A SURVEY ON DATA MINING TECHNIQUES AND ITS APPLICATIONS

Tanu Minhas[1], Nancy Sehgal[2]
[1] Research Scholar,Department of Computer Science & Engineering,Baddi University of Emerging Sciences and Technology ,Himachal Pradesh,
[2] Astt. Professor,Department of Computer Science & Engineering,Baddi University of Emerging Sciences and Technology ,Himachal Pradesh,

ABSTRACT: Nowadays, a very huge volume of data is generated by medical field everyday at a rapid rate. This requires a need for new techniques and tools to analyze large data sets to gain knowledge. This growing need gives a perspective for a new research field called Knowledge Discovery in Databases (KDD) or Data Mining. It is the process of analyzing data to find interesting, and hidden patterns, trends, descriptive and predictive models from large amount of data. Various tasks are included in data mining process. This review paper analyses various clustering and classification techniques which are used to predict previously unknown class of objects and various data mining applications.

Keyword: - Data Mining, Clustering, Classification, Data Mining Applications

---

1. INTRODUCTION

Data Mining is known as the process of analyzing data to extract interesting patterns and knowledge from large amount of raw data such as data warehouse, data repositories and large data base etc. Data mining is used to discover knowledge out of data and presenting it in a form that is easily understood to human beings. Data mining is used for analysis purpose to analyze different type of data by using available data mining tools. Various techniques and methods are provided by the data mining process for the transformation of data into useful information used for decision making in future. This information is currently used for wide range of applications like customer retention, education system, production control, healthcare, market basket analysis, manufacturing engineering, scientific discovery and decision making etc [2]. Data Mining is also known as KDD (Knowledge discovery from the database) process. The main goal of the KDD process is to extract knowledge from data in the context of large databases.

1.1 KDD PROCESS

The Discovery of knowledge in Databases process includes the following steps to gain unique knowledge. These are:

1. In the first phase cleansing of data is done which comes under data Pre-Processing. Data which is not relevant contain noise and inconsistent value is removed.
2. In the Data Integration phase, data from heterogeneous sources is combined.
3. In the Selection phase, applicability of analyzed data is taken into consideration.
4. In the Transformation step, data is transformed into forms appropriate for mining with respect to various mining techniques.
5. Under Data Mining phase, intelligent methods are applied in order to extract data patterns.
6. Within severely unique patterns, which includes acquaintance are recognized. This step involves, evaluating the required patterns.
7. In this last phase the exposed results, which includes knowledge are represented.

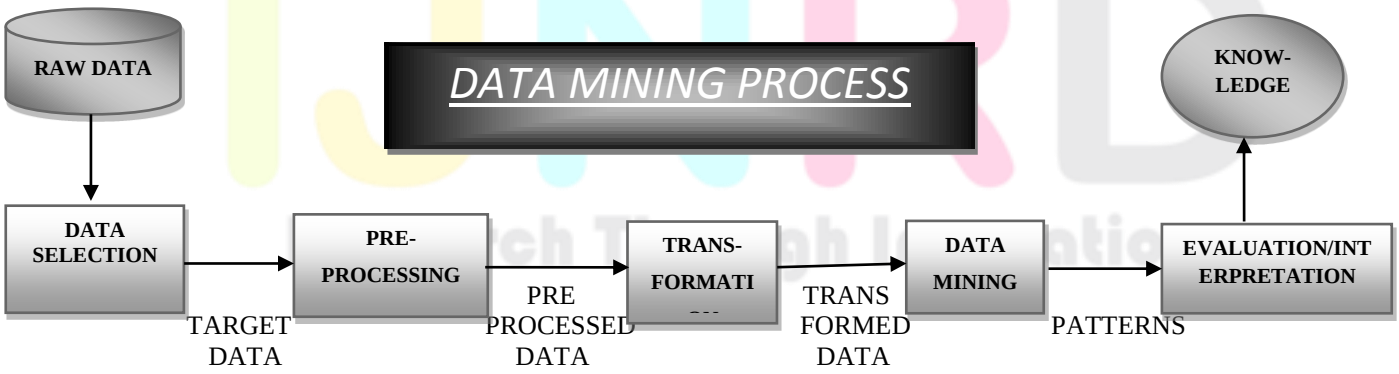

Fig 1-: Data mining as a Core process in KDD

1.2 DATA MINING LEARNING APPROCHES

Data mining uses the two major functions to mine the data.These functions are: supervised learning and unsupervised learning [6].

