GOVERNMENT POLYTECHNIC COLLEGE MATTANNUR-670702

(Department of Technical Education, Kerala)

SEMINAR REPOPORT ON

EARTHQUAKE PREDICATION USING DATA MINING

SUBMITTED BY

SANALRAJ M

(Reg.No:19041677)

DEPARTMENT OF ELECTRONICS ENGINEERING

2021-22

GOVERNMENT POLYTECHNIC COLLEGE
MATTANNUR-670702

(Department of Technical Education, Kerala)

CERTIFICATE

Certified that seminar work entitled “ ***EARTHQUAKE PREDICTION USING DATA MINING***”*is a bonafide work carried out by* **“*SANALRAJ M*”** *in partial fulfilment for the award of Diploma in Electronics Engineering from Government Polytechnic College Mattannur during the academic year* 2021-2022.

SeminarCo-ordinator **Head ofSection**

InternalExaminer **ExternalExaminer**

DECLARATION

I hereby declare that the report of *the* ***EARTHQUAKE PREDICTION USING DATA MINING*** work entitled which is being submitted to the Govt. Polytechnic College Mattannur, in partial fulfilment of the requirement for the award **of *Diploma in Electronics Engineering*** *i*s a confide report of the work carried out by me. The material in this report has not been submitted to any institute for the award of any degree.

Place:Mattannur **SANALRAJ M**

Date:

ACKNOWLEDGMENT

I would like to take this opportunity to extend my sincere thanks to people who helped me to make this seminar possible. This seminar will be incomplete without mentioning all the people who helped me to make itreal.

Firstly, I would like to thank GOD, almighty, our supreme guide, for bestowing his blessings upon me in my entire endeavor.

I would like to express my deepest gratitude **Mr. M C PRAKASHAN** (Principal GPTC, Mattannur), **Mr. GEORGE KUTTY P P (**Head of Department of Electronics Engg.), for the help rendered by him to prepare and present this Seminar in proper way.Moreover I am very much indebted to **Mr. SREEJITH A** (Lecturer, Electronics Engg, seminar co-ordinator), for their advice.

I am also indebted to all my friends and classmates who have given valuable suggestion and encouragement.

SANALRAJ M

ABSTRACT

Data mining consists of evolving set of techniques that can be used to extract valuable information and knowledge from massive volumes of data. Data mining research & tools have focussed on commercial sector applications. Only a fewer data mining research have focused on scientific data. 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 (eg Earthquake rupture). The data mining techniques help researchers to predict the changes in the intensity of volcano. This paper uses predictive statistical models that can be applied to areas such as seismic activity or the spreading of fire. The basic problem in this class of systems is dynamic, usually unobservable with respect to earthquake

The space-time patterns associated with time, location and magnitude of the sudden events from the force threshold are observable. This paper highlights 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 earthquake and warning the scientific, hence alerting the public.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

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 in network technologies have enabled communication of large volumes of data across geographically distant hosts. This results in an need of tools & Technologies for effectively analyzing scientific data sets with the objective of interpreting the underlying physical phenomena. Data mining applications in geology and geophysics were among the first to pursued and have achieved significant success in such areas as weather prediction, mineral prospecting, ecology, modeling and predicting earthquake from satellite maps.

An interesting aspect of many of these applications is that they combine spatial and temporal aspects both in the data and in the phenomena being mined. Data sets in these applications come both from observations and simulation. Investigation 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 be extracted. 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 co-ordinator in the feature space. The seismic parameters in the form of time series can be analyzed by using various pattern recognition techniques.Using statistical or pattern recognition methodology usually performs this extraction process. Thus this paper gives insight of mining the scientific data.

CHAPTER 2

DATA MINING DEFINITION AND GOALS

2.1 Data mining definition

• Data mining is defined as an information extraction activity whose goal is to discover hidden facts contained in databases.

• It refers to finding out new knowledge about an application domain using data on the domain usually stored in a database. The application domain may be astrophysics, earth science solar system science.

