ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 6, Issue 1, January 2019

Prediction Using Data Mining Techniques and Tools

A.Sahaya Arthy
Research Scholar, Department of Computer Science
Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, India

Abstract: **- Data mining is a process of extracting knowledge by analyzing various past ancient data bases and predict the end result or take appropriate decision based on prediction. This data mining process has many challenging issues while performing research, analyzing raw data of the past records and predicting may also lead to negative decision result few times. Direct application of methods and techniques developed under related studies in machine learning, statistics and database systems cannot solve these problem. It is required to perform dedicated and appropriate analytical studies to invent new data mining methods or to develop integrated unique techniques for efficient and effective data mining, whereas data mining itself has formed an independent unique and innovative field of study. Data mining is a widely used platform to perform various decisions making in many Industry like banking, finance, agriculture, communication, telecom, Military Service, Police department, other government departments, engineering, Medication &hospitals, law & order etc. This paper deals with detailed study of Data Mining, its techniques, tasks and related tools.**

Keywords: - Data mining Techniques; Data mining algorithms; Data mining tools.

I. INTRODUCTION

Data mining is an extraction of hidden predictive information and acquires knowledge by analyzing a large database. Technically, data mining is the process of discovering correlations or patterns among large number of fields from relational databases. It is a strong tool used widely because it can provide with relevant information that anyone can use for their own advantage. When we have the right data, then we have to apply it in the right manner and then we will be able to get beneficial appropriate result. Now a day's getting information is relatively easy as we have many sources to grab the relevant data. But to achieve certain desired goals it is required to get the most relevant information. This is where data mining becomes a powerful tool that we will become very familiar with. Data mining gives us the power to predict certain behaviors within a system. Data mining involves the high standard detection, association, regression, rule learning, classification, and summarization and clustering. The ultimate goal of this technique is to find various patterns that were previously unknown. Hence the mined results should be valid, novel, useful, and understandable. Once we find these patterns, we can use it to solve many problems.

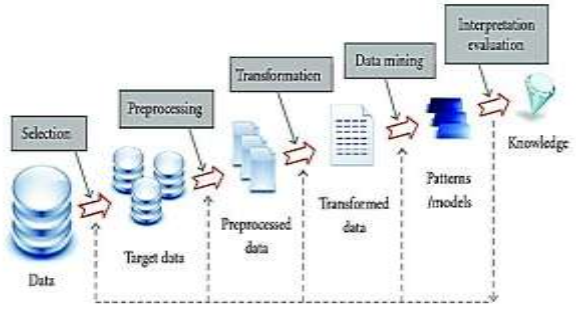

Figure 1. Knowledge Discovery in Database processes

The term Knowledge Discovery in Databases, or KDD for short, refers to the broad process of finding knowledge in data. In KDD the most important step is data mining. KDD will convert the low level data into high level data. Data mining is filed in which useful outcome that is being predicted by analyzing large database. It uses readily built tools to get out the useful hidden patterns, trends and prediction of future which can be obtained using the techniques. Data mining involves various models to discover patterns which consist of various components. [2]

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 6 Issue 1, January 2019

Three steps involved are

- Exploration
- Pattern identification
- Deployment

2. Data Mining Types

Predictive data mining: It produces the model of a system described by the given data. It uses some variables or fields in the data set to predict unknown or future values of other variables of interest.

Descriptive Data Mining: It produces new, significant information based on the available data set. It focuses on finding different patterns describing the data that can be interpreted by human.

3. Data Mining Tasks

- Data processing [descriptive]
- Prediction [predictive]
- Regression [predictive]
- clustering [descriptive]
- Classification [predictive]
- Link analysis associations [descriptive]

4. Prediction on Data Mining Techniques:

There are several data mining techniques which has been developed and used in data mining projects including association, classification, clustering, prediction and neural network.

Classification Analysis Technique:

Classification consists of predicting a certain outcome based on the given input. This type of technique results are purely depends on the input and output will be based on that only. In order to predict the outcome, the algorithm processes a training set containing a set of unique attributes and the respective outcome, usually called destination or prediction attribute. The algorithm tries to discover correlation between the attributes and to predict the outcome. Then the algorithm is given a data set which is not seen before called prediction set, which contains the same set of attributes, except for the prediction attribute – not yet known. The algorithm analyses the input and produces a prediction. The accuracy of the prediction defines the efficiency of the algorithm used, for example in a medical database the training set will have relevant patient information previously recorded and the prediction attribute has to be whether the patient had a heart problem or not. Table 1 below illustrates the training and prediction sets of such database.

