Volume 8, No. 4, May 2017 (Special Issue)
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
REVIEW ARTICLE
Available Online at www.ijarcs.info

A Survey on Data Mining and Its Current Research Directions

Er. Anita Devi
Research Scholar, Department of CSE
SBBSUIET, Padhiana, PunjabS
neetusalhan207@yahoo.com

Er. Jasjeet Kaur
AssistantProfecerDepartment of CSE
BBSUIETPadhiana, Punjab
jparmar21nice@gmail.com

***Abstract*:** Data mining is extracts the knowledge from a large amount of data which stores in multiple heterogeneous data base. Knowledge is converting the message through direct or indirect.A survey of various mining techniques. These techniques are including association, correlation, clustering and neural network. It is also conducts a formal review of the application of data mining such as the education sector, marketing, fraud detection, manufacturing and telecommunication. Data mining are based on the data models and clusters. These are divided for the different paths for research the information.

Keywords: Data, Mining, knowledge, Database, cluster, regression, technique.

I. INTRODUCTION

Nowaday's databases are comprised of terabytes or more data in it. As they are able to accommodate huge mass of hetero generous data, different variety of strategic information lies hidden inside it. So, through effective data mining only we can able to draw meaningful conclusions which are the basic purpose of data mining. As data arebeing accumulated continuously as well as rapidly whether it is a research field, education, and market products, medical science, electronic information, media, entertainment etc. It is difficult to get faster and appropriate information by traditional manual analysis which is tedious as well as very cumbersome. So, data mining is used basically i) to reduce costs through proper detection and prevention of waste and fraud, ii) obtaining appropriate and up-to-date information and iii) increase revenues through improved marketing strategy. We can take an example of research of finding presence of water in the planet Mars. Scientists receive different data from Mars through satellites. Those data are changing time to time as the satellite provides new sets of data in different time duration. The job is very challenging as well as required effective research analysis. So through an effective data mining analysis scientists can able to find the outcome more scientifically which is not that possible or easier in Traditionalanalysis. So data mining finds patterns and relationships in datawhich are stored in databases, data warehouses, or other information repositories providing more advanced andeffective information which is augmented by utilizing models equipped with sophisticated techniques. In other way, Data Mining allow to [1] explore massive data in such a way that its end result is obtaining of fruitfulknowledge information multidisciplinary field,drawing work from various areas like i) Artificial intelligence, ii) Neural networks, iii) Machine learning, iv)Database technology, v) statisticsvi)pattern recognition,vii) signal processing, viii) spatial data analysis, ix) Business, x) Economics, xi) Bio-informatics etc. Data mining can be done in different types of data but one of the interesting facts is that it can process spatial data which is used for geographical, chip design, medical and satellite image databases.

Data mining is also known as [2] **KDD i.e.** Knowledge Discovery in databases. KDD is the automated extraction of novel, understandable and potentially useful patterns implicitly stored in large databases, data warehouses and other massive information repositories comprising textual, Numerical,graphical, spatial data.

A. *Pre mining tasks*

1) Data Cleaningbasically emphasizes on removing noisy and inconsistent data.
2) Data Integrationaccumulates data from various sources to a single location as well as into a common format.

B. *Post mining tasks*

1) Pattern Evaluationfocuses on identifying the truly interesting patterns that represents knowledge.
2) Knowledge Presentationprovides discovered rules using visualization and different knowledge representationtechniques

II. BASIC TAXONOMY OF DATA MINING MODELS

Generally two types of models can be classified or built in data mining:-

Predictive models**:**These models predict or forecastexplicit value of a particular attribute i.e. when the modelpredicts according to class membership it is call asclassification model or simply classifier. If the modelpredicts a number

from a wide range of possible values,then it is referred as regression model

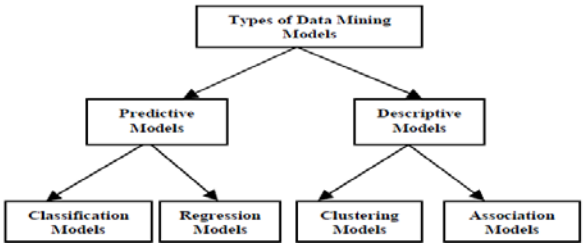

Fig 1.Taxonomoy of data mining model

Descriptive models**:**These models describe patterns in existing data and basically used to create meaningful subgroups. When it clumps together similar things, events or people into groups are called clusters which reduce data complexity. When the descriptive model engross determinations of likeness i.e. how frequently two or more things associate together at that time the model is known asAssociation models.