• It’s a variety of techniques to identify nuggets of information or decision making knowledge in bodies of data and extracting these in such a way they can be put to use in the areas such as decision support, prediction ,forecasting and estimation.

2.2 Data mining goals

• Bring together representatives of the data mining community and the domain science community so that they can begin to understand the currents capabilities and research objectives of each others communities related to data mining.

• 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.

CHAPTER 3

DIFFERENT TYPES OF MINING

3.1 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.

3.2 Relationship based mining

- Spatial Associations: Identify events (eg astronomical objects) at the same location. (eg 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.

User requirements for data mining in large scientific databases

- Cross identifications: Refers to the classical problem of associating the source list in one database to the source list in another.
- Cross correlation: Refers to the search for correlations, tendencies, and trends between physical parameters in multidimensional data usually across databases.
- Nearest neighbor identification. Refers to the general application of clustering algorithms in multidimensional parameter space usually within a database.

CHAPTER 4

DATA MINING TECHNIQUE

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

4.1 Statistics:

♦Can be used in several data mining stages

• Data cleansing ie the removal of erroneous or irrelevant data known as outliers.

• EDA Exploratory data analysis eg frequency counts histograms.

• Data selection sampling facilities and so reduce the scale of computation

• Attribute redefinition eg bodies mass index, BMI which is weight/height2.

• Data analysis –measures of association and relationships between attributes interestingness of rules, classification etc.

4.2 Visualization:

♦ Enhances EDA , makes patterns more visible

4.3 Clustering

♦ Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
♦ Clustering based on the principle: maximizing the intra-class similarity and minimizing the intra class similarity
♦ Clustering and segmentation is basically partitioning the database so that each partition or group is similar according to some criteria or metric.
♦ Data mining applications make use of clustering according to similarity eg to segment a client/ customer base
♦ It provides subgroups of population for further analysis or action –very important when dealing with large databases.

4.4 Association (correlation and causality)

♦ Multi-dimensional vs. single-dimensional association age(X, “20..29”) ^ income(X, “20..29K”) à buys(X, “PC”) [support = 2%, confidence = 60%] contains (T,“computer”) à contains(x, “software”) [1%, 75%]

4.5 Classification and Prediction

♦ Finding models (functions) that describe and distinguish classes or concepts for future prediction e.g., classify countries based on climate, or classify cars based on gas mileage
♦ Presentation: decision-tree, classification rule, neural network
♦ Prediction: Predict some unknown or missing numerical values

4.6 Outlier analysis

♦ Outlier: a data object that does not comply with the general behavior of the data It can be considered as noise or exception but is quite useful in fraud detection, rare events analysis