Training set- Table 1

Age	Heart rate	Blood pressure	Heart problem
65	78	150/70	yes
37	83	112/76	no
71	67	108/65	no

Prediction set

Age	Heart rate	Blood pressure	Heart problem
43	98	147/89	?
65	58	106/63	?
84	77	150/65	?

Table 1 – Training and Prediction Sets for Medical Database.

In several types of knowledge representation present in the literature, classification normally uses prediction rules to express knowledge. Prediction rules are mostly expressed in the form of IF-THEN rules, where the antecedent (IF part) consists of a conjunction of conditions and the rule consequent (THEN part) predicts a certain prediction attribute value for an item that satisfies the antecedent. [4] Using the example above, a rule predicting the first row in the training set may be represented as following:

IF (Age=65 AND Heart rate>70) OR (Age>60 AND Blood pressure>140/70) THEN Heart problem=yes

In most of the cases prediction rule is extremely larger than the example above. Conjunction has a nice property for classification; each condition separated by OR's defines smaller rules that captures relationship between attributes. Satisfying any of these smaller rules means that the effect is the prediction. Each smaller rule is formed with AND's which narrow down relationships between attributes. How predictions are done is measured in percentage of predictions hit against the total number of predictions. A probable rule ought to have a hit rate greater than the occurrence of the prediction attribute. In other words, if an algorithm is trying to predict rain in a rainy season and it rains 80% of the time, the algorithm could easily have a hit rate of 80% by just predicting rain all the time. Therefore, 80% is the base prediction rate that any algorithm should achieve in this case. The perfect solution is a rule with 100% prediction hit rate, which is very hard, when impossible, to achieve.[4]

Association Rule Learning Technique

An association rule represents a very promising technique to improve heart disease prediction in medical industry. Unfortunately, when association rules are

applied on a medical data set, they produce an extremely large number of rules. Many such rules are medically irrelevant and the time required to find them can be impractical. Four constraints were proposed to reduce the number of rules: item filtering, attribute grouping, maximum item set size, and antecedent/consequent rule filtering. A more important issue in general is association rules are mined on the entire data set without validation on an independent sample. To solve these, the author has introduced an algorithm that utilizes search constraints to reduce the number of rules, searches for association rules on a training set, and finally validates them on an independent test set. Instead of using only support and confidence, one more parameters, that is lift have been used as the metrics to evaluate the medical significance and reliability of association rules. Doctors predominantly use to explicit results. Sensitivity is defined as the probability of identifying sick patients accurately, whereas specificity is defined as the probability of identifying healthy individuals accurately. Along with confidence Lift is also used to understand sensitivity and specificity. An algorithm has three major steps to find predictive association rules in a medical data set. First, a medical data set with categorical and numeric attributes is transformed into a transaction data set. Second, four constraints mentioned above are incorporated into the search process to find predictive association rules with medically relevant attribute combinations in the data. Third, train and test approach is used to validate association rules.

Types of Association Rule

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule
- Fuzzy Association Rules Mining

Decision Tree algorithms

It includes CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and C4.5. These algorithms differ in selection of splits, when to stop a node from splitting, and assignment of class to a non-split node. CART mainly uses Gini index to measure the impurity of a partition or set of training data computing or tuples. It can handle high dimension categorical data. Decision Trees can also handle continuous data (as in regression) but they must be converted to categorical data. The decision tree is built from the very small training set. We will refer to a row as a data instance. The data set contains three predictor attributes, namely Age, Gender, Intensity of symptoms and one goal attribute, namely disease whose values (to be predicted from symptoms) indicates whether the patient have the disease or not.

Neural Network

Neural networks got modeled after the cognitive processes of the brain. They are capable of predicting new observations from existing. This network consists of interconnected processing elements also called units, nodes, or neurons. The neurons within the network work together in parallel to produce an output function. Since the computation is performed by the collective neurons, a neural network can still produce the output function even if some of the individual neurons are malfunctioning. In general, each neuron within a neural network has an associated activation number. Each connection between neurons has a weight associated with it. These quantities simulate their counterparts in the biological brain firing rate of a neuron and strength of a synapse a junction between two nerves. The activation of a neuron depends on the activation of the other neurons and the weight of the edges that are connected to it. The neurons within a neural network are usually arranged in layers. The number of layers within the neural network and the number of neurons within each layer normally matches the nature of the investigated phenomenon. After the size has been determined the network is usually then subjected to training. Then the network receives a sample training input with its associated classes. Then it applies an iterative process on the input in order to adjust the weights of the network so that the future predictions are optimal. After the training phase, the network is ready to perform predictions in new sets of data. Neural networks can often produce very accurate predictions. However, one of their greatest criticisms is that they represent a "black-box" approach to the research. They do not provide any insight into the underlying nature of the scenario.

