International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 08 | Aug -2017 www.irjet.net p-ISSN: 2395-0072

DATA MINING – A PERSPECTIVE APPROACH

Arockia Panimalar.S [1], Rubasri. K[2]

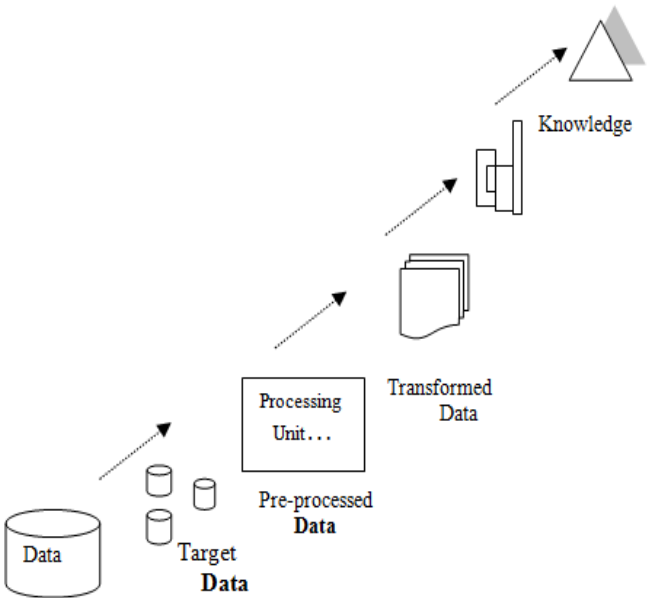
[1] *Assistant Professor, Department of BCA & M.Sc SS, Sri Krishna Arts and Science College, Coimbatore, India*
[2] *III BCA , Department of BCA & M.Sc SS, Sri Krishna Arts and Science College, Coimbatore, India*

Abstract - *In this paper, we have to focus on data mining concept and its tools and technology which help us for a market perspective to consume a proper decision and get a proper result. Data mining is a consistent process that is used to analyze large amounts of information that can be in the form of document in order to find important data. The goal of information mining is to find patterns that were previously unknown. Once you have found out those rules, you can use them to solve a number of complex problems. Data mining [sometimes called data or knowledge discovery from data (KDD)] is the process of analyzing data from the huge amount of information and summarizing it into useful information. Data mining is one of a number of analytical tools for analyzing data. It grants users to search and analyze data from many different sources and transform into decision making data from which user can take a decision. It is in the process of discovering patterns among dozens of fields in large relational databases. Data mining is a powerful tool because it can furnish relevant information. But it is not so easy to find relevant information that can help you to take proper decision. This is where data mining becomes a powerful tool that will help to extract useful information.*

Key Words*: *Data Mining, KDD, Data Mining Task, Data Preprocessing

1. INTRODUCTION

In the 21st century the human beings are used in the different technologies to adequate in the fellowship. Each and every day the human beings are using the huge data and these data are in the different fields. It may be in the form of documents, may be graphical formats, may be the video, and may be recorded (varying array). As the data are useable in the different formats so that the proper action to be taken. Not merely to analyze these data, but also take a good decision and maintain the data. As and when the customer will require the data should be retrieved from the database and make the best decision. This system is really we called as a data mining or Knowledge Hub or basically KDD (Knowledge Discovery Process). The most important reason that drew in a great deal of attention in information technology the discovery of useful information from large collections of data industry towards the field of Data mining is due to the perception we are data rich but information poor. In that respect is a huge volume of data, but we hardly able to turn them into useful information and knowledge for managerial decision making in business. To produce data it requires enormous accumulation of information. It might be diverse configurations like sound/video, numbers, content, figures, and hypertext designs. To take finish preferred standpoint of the information; the information recovery is just insufficient, it requires a device for programmed rundown of information, extraction of the substance of data put away, and the revelation of examples in crude information.



With the huge measure of information put away in documents, databases, and different stores, it is progressively imperative, to grow capable instrument for investigation and translation of such information and for the extraction of fascinating learning that could help in basic leadership. The main answer for all above is Data Mining. Data mining is the extraction of concealed prescient data from expansive databases. It is an intense innovation with remarkable potential to enable associations to concentrate on the most essential data in their information distribution centers. Data mining instruments foresee future patterns and practices, encourages associations to make proactive learning driven conclusions. The robotized, imminent examinations offered by data mining move past the investigations of past occasions gave by planned apparatuses ordinary of choice emotionally supportive networks. Data mining apparