\text{for}\,x_i\,\text{in}\,C¢ \qquad a(x_i) = \bar{d}(\{x_i\}, C¢ - \{x_i\})$$

$$\text{for}\,x_i\,\text{in}\,C² \qquad a(x_i) = \bar{d}(\{x_i\}, C² - \{x_i\})$$

IX. DATA MININGAPPLICATIONS

- In Scientific discovery – super conductivity research, For Knowledge Acquisition.
- In Medicine – drug side effects, hospital cost analysis, genetic sequence analysis, prediction etc.
- In Engineering – automotive diagnostics expert systems, fault detectionetc.,
- In Finance – stock market perdition, credit assessment, fraud detectionetc.

X. FUTUREENHANCEMENTS

The future of data mining lies in predictive analytics. The technology innovations in data mining since 2000 have been truly Darwinian and show promise of consolidating and stabilizing around predictive analytics. Nevertheless, theemerging market for predictive analytics has been sustained by professional services, service bureausand profitable applications in verticals such as retail, consumer finance, telecommunications, travel and leisure, and related analytic applications. Predictive analytics have successfully proliferated into applications to support customer recommendations, customer value and churn management, campaign optimization, and fraud detection. On the product side, success stories in demand planning, just in time inventory and market basket optimization are a staple of predictive analytics. Predictive analytics should be used to get to know the customer, segment and predict customer behavior and forecast product demand and related market dynamics. Finally, they are at different stages of growth in the life cycle of technologyinnovation.

XI. CONCLUSION

The problem of earthquake prediction is based on data extraction of precursory phenomena and it is highly challenging task various computational methods and tools are used for detection of pre-cursor by extracting general information from noisydata.

By using common frame work of clustering we are able to perform multiresolutional analysis of seismic data starting from the raw data events described by their magnitude spatio-temporal data space. This new methodology can be also used for the analysis of the data from the geological phenomena e.g. We can apply this clustering method to volcanic eruptions.

REFERENCES

[1]. W.Dzwinel et al Non multidimensional scaling and visualization of earth quake cluster over space and feature space, nonlinear processes in geophysics 12[2005]pp1-12.

[2]. C.Lomnitz.FundamentalsofEarthquakeprediction[1994].

[3]. B.Gutenberg& C.H. Richtro, Earthquake magnitude, intensity, energy & acceleration bulseismsoc. Am 36, 105-145 [1996].

[4]. C.Brunk, J.Kelly&Rkohai "MinesetAn integrate system for Visual Data Mining" 1997.

[5]. MAdFraud: Investigating Ad Fraud in Android Applications.

[6]. Mining Personal Context-Aware Preferences for Mobile Users.

[7]. A Flexible Generative Model for Preference Aggregation.

[8]. D. M. Blei, A. Y. Ng, and M. I. Jordan. Lantentdirichlet allocation.*Journal of Machine Learning Research*, pages 993–1022, 2003.

[9]. Y. Ge, H. Xiong, C. Liu, and Z.-H. Zhou. A taxi driving fraud detection system. In *Proceedings of the 2011 IEEE 11th International Conference Neural Networks*, ICNN '11, pages 181–190, 2011.

[10]. N. Spirin and J. Han, "Survey on webspam detection: Principles and algorithms," SIGKDD Explor. Newslett., vol. 13, no. 2, pp. 50–64,May2012.

[11]. E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, "Detecting product review spammers using rating behaviors," in Proc.19thACMInt. Conf. Inform. Knowl.Manage., 2010, pp. 939–948.

[12]. G. Ateniese, R. Di Pietro, L. V. Mancini, and G. Tsudik,

[13]. "Scalable and efficient provable data possession," in Proc.SecureComm, 2008, Art. ID 9.C. C. Erway, A. Küpçü, C. Papamanthou, and R.Tamassia, "Dynamic provable data possession," in Proc. CCS,2009, pp. 213–222.

[14]. E. Esiner, A. Küpçü, and Ö. Özkasap, "Analysis and optimization on FlexDPDP: A practical solution for dynamic provable data possession," Intelligent Cloud Computing(Lecture Notes in Computer Science), vol. 8993. Berlin, Germany: Springer-Verlag, 2014, pp. 65–83.

[15]. E. Zhou and Z. Li, "An improved remote data possession checking protocol in cloud storage," in Algorithms andArchitectures for Parallel Processing (Lecture Notes),105- 110.

[16] Opinion spam and analysis. In Proceedings of the 2008 International Conference on Web Search and Networking, WSDM '08, pages 219–230, 2008.