Clustering Analysis

Clustering is a process of grouping physical or abstract objects into classes of similar objects or unsupervised classification. Clustering analysis helps majorly to construct meaningful partitioning of a large set of objects based on a "divide and conquer" methodology which decomposes a large scale system into smaller components to simplify the design and implementation. Clustering is identifying similar groups from an unstructured data. Clustering is the task of grouping a set of objects in such a way that object in same group are more similar to each other to those in other group. Once the clusters are decided the objects are labeled with their corresponding clusters, and common features of the objects in cluster are summarized to form a class description. For example, a hospital may cluster its patients in to several groups based on the similarities of

their name, age, sex, disease, symptoms, blood pressure, sugar level etc., and the common characteristics of the patients in a group can be used to describe that group of patients. The doctor has to understand the patients to predict their disease and provide medication.

Types of clustering methods

- Hierarchical Methods,
- Partitioning Methods,
- Density Based Methods,
- Grid-Based Methods and
- Model Based Algorithms

Categorization of Clustering Algorithms

Algorithms are the key step for solving multiple techniques. In these clustering techniques, various algorithms are currently live and still lot more are evolving, but in general the algorithm for clustering is neither straight nor canonical.

Hierarchical methods:

- Agglomerative Algorithms and
- Divisive Algorithms.

Partitioning methods:

- Relocation Algorithms,
- Probabilistic Clustering,
- K-Medoids Methods and
- K-Means Methods.

Density-based algorithms:

- Density-based connectivity clustering and
- Density functions clustering.

Grid-based methods:

- Methods based on co-occurrence of categorical data,
- Constraint-based clustering,
- Clustering algorithms used in machine learning,
- Gradient descent and artificial neural networks and
- Evolutionary methods.

Model Based Algorithms:

- Algorithms for high dimensional data,
- Subspace Clustering,
- Projection Techniques and
- Co-Clustering Techniques.

Prediction Techniques

Regression technique shall be adapted for prediction. Regression analysis shall be used to model the relationship between one or more independent variables and dependent variables as well. In data mining independent variables are attributes already known and the response variables are what we want to predict. Unfortunately, many real scenario problems are not simply prediction. For instance, sales volumes, stock prices, petrol prices, crude oil prices, gold rates and product failure rates are all very difficult to predict because they are very dynamic and may depend on complex interactions of multiple predictor variables. Therefore, decision trees, or neural nets are required to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (To classify categorical response variables) and regression trees (To forecast continuous response variables). Neural networks also can create both classification and regression models.

TOOLS FOR DATA MINING TECHNIQUES

There are many open source tools available for data mining. Some of the tools work for clustering, some for classification, regression, association and some for all. There are various algorithms for each technique as discussed above. This section describes features of different tools and which tool can be used to implement which algorithm.

1. Rapid Miner

Rapid Miner is a data science software platform that provides an integrated environment for data preparation, machine learning, deep learning, text mining and predictive analysis. It is one of the apex leading open source systems for data mining.

2. Oracle Data Mining

It represents Oracle's Advanced Analytics Database. Leading companies in the Market use it to maximize the potential of their data to get accurate predictions.

3. IBM SPSS Modeler

It helps to generate data mining algorithms with minimal or without programming. It is widely used in anomaly detection, Bayesian networks, CARMA, Cox regression and basic neural networks use multilayer perceptron with back-propagation learning.

4. KNIME

Konstanz Information Miner is an open source data analysis platform. It is a platform that helps to make predictive intelligence accessible to users without experience also.

5. Python

It is available as a free and open source language; commonly business-use case-data visualizations are straightforward as long as you are comfortable with basic programming concepts like variables, data types,

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 6 Issue 1, January 2019

6. Orange
It is an open source data visualization, machine learning and data mining toolkit. It is a component based visual programming software package for data visualization, machine learning, data mining and data analysis.

7. Kaggle
It is the world's largest community of data scientist and machine learners. Kaggle is a platform that helps to solve the most difficult problems, recruit strong teams and accentuate the power of data science in the Industry.

8. Rattle
Rattle GUI is an open and free software package which provides a graphical user interface for data mining using R Statistical programming language provided by Toga ware.

9. Weka
Weka is a platform which supports several standard data mining tasks, more specifically data pre-processing, clustering, classification, regression, visualization, and feature selection. This is developed in Java and the algorithms can either be applied directly to a dataset or called from Java code, because of the Java API.