A. Supervised learning:

Supervised learning is also called directed or predictive data mining. In this approach variables under investigation can be divided into two groups: input objects and desired output data sets. To proceed with directed data mining techniques the values of the output data sets are provided which are used to train the machine and get the desired outputs. Classification comes under supervised data mining.

B. Unsupervised learning:

Unsupervised learning is also called undirected or descriptive data mining. In this approach all variables are treated in same way, there is no distinction between input objects and desired output data sets. The main target of unsupervised data mining is to find hidden structures and relation among data. Clustering comes under the process of unsupervised data mining technique.

2. TECHNIQUES OF DATA MINING

Data Mining uses several core techniques to describe the type of mining and data recovery operation.

2.1 Clustering

Clustering is the process in which grouping of data objects is done in such a way that objects in the same group (cluster)are more similar to each other than to those in other groups (clusters). Clustering mainly focus on inter cluster property rather than intra cluster properties. Clustering does not rely on predefined classes and training. Unsupervised clustering is different from pattern reorganization in the area of statistics known as discriminate analysis and decision analysis which classify the objects from a given set of object [7]. Data clustering [9], is an unsupervised learning method aims at creating the groups of data objects or clusters, in a form that objects in the same cluster are highly similar and objects in different clusters are quite dissimilar. Cluster analysis is one of the traditional topics in the data mining field. It is the first step in the direction of exciting knowledge discovery. Clustering is the procedure of grouping data objects into a set of disjoint classes, called clusters. Now objects within a class have high resemblance to each other in the meantime objects in separate classes are more unlike.Cluster analysis [10] has been widely used in several applications, including market research, pattern recognition, data analysis, machine learning, image processing, information retrieval, data compression and computer graphics.Clustering is a method used to group similar data objects, but it differs from categorization of data objects are clustered on the fly instead of through the use of predefined topics. There are many clustering algorithms used for clustering. The major fundamental clustering methods can be classified into following categories [8]:

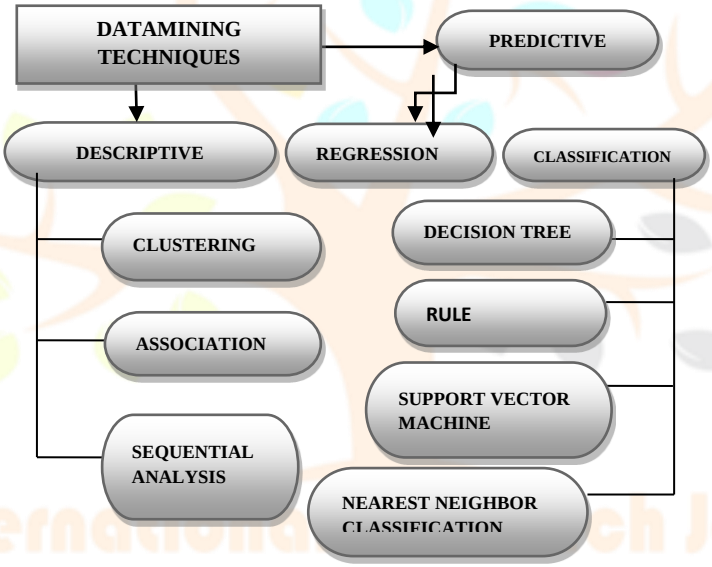

Fig -2: Techniques of Data Mining

2.1.1 Clustering Methods

Partitioning Method: - The general approach for partitioning method includes high inter cluster property and low intra cluster property i.e. data items within a cluster is highly similar and dissimilar with distinct clusters. Most partitioning methods are distance-based. In this construct a partition of a data set containing n objects into a set of k clusters, so to minimize a criterionθ. Here k is a input parameters. E.g. K-mean and K-centriod [11].