III. CLASSIFICATION OF DATA MINING SYSTEMS

Classification of data mining systems is performed according to types of databases are mined i.e. relational, transaction, spatial databases etc., what types of knowledge comprised of mining functionalities like association, customization, clustering etc. are applied to it, what types of data mining techniques are being deployed like machineLearning ,neural networks, pattern recognition etc.

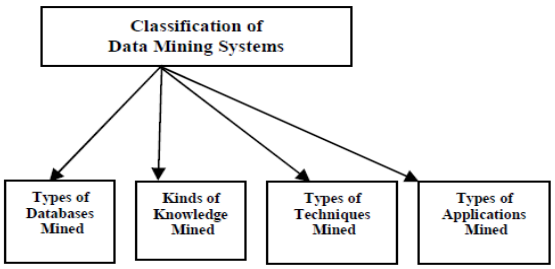

Fig 2. Classification of data mining system

IV. RELATED WORK

Though varieties of data mining techniques, like association rule mining, cluster analysis etc. are being used so often in different applications, in this section we have specified most of the work related to association rule mining and others. AIS Algorithm was the first association rule mining algorithm which was proposed in [3] where generation of candidate itemsets is done at the time of scanning the database through extending known-frequent itemsets with items from each transaction. It is a multipassalgorithm. An estimate of the supports of these candidates is used to steerwhetherthese candidates are needed to be extended further for production of more candidates.The chief drawback of this algorithm is that it generates too many candidates which are small and requires much more space and is thereby inefficient.

Apriority algorithm which is introduced in [4] that is more efficient than AIS by an order of magnitude. The foremost advantage of Apriority is that it incorporates the subset frequency based pruning optimization that means, it only process any itemset whose subsets are frequent also. It utilizes a data structure that is known as hashtree which is used for storing the counters of candidate itemsets. The main drawbacks of Apriori are I) it performs n passes overthe database, where n is the length of the longest frequent itemset. The counts of candidate itemsets of length k are obtained in kith pass, ii) it follows a tuple-by-tuple approach where counters of candidate itemsets are updated after reading in individual transaction of whole database so that much redundant work are performed after each and individual transaction. Based on this algorithm, lots of newalgorithms are deliberated with enhancements ormodifications.

An algorithm called DIC Algorithm introduced in [5] where candidates are generated as well as removed afterevery m transactions where M is used as a parameter in thealgorithm. The algorithm utilizes multi-pass technique buttypically it can be completed within two passes. Thedrawback of this is that, it also follows a tuple-by-tupleapproach.

V. CURRENT RESEARCH DIRECTIONS

In[6].A novel approach is specified to build data mining models from perturbed data which preserves privacy of the data mining. In this paper, before the data is sent to the data miner, random noise from a known distribution is added to the privacy sensitive data. Consequently an approximation to the original data distribution from the perturbed data isreconstructed by the data miner and reconstructedDistribution is used for data mining purposes. As noise is added here, loss of information versus preservation of privacy both the concepts are made trade off in the perturbation based approaches. The proposed model is based on an individually adaptable perturbation model, which enables the individuals to choose their own privacy levels. Inthe focus is laid on mining non derivable frequent itemsets in an incremental fashion. In their approach researchers have designed a compact data structure NDFITwhich can able to maintain a dynamically selected set for itemsets more efficiently. An optimized algorithm named NDFLoDs is proposed for generating non-derivable frequent itemsets over stream sliding window and the method has an edge over previous approaches. In a prototype multidimensional association mining system is proposed where users can build effective data mining models through the intelligence support of the ontology's that can prevent ineffective pattern generation, discover concept extended rules, and provide an active knowledge re-discovering mechanism. For improving thewarehouse mining process more effective, in theproposed technique, schema ontology, schema constraint ontology, domain ontology and user preference ontology is incorporated also for making the model more perfect oneIn [7], a system model is proposed

CONFERENCE PAPERS
National Conference on Emerging Trends on Engineering & Technology (ETET-2017)
On 21st April 2017
University Inst. of Engg. & Tech. & University Inst. of Computer, SBBS University, Punjab (India)

where the personal route of a user is predicted using a probabilistic model built from the historical trajectory data. A novel mining algorithm called CRPM (Continuous Route Pattern Mining) is proposed here through which route patterns are extracted from personal trajectory data. The advantage of this approach is that it can withstand different kinds of disturbance in trajectory data. Apart from that a clientserver architecture is employed which augments both the privacy of personal data to a certain extent and reducing the computational load on mobile devices also.