4.7 Trend and evolution analysis

♦ Trend and deviation: regression analysis

♦ Sequential pattern mining, periodicity analysis

♦ Similarity-based analysis

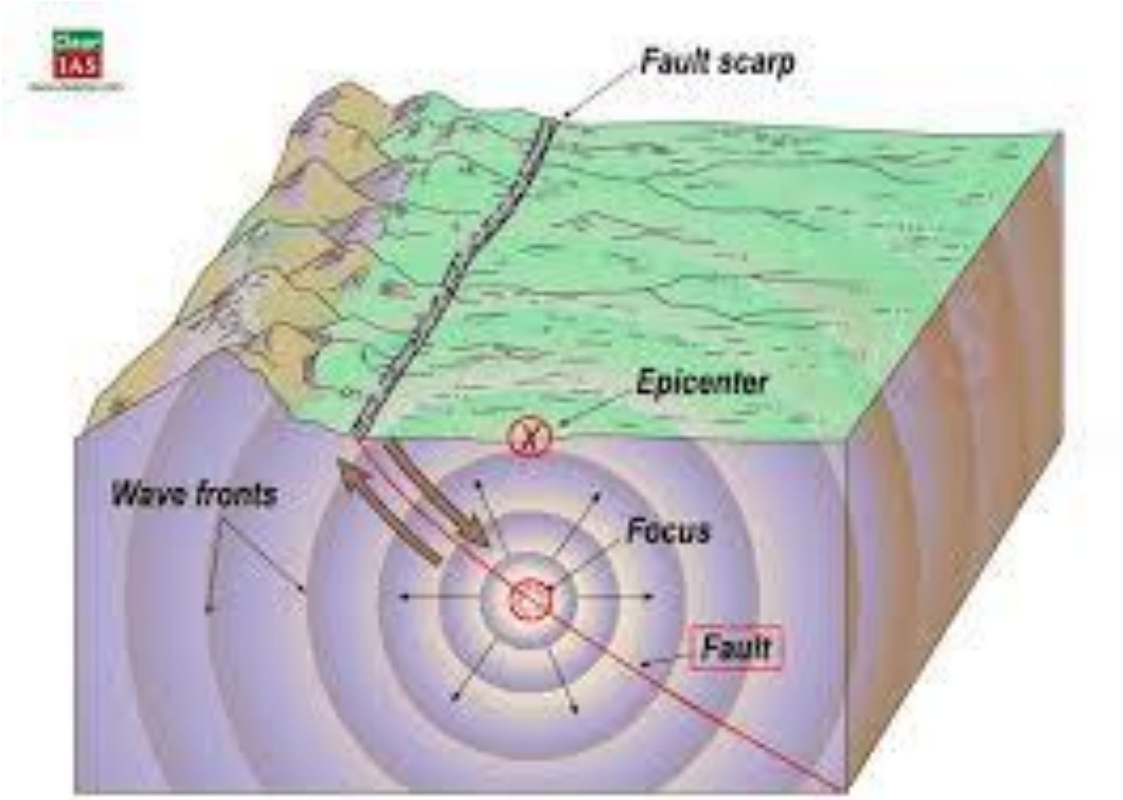

Fig 4.1 Earthquake analysis

CHAPTER 5

EARTH QUAKE PREDICTION.

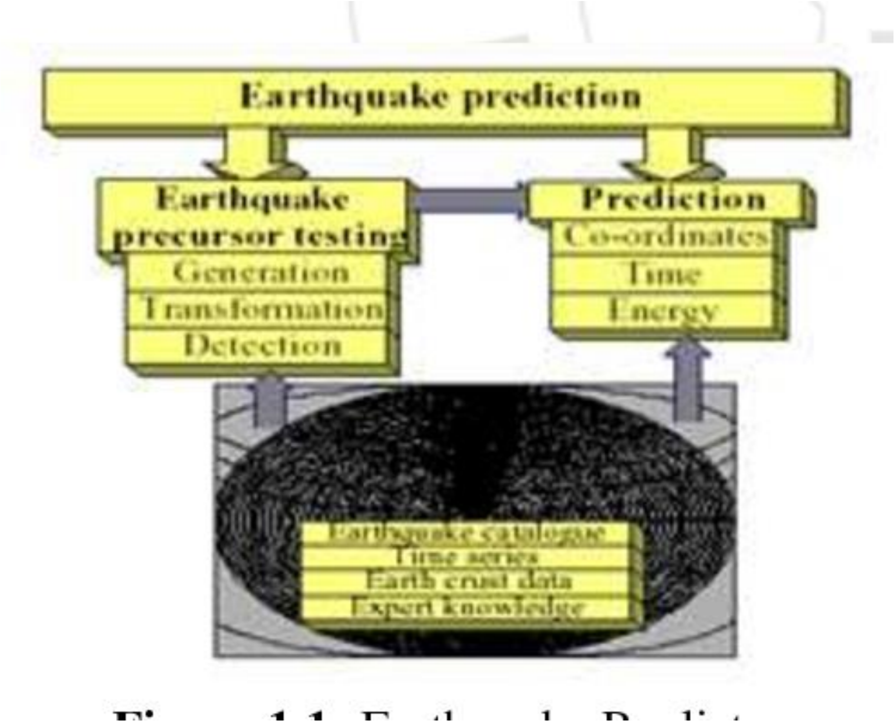

Fig 5.1 Earthquake prediction

(i) Ground water levels

(ii) Chemical changes in Ground water

(iii) Radon Gas in Ground water wells.