K-Mean Clustering: - The k-means clustering algorithm is the basic algorithm which is based on partitioning method and is used for many clustering tasks. It uses k as a parameter, divide n objects into k clusters so that the objects in the same cluster are similar to each other but dissimilar to other objects in other clusters. The aim of K-Means clustering is to reduce total intra-cluster variance, or, the squared error function [2].

Hierarchical Methods:-In this method data objects are decomposed into hierarchical fashion i.e. levelling of data is done in order to form the cluster. It can be classified as being either Agglomerative or Divisive based on which hierarchical decomposition is formed. Agglomerative approach is the bottom up approach where in the initial step each data object is considered as a separate group (cluster). It then merges groups close to one another until all the groups are merged into one. On the other hand divisive approach is the top down approach where in the initial step all data objects are considered as a same group (cluster). It then start splitting of the large cluster into smaller one, and continue this process until each data object is clustered as a separate group. It is the reverse process of agglomerative approach. It is rigid in nature.

Density Based Methods:-Most partitioning methods cluster objects based on distance between objects. Spherical shaped clusters can be discovered by these methods and encounter difficulty in discovering clusters of arbitrary shapes. So for arbitrary shapes new methods are used known as density-based methods which are based on the notion of density. It helps to discover arbitrary shape clusters. It also handles noise in the data. It is one time scan. It requires density parameters also [13].

Grid Based Methods:-In this method data objects are represented as a grid. Here object space is quantized into a finite number of cells that form a grid structure. This method possesses fast processing time and is independent of the number of data objects. It depends only on the number of cells in each dimension in the quantized space. In grid based methods objects together form grid [11].

2.2 Classification

Classification is a supervised learning method of data mining process. It is a process of data organization into predefined classes. Classification is a data analysis task, where to predict the categories or class labels a model or classifier is constructed. Classification is a two phase process. In the first phase, by analyzing the data tuples from training data having a set of attributes a model is built. The value of class label attribute is known for each tuple in the training data. On training data Classification algorithm is applied to create the model or classifier. In the second phase of classification accuracy of the classifier or model is checked and to check this accuracy test data is used. If the accuracy of the classification is acceptable then it can be used to classify the unknown data objects into labeled classes. These classified models can be used as predictor. The major task of classification is to build a classifier model that can be applied to unclassified data sets in order to classify and predict the future results based on this classification. Machine learning algorithm uses classification techniques as an important component in order to extract rules and patterns from data that could be used for prediction. Classification techniques are used to classify the unclassified data records into one among a set of predefined classes. Examples of classification include:

- Classification of patient's disease based on their symptom and some important test value.
- Classification of students based on their CGPA.
- Classification of home telephone lines which are used for internet connection.

2.2.1 Classification Methods

Decision Tree: - It is a framework that consist a root node, branches, and leaf nodes. The upmost node in the tree is the root node and each leaf node holds a class label. Each internal node indicates a test on an attribute; each branch indicates the result of a test. It does not require any field insight and easy to understand.

Rule Induction: - Rule based classification algorithm also known as divide-and-conquer approach. This technique is a repetitive process which works in two steps, in first step it generates a rule that cover up a subset of the training examples and in second step it removes all examples covered by the rule from the training set. This process is repeated until all examples are covered [15]. It basically aims at understanding data structure and providing understandable explanation rather than only black box prediction.

Support Vector Machine: - An SVM is a supervised machine learning algorithm that analyzes the data for both classification and regression. SVMs are more commonly used in classification problems. SVMs are based on the idea of finding a hyper plane that best divides a dataset into two classes. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on on which side of the gap they fall.

Nearest Neighbor Classification: - The k-nearest neighbor's algorithm (k-NN) is a technique to classify objects based on closest training examples in the feature space. It is an instance-based learning, or lazy learning which can also be used for regression. The k-NN algorithm is simplest among all machine-learning algorithms. By locations and labels of the training samples the object space is divided into segments. A class c is assigned to the point in the space only if it is the most frequent class label among the k nearest training samples. Normally Euclidean distance is used as the distance metric; however this will only work with numerical values. In case of text classification another metric, such as the overlap metric (or Hamming distance) can be used.