10. Teradata
Teradata analytical platform delivers the best functions and leading engines to enable users to leverage their options to choose the tools and languages at scale across different data types.

Comparative Analysis of Data Mining Tools:
In this section, the best available data mining tools [7] were taken and comparative study was made by considering parameter license, technical specification, features and describing in table.

Table2
Features of the most commonly used data mining tools.

Software tool	Types	Features
WEKA	Machine Learning based	It is java based open source data mining tool, based on various data mining and machine learning algorithms.
ORANGE	Machine Learning, Data mining, data visualization	It is open source data mining tool, based on Visual programming, visualization, large tool box, plate form independent, and interaction and data analysis.
Rapid Miner	Statistical analysis, predictive analysis'	It is an important predictive analytic platform. It is user friendly, rich library of data science and has many machine learning algorithms.
Data Melt	Statistical analysis , numeric and symbolic computations, scientific visualization	It is based on cluster analysis, linear regression, neural networks, curve fitting, fuzzy system, analytic calculations & interactive visualizations.
Apache Mahout	Machine Learning based	It is a library of machine learning algorithms which is used in clustering, classification & frequent pattern data mining. It is also used in a distributed mode with integration of Hadoop
ELKI	cluster analysis and outlier detection	It is java based open source data mining tool and licensed under AGPLv3.It focuses on outlier detection and cluster analysis with a compilation of several algorithms from both these domains.
KNIME	Enterprise reporting, Business Intelligence	It is based on java and built upon Eclipse. It is scalable, highly extensive, and capable of data visualization and import –export workflow.
MOA	Machine Learning based	It can handle large volumes of real time data streams at a very high speed and can be used through GUI, command line, or Java API.

KEEL	Machine Learning based	It is java based tool and licensed under GPLV3and based on clustering, classification, and association. It has a very user friendly GUI
Rattle or R	Statistical computing	It is a statistical programming language and R program can run on Mac OS Linux, and Windows, it is based on statistics, clustering, modeling and visualization.

II. CONCLUSION

Data mining offers an appropriate way to discover hidden patterns by various prediction analysis within large amounts of data. These hidden patterns can be potentially used to predict future behavior. The availability of new data mining algorithms is relatively more however; it should be met with caution. There have been a large number of data mining algorithms rooted in this field to perform different data analysis tasks and take multiple decisions across various decision making panel and different industries. This paper dealt with a detailed study on prediction of Data Mining, its techniques, tasks and related tools. Data mining has wide application domain almost in every industry where the data is generated that's why data mining is considered as one of the most important frontiers in database and information systems, data base management systems and one of the most promising interdisciplinary developments in Information Technology. Data mining helps researchers to analyze past data sets, practices, progress, and nature of performances and helps them to predict results, forecast budgets and also to take various constructive decisions in the industry to get better results towards growth.

REFERENCES

[1] Kumbhare, Trupti A., and Santosh V. Chobe. "An overview of association rule mining algorithms." International Journal of Computer Science and Information Technologies 5.1 (2014): 927-930.

[2] Ramageri, Bharati M., and B. L. Desai. "Role of data mining in retail sector." International Journal on Computer Science and Engineering 5.1 (2013): 47.

[3]Ilayaraja, M., and T. Meyyappan. "Mining medical data to identify frequent diseases using Apriori algorithm." Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on. IEEE, 2013.

[4] Voznika, Fabricio, and Leonardo Viana. "Data Mining Classification." (2007).

[5] Tamilselvi, R., and S. Kalaiselvi. "An Overview of Data Mining Techniques and Applications." Int J Sci Eng. Res 1.1-3 (2013): 506-9.

[6] Karur Parminder and Qamar Parvez Rana. "Comparison of various data mining tools", International Journal of Engineering Research & Technology (IJERT)ISSN: 2278-0181IJERTV3IS100246 Vol. 3 Issue 10, October- 2014

[7] Rawat, Keshav Singh, and I. V. Malhan. "Comparative Analysis of Data Mining Techniques, Tools and Machine Learning Algorithms for Efficient Data Analytics." IOSR Journal of Computer Engineering ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 4,

[8] Doddi, Achla Marathe, SS Ravi, David C. Torney, Srinivas. "Discovery of association rules in medical data." Medical informatics and the Internet in medicine 26.1 (2001): 25-33.

[9] Gera, Mansi, and Shivani Goel. "Data mining-techniques, methods and algorithms: A review on tools

and their validity." International Journal of Computer Applications 113.18 (2015).

[10] Verma, Manish, et al. "A comparative study of various clustering algorithms in data mining." International Journal of Engineering Research and Applications (IJERA) 2.3 (2012): 1379-1384.