5.1 Ground Water Levels

Changing water levels in deep wells are recognized as precursor to earth quacks. The pre-seismic variations at observations wells are as follows.

1. A gradual lowering of water levels of period of months or years
2. An accelerated lowering of water levels in the final few months or weeks preceding the earth quake.
3. A rebound where water levels begin to increase rapidly in the last few days or hours before the main shock.

5.2 Chemical Changes in Ground water

1. The Chemical composition of ground water is affected by seismic events.
2. Researches at the university of tokyo tested the water after the earth quake, the result of the study showed that the composition of water changed significantly in the period around earth quake.
3. They observed that chloride concentration were almost constant.
4. Levels of sulfate also showed a similar rise.

5.3 Radon Gas in Ground water wells.

1. Increase levels of radon gas in wells is a precursor of earthquakes recognized by research group.
2. Although radon has a relatively short half 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 earthquake predictions.
3. The detection and effective earthquake procedures in using the precursors for estimation of time, place and energy for estimation of time, place and energy of expected earthquake.
4.There are no effective solution to the problem.
5.Earth quake catalogues, geo-monitoring time series data about stationery seismo-tectonic properties of geological environment and expert knowledge and hypotheses about earthquake precursors are used to solve this problem

CHAPTER 6

DATA MINING AND KNOWLEDGE EXTRACTION

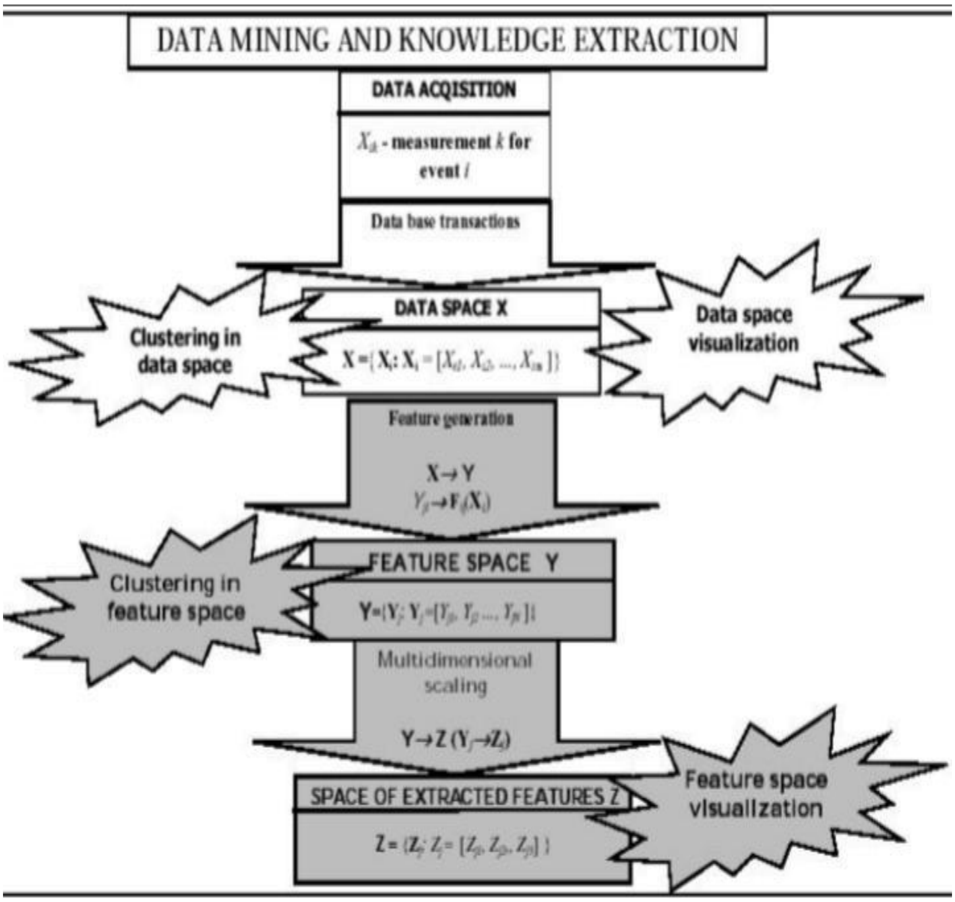
DATA MINING AND KNOWLEDGE EXTRACTION
DATA ACQISITION
X_{ik} - measurement k for event i
Data base transactions
DATA SPACE X
$X = \{X_i: X_i = [X_{i1}, X_{i2}, ..., X_{in}]\}$
Clustering in data space
Data space visualization
Feature generation
$X \rightarrow Y$
$Y_j \rightarrow F_j(X_i)$
FEATURE SPACE Y
$Y = \{Y_j: Y_j = [Y_{j1}, Y_{j2}, ..., Y_{jn}]\}$
Clustering in feature space
Multidimensional scaling
$Y \rightarrow Z\ (Y_j \rightarrow Z_j)$
SPACE OF EXTRACTED FEATURES Z
$Z = \{Z_j: Z_j = [Z_{j1}, Z_{j2}, Z_{j3}]\}$
Feature space visualization

Fig 6.1 Data mining and knowledge extraction

This paper proposes a novel multi-resolution approach, which combines local clustering techniques in the data space with a non-hierarchical clustering in the feature space. The raw data are represented by n-dimensional vectors Xi of measurements Xk. The data space can be searched for patterns and be visualized by using local or remote pattern recognition and 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 both the noise and other random components. The new features Yl form an N-dimensional feature space. We use multidimensional scaling procedures for visualizing the multi-dimensional events in 3-D space. This transformation allows for a visual inspection of the N-dimensional feature space. The visual analysis helps greatly in detecting subtle cluster structures, not recognized by classical clustering techniques, selecting the best pattern detection procedure used for data clustering, classifying the anonymous data and formulating new hypotheses.

Clustering schemes Clustering analysis is a mathematical concept whose main useful role is to extract the most similar (or dissimilar) separated sets of objects according to a given similarity (or dissimilarity) measure. This concept has been used for many years in pattern recognition. Nowadays clustering and other feature extraction algorithms are recognized as important tools for revealing coherent features in the earth sciences, bio-informatics and in data mining. Depending on the data structures and goals of classification, different clustering schemes must be applied.In our new approach we use two different classes of clustering algorithms for different resolution levels. In data space we use agglomerative schemes, such as modified mutual nearest neighbor algorithm (mnn). This type of clustering extracts better 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. The non-hierarchical clustering algorithms are used mainly for extracting compact clusters by using global knowledge about the data structure. We use improved k means based schemes, such as a suite of moving schemes, which uses the k-means procedure plus 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:

$$\mathbf{j(w, n)=\sum J \sum ieCj |xi\text{-}zj|^{2}}$$

where zj 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 many times the clustering procedures for 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, z) minimum is selected

CHAPTER 7

TYPES OF CLUSTERING

7.1 Hierarchical clustering methods

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 bottom-up 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 characterised 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 top-down 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 since display of the full hierarchy is not really appropriate for data sets of more than a few hundred compounds.