VI. Data Mining Techniques

Data mining means collecting relevant information from unstructured data. So it is able to help achieve specific objectives. The purpose of a data mining effort is normally either to create a descriptive model or a predictive model .A descriptive model presents, in concise form, the main characteristics of the data set. The purpose of a predictive model is to allow the data miner to predict an unknown (often future) value of a specific variable; the target variable [7]. The goal of predictive and descriptive model can be achieved using a variety of data mining techniques as shown in figure 2[8].

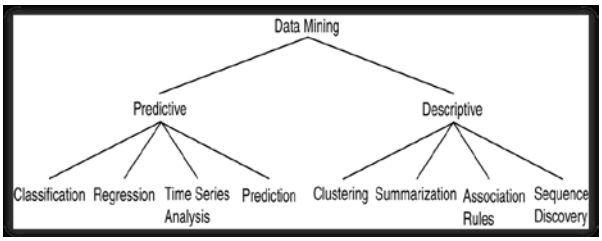

Fig. 3 Data Mining technique

1) *Classification*: Classification based on categorical (i.e. discrete, unordered). This technique based on the supervised learning (i.e. desired output for a given input is known) .It can be classifying the data based on the training set and values (class label). These goals are achieve using a decision tree, neural network and classification rule (IF- Then).for example we can apply the classification rule on the past record of the student who left for university and evaluate them. Using these techniques we can easily identify the performance of the student.
2) *Regression***:** Regression is used to map a data item to a real valued prediction variable [8]. In other words, regression can be adapted for prediction. In the regression techniques target value are known. For example, you can predict the child behavior based on family history.
3) *Time Series Analysis:* Time series analysis is the process of using statistical techniques to model and explain a time-dependent series of data points. Time series forecasting is a method of using a model to generate predictions (forecasts) for future events based on known past events . For example stock market.
4) *Prediction:* It is one of a data mining techniques that discover the relationship between independent variables and the relationship between dependent and independent variables [9].Prediction model based on continuous or ordered value.
5) *Clustering:* Clustering is a collection of similar data object. Dissimilar object is another cluster. It is way finding similarities between data according to their characteristic. This technique based on the unsupervised learning (i.e. desired output for a given input is not known). For example, image processing, pattern recognition, city planning.
6) *Summarization:*Summarization is abstraction of data. It is set of relevant task and gives an overview of data. For example, long distance race can be summarized total minutes, seconds and height.
7) *Association Rule***:** Association is the most popular data mining techniques and fined most frequent item set. Association strives to discover patterns in data which are based upon relationships between items in the same transaction. Because of its nature, association is sometimes referred to as "relation technique". This method of data mining is utilized within the market based analysis in order to identify a set, or sets of products that consumers often purchase at the same time [10].
8) *Sequence Discovery:*Uncovers relationships among data [11]. It is set of object each associated with its own timeline of events. For example, scientific experiment, natural disaster and analysis of DNA sequence.

VII. DATA MINING APPLICATION

Various field adapted data mining technologies because of fast access of data and valuable information from a large amount of data. Data mining application area includes marketing, telecommunication, fraud detection, finance, and education sector, medical and so on. Some of the main applications listed below:

1) *Data Mining in Education Sector:* We are applying data mining in education sector then new emerging field called "Education Data Mining". Using these term enhances the performance of student, drop out student, student behavior, which subject selected in the course. Data mining in higher education is a recent research field and this area of research is gaining popularity because of its potentials to educational institutes. Use student's data to analyze their learning behavior to predict the results [12].

2) *Data Mining in Banking and Finance:*Data mining has been used extensively in the banking and financial markets [13]. In the banking field, data mining is used to predict credit card fraud, to estimate risk, to analyze the trend and profitability. In the financial markets, data mining technique such as neural networks used in stock forecasting, price prediction and so on.

3) *Data Mining in Market Basket Analysis:* These methodologies based on shopping database. The ultimate goal of market basket analysis is finding the products that customers frequently purchase together. The stores can use this information by putting these

products in close proximity of each other and making them more visible and accessible for customers at the time of shopping [14].