3. LITERATURE REVIEW

K.Rajalakshmi, Dr. S S Dhenakaran and N. Roobini (2015) represents a paper on "Comparative Analysis of K-Means Algorithm in Disease Prediction". In this paper [2] they explained that various analytical tools are used to handle large amount of data generated by medical field. Data mining is also required to turn this data into knowledge. In this paper K-means algorithm is used to analyze various disease predictions techniques. This prediction system reduces the human effects and cost effective one.

Sanjay Chakraborty, Prof. N.K Nigwani and Lop Dey (2014) represented a paper on "Weather Forecasting using Incremental K-means Clustering". In this paper[13] generic approach of incremental K-mean clustering is considered for weather forecasting. This paper generally uses typical K-means clustering on the main air pollution database and a list of weather category will be developed based on the peak mean values of the clusters. Whenever new data are coming, the incremental K-means is used to club data into those clusters where weather category has been previously defined. Thus it is able to predict weather information of future.

Akhilesh Kumar Yadav, Divya Tomar, Sonali Agarwal (2013), represents a paper on "Clustering of Lung Cancer Data Using Foggy K-Means". The aim of this paper[12] is to evolve a new approach based upon foggy k-mean clustering. The result of the experiment illustrate that foggy k-means clustering algorithm has outstanding result on datasets which are real as compared to simple k-means clustering algorithm and provides a enhanced result to the real world problem.

Daljit Kaur and Kiran Jyot, (2013) represents a paper on "Enhancement in the performance of K-means Algorithm". This paper[14] proposes a method to make the algorithm more adequate and potent. The proposed technique reduces the complexity and deed of numerical calculation but it maintains the ability of implementing k-means algorithm. It also unfolds the problem of dead unit.

Bala Sundar V, T Devi and N Saravan, (2012) presented a paper on "Development of a Data Clustering Algorithm for Predicting Heart". In this paper [7] they examined the conclusion of the accuracy of the result by using k-mean clustering technique in prediction of heart disease diagnosis with real and artificial datasets. The research output shows that the integration of clustering gives promising results with maximum precision rate and robustness.

Chew Li Sa, Bt Abang Ibrahim D.H, Dahliana Hossain E.H, and Bin Hossin, M.,(2013) presented a paper on "Student Performance Analysis System".In this paper[10] they proposed a system named Student Performance Analysis System (SPAS) to keep track of student's result in a particular university. The proposed project offers a system which predicts performance of the students on the basis of their result on the basis of analysis and design. The planned scheme offers student performance prediction through the rules generated via data mining technique. The mining approach used in this proposal is classification, which classifies the students based on their grade.

4. DATA MINING APPLICATIONS

Future healthcare

Data mining holds high possibility to enhance healthcare systems. It uses data and analytics to determine optimum practices that enhance care and decrease costs. Researchers use data mining techniques like multi-dimensional databases, machine learning, soft computing, data visualization and statistics. To predict the volume of patients in every category data mining can be used. Processes are established to make sure that the patients receive appropriate care at the right place and at the right time.

Education

Educational Data Mining is a new emerging field which concerns with developing techniques that discover knowledge from huge raw data rising from educational environments. The objectives of EDM are identified as predicting students' future learning behaviour, studying the effects of educational support, and advancing scientific knowledge about learning. Data mining can be used by an educational institution to take exact decisions and also to predict the student's result. With these outcomes the institution can concentrate on what to teach and how to teach. Learning pattern of the students can be captured and used to develop methods to teach them.

Fraud detection

Due to the action of frauds billions of currency has been vanished. Traditional approach of fraud detection is time consuming and complicated. Data mining helps in providing meaningful patterns and relevant information. Any information that is valid and useful is knowledge. A perfect fraud detection system should preserve information of all the users. A supervised approach consists of collection of sample records. These records are further classified as fraudulent or non-fraudulent. A model is built using this data and the algorithm is made to determine whether the record is fraudulent or not.

Intrusion detection

Any process that will deal the integrity and confidentiality of a resource is an intrusion. The preventive measures to bypass an intrusion comprise user authentication, escape programming errors, and information protection. Data mining can help to enhance intrusion detection by adding a level of target to anomaly detection. It helps an analyst to differentiate an activity from common everyday network activity. Data mining also helps to extract data which is more related to the problem.