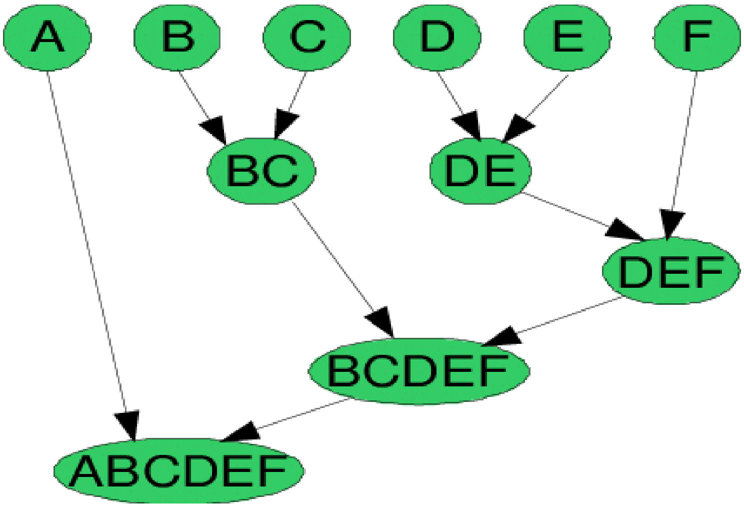

Fig 7.1 Hierarchical clustering

7.2 Non hierarchical clustering methods

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 thus been 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 methods (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 scompounds to a user-defined number of seed clusters and then iteratively reassign compounds to see if better clusters result. Such methods are prone to reaching local optima rather than a global optimum, and it is generally not possible to determine when or whether the global optimum solution has been reached; nearest neighbour methods, such as the Jarvis-Patrick method, assign compounds to the same cluster as some number of their nearest neighbours. 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, high-dimensional datasets such as those used in chemical applications.

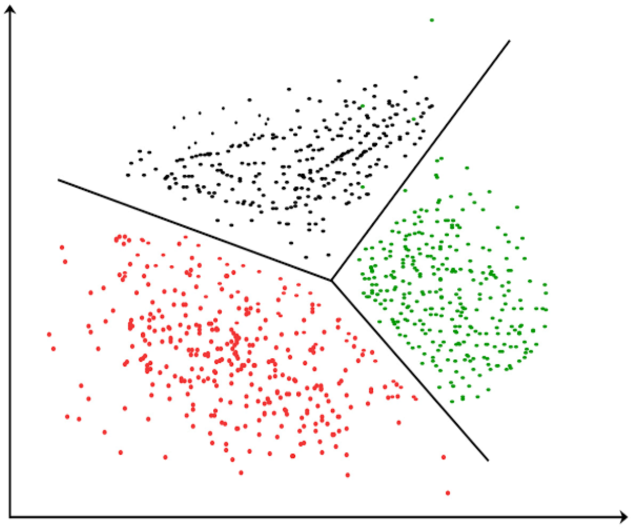

Fig 7.2 Non-Hierarchical clustering

CHAPTER 8

DATA MINING APPLICATIONS

- 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 detection etc.,
- In Finance – stock market perdition, credit assessment, fraud detection etc.
- Knowledge Acquisition

CHAPTER 9

FUTURE ENHANCEMENT

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, the emerging market for predictive analytics has been sustained by professional services, service bureaus and 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 behaviour and forecast product demand and related market dynamics. Finally, they are at different stages of growth in the life cycle of technology innovation.

CHAPTER 10

CONCLUSION

In earthquake prediction there are no simple conclusion which can be drawn. The problem of earthquake prediction is based on data extraction of precursory phenomena and it is highly challengingly task various computational methods and tools are used for detection of pre-cursor by extracting general information from noisy data.

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 noisy data.

By using common frame work of clustering we are able to perform multi-resolutional 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 eg. We can apply this clustering method to volcanic eruptions

CHAPTER 11

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. Fundamentals of Earthquake prediction [1994]

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

[4] C.Brunk, J.Kelly & Rkohai "Mineset An integrate system for data access, Visual Data Mining & Analytical Data Mining", proceeding of the 3rd conference on KDD 1997.

[5] Andenberg M.R.Cluster Analysis for application, New York, Academic, Press 1973.

[6] "Predicting the Earthquake using Bagging Method in Data Mining", S.Sathiyabama,K.Thyagarajah, D. Ayyamuthukumar

[7] "A Bagging Method using Decision Trees in the Role of Base Classifiers", Kristína Machová, František Barčák, Peter Bednár

[8] "Cluster Analysis, Data-Mining, Multi-dimensional Visualization of Earthquakes over Space, Time and Feature Space", Witold Dzwinel, David A. Yuen, Krzysztor Boryczko, Yehuda Ben-Zion, Shoichi Yoshioka, Takeo Ito

[9]http://cse.stanford.edu/class/sophomore-college/projects-00/neural-networks/Architecture/feedforward.html

[10]www.dmreview.com

[11]www.aaai.org/Press/Books/kargupta2.php

[12]www.forrester.com

[13]www.ftiweb.com