4) *D*ata Mining in Earthquake Prediction:Predict the earthquake from the satellite maps. Earthquake is the sudden movement of the Earth's crust caused by the abrupt release of stress accumulated along a geologic fault in the interior. There are two basic categories of earthquake predictions: forecasts (months to years in advance) and short-term predictions (hours or days in advance) [15].

5) *DataMining in Telecommunication*:Thetelecommunications field implement data mining technology because of telecommunication industry have the large amounts of data and have a very large customer, and rapidly changing and highlycompetitiveenvironment.Telecommunication companies uses data mining technique to improve their marketing efforts, detection of fraud, and better management of telecommunication networks.

6) *Data Mining in Agriculture:*Data mining than emerging in agriculture field for crop yield analysis a with respect to four parameters namely year, rainfall, production and area of sowing. Yield prediction is a very important agricultural problem that remains to be solved based on the available data. The yield prediction problem can be solved by employing Data Mining techniques such as K Means, K nearest neighbor (KNN), Artificial Neural Network and support vector machine (SVM) .

7) *Data Mining in Cloud Computing*: Data Mining techniques are used in cloud computing. The implementation of data mining techniques through Cloud computing will allow the users to retrieve meaningful information from virtually integrated data warehouse that reduces the costs of infrastructure and storage .Cloud computing uses the Internet services that rely on clouds of servers to handle tasks The data mining technique in Cloud Computing to perform efficient, reliable and secure services for their users.

VIII. Data Mining Algorithms and Techniques

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbour method etc., are used for knowledge discovery from databases.

A. Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applicationsare particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

Types of classification models*:*

- Partitioning Methods
- HierarchicalAgglomerative (divisive) methods
- Density based method
- Grid-based methods
- Model-based methods

B. Clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern andA correlation among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as reprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality.

Types of clustering methods

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

IX. METHODOLOGIES OF DATA MINING

A. Neural Network

Neural Network or an artificial neural network is a biological system that detects patterns and makes predictions. The greatest breakthroughs in neural network in recent years are in theirapplication to real world problems like customer response prediction, fraud detection etc. Data mining techniques such as neural networks are able to model the relationships that exist in data collections and can therefore be used for increasing business intelligence across a variety of business app locations. Neural Networks are used in a variety of applications. Artificial neural network have become a powerful tool in tasks like pattern recognition, decision problem or predication applications. It is one of the newest signals processing technology. ANN is an adaptive, non linear system that learns to perform a function from data and that adaptive phase is normally training phase where system parameter is change during operations. After the training is complete the parameter are fixed. If there are lots of data and problem is poorly understandable then using

CONFERENCE PAPERS
National Conference on Emerging Trends on Engineering & Technology (ETET-2017)
On 21st April 2017
University Inst. of Engg. & Tech. & University Inst. of Computer, SBBS University, Punjab (India)

ANN model isAccurate,thenon linear characteristics of ANN provide it lots of flexibility to achieve input output map.

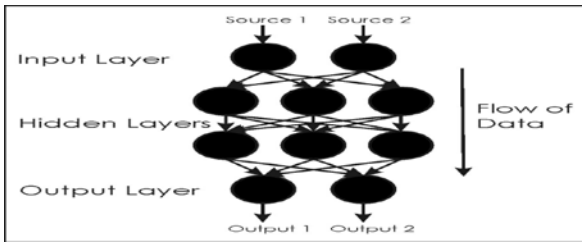

Fig: 4 Neural Network with hidden layers

B. Decision Trees

A decision tree is a flow chart like structure where each node denotes a test on an attribute value, each branch represents anOutcome of the test and tree leaves represent classes or class distribution. A decision tree is a predictivemodel most often used for classification. Decision trees partition the input space into cells where each cell belongs to one class. The partitioningis represented as a sequence of tests. Each interior node in the decision tree tests the value of some input variable, and thebranches from the node are labelled with the possible results of the test. The leaf nodes represent the cells and specify the classto return if that leaf node is reached. The classification of a specific input instance is thus performed by starting at the rootnode and, depending on the results of the tests, following the appropriate branches until a leaf node is reached .Decisiontree is represented in figure 2.

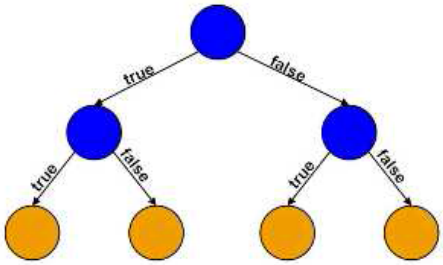

Fig 5 Decision tree

Decision tree is a predictive model that can be viewed as a treewhere each branch of the tree is a classification question andleaves represent the partition of the data set with theirclassification. The author defines a Decision Tree as aschematic tree-shaped diagram used to determine a course of action or show a statistical probability. Decision trees can be viewed from the business perspective as creating a segmentation of the original data set. Thus marketing managersmake use of segmentation of customers, products and salesregion for predictive study. These predictive segments derived from the decision tree also come with a description of the characteristics that define the predictive segment. Because of their tree structure and skill to easily generate rules the method is a favoured technique for building understandable models.

CONCLUSION

Data mining can be concluded that data mining is a very demanding and most sought after area now a days. Data mining enhance understanding by showing which factors most affect specific outcome. For any development analysis purpose effective study of data provides an effective outcome which is possible through perfect data mining. In today's advanced world, by implementing various advanced data mining techniques. We can also to obtain effective data miningoutputs provides immenseknowledge through which a systemworks perfectly and reasonability according to their own requirement and extreme satisfaction. Not only that,it provide definite and appropriate results thought which more and more appropriate and valuable information can be found out. In turn, one can able to explore in a wide range of application in a varieties way of true aspects.

REFERENCES

[1]. V. Pudi and J. Haritsa, "Quantifying the utility of the past in mining large databases Informationsystems,"

[2]. J. Gennari, P. Langley, and D. Fisher, "Models of incremental concept formation,"artificiall intelligence

[3]. Agrawal, R., Imielinski, T., and Swami, A. N. 1993. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management

[4]. HToivonen sampling large databases for association rules. In Proc. Of Intl. Conf. on very Large Databases ((VLDB)

[5]. S. Brin, R. Motwani, J. Ulman, and S. Tsur, Dynamic item set counting implication rules for market basket data. In Proc. Of ACM SIGMOD Intl. Conf. on Management

[6]. JoseZubcoff, Juan Trujillo, A UML 2.0 profile to designAssociation Rule mining models in the multidimensionalconceptual modeling of data warehouses, Data &Knowledge Engineering

[7]. George Gigli, ÉloiBossé, George A. Lampropoulos, An optimized architecture for classification combining data

[8]. HIANCHYE KOH , School of Business, SIM University, Singapore

[9]. Sirgo, J., Lopez, A., Janez, R., Blanco, R., Abajo, N., Tarrio, M., Perez, R., "A Data Mining Engine based on Internet, Emerging Technologies and Factory Automation," Proceedings ETFA '03, IEEEz

[10]..Bianca V. D.,PhilippeBoula de Mareüil and Martine Adda-Decker, "Identification of foreign-accented French using data mining techniques, Computer Sciences Laboratory for Mechanics and Engineering Sciences(LIMSI)".Websitewww.limsi.fr/Individu/bianca/article/Vieru&Boula&Madda_ParaLing07.pdf

[11].Bianca V. D.,PhilippeBoula de Mareüil and Martine Adda-Decker, "Identification of foreign-accentedFrench using data mining techniques, Computer Sciences Laboratory for Mechanics and Engineering Sciences(LIMSI)".Websitewww.limsi.fr/Individu/bianca/article/Vieru&Boula&Madda_ParaLing07.pdf

[12].R. Andrews, J. Diederich, A. B. Tickle," A survey and critique of techniques for extracting rules from trained artificial neural networks", Knowledge-BasedSystems.

[13].LoirRoach and OdeMaimon,"Data Mining with Decision Trees: Theory and Applications(Series in

[14].Machine Perception and Artificial Intelligence)", ISB

[15].Venkatadri.M and Lokanatha C. Reddy, "A comparative study on decision tree classification algorithm in data mining" , International Journal Of ComputerApplications In Engineering ,Technology And Sciences

CONFERENCE PAPERS
National Conference on Emerging Trends on Engineering & Technology (ETET-2017)
On 21st April 2017
University Inst. of Engg. & Tech. & University Inst. of Computer, SBBS University, Punjab (India)