Financial Banking

A large amount of data is assumed to be produced with new transactions with the advent of computerised banking everywhere. Data mining can commit to resolve business problems in banking and finance by finding patterns, causalities, correlations and by using various analytical tools in business knowledge and market prices that are not instantly possible to managers because the data generated too quickly to screen by experts is in bulk. The managers may find these information for better segmenting, targeting, acquiring, retaining and maintaining a profitable customer.

Research Analysis

Antiquity demonstrates that we have sub stained radical modification in research. Data mining is helpful in data cleaning, data pre-processing, integration, transformation and evolution of databases. The researchers can find any similar data from the database that might bring any modification in the research. Recognition of any co-occurring sequences and the correlation between any activities can be known. Data visualisation and visual data mining provide us with a clear view of the data.

5. CONCLUSION

Data mining has great concern about finding the patterns, prediction, knowledge discovery etc., in distinct business domains. Data mining approaches such as classification, clustering etc., helps in finding the patterns and knowledge to agree upon the future trends in businesses to expand. Data mining has wide application area almost in every industry where the data is generated in bulk. In future we wish to propose improvement in k-mean clustering algorithm that is defined in this paper and the improved algorithm will be implemented using Matlab.

6. REFERENCES

[1] B V Sumana,T.Santhanam,"Prediction of Diseases by Cascading Clustering and classification", 2014 International Conference on advances in Electronics Computer and Communication.

[2] K.Rajalakshmi, Dr.S.S.Dhenakaran, N.Roobin "Comparative Analysis of K-Means Algorithm in Disease Prediction", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 4, Issue 7, July 2015.

[3] Madhu Yedla, T M Srinivasa, "Enhancing K-means Clustering Algorithm with Improved Initial Center", International Journal of Computer Science and Information Technologies, Vol. 1 (2) 2010, page 121-125.

[4] Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khusro and Huma Javed, "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity," Middle-East Journal of Scientific Research, pages 959-963, 2012.

[5] K. A. Abdul Nazeer, M. P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm, Proceedings of the World Congress on Engineering , Vol IWCE 2009, July 1 - 3, 2009, London, U.K.

[6] K.Kameshwaran, K.Malarvizhi, "Survey on Clustering Techniques in Data Mining", IJCSIT, Vol. 5, 2014, 2272-2276.

[7] Bala Sundar V,T Devi, N Saravan, "Development of a Data Clustering Algorithm for Predicting Heart", International Journal of Computer Applications (0975 – 888) Volume 48– No.7, June 2012.

[8] Qasem A. Al-Radaideh, Adel Abu Assaf 3eman Alnagi, " Predictiong Stock Prices Using Data Mining Techniques", The International Arab Conference on Information Technology (ACIT'2013).

[9] Abdelghani Bellaachia, Erhan Guven, "Predicting Breast Cancer Survivability Using Data Mining Techniques", Washington DC 20052, 2010.

[10] Chew Li Sa; Bt Abang Ibrahim, D.H.; Dahliana Hossain, E.; bin Hossin, M., "Student performance analysis system (SPAS)," in Information and Communication Technology for The Muslim World (ICT4M), 2014 The 5th International Conference on , vol., no., pp.1-6, 17-18 Nov. 2014.

[11] Oyelade, O. J, Oladipupo, O. O and Obagbuwa, I. C, "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", International Journal of Computer Science and Information Security, Vol. 7, o. 1, 2010.
[12] Sanjay Chakrabotry, Prof. N.K Nigwani and Lop Dey "Weather Forecasting using Incremental K-mean Clustering", 2014.
[13] Akhilesh Kumar Yadav, Divya Tomar, Sonali Agarwal, "Clustering of Lung Cancer Data Using Foggy K-Means", International Conference on Recent Trends in Information Technology (ICRTIT) 2013.
[14] Daljit Kaur and Kiran Jyot, "Enhancement in the Performance of K- means Algorithm", International Journal of Computer Science and Communication Engineering, Volume 2 Issue 1, 2013.
[15] Phyu, Thair Nu. "Survey of classification techniques in data mining."International MultiConference of Engineers and Computer Scientists, 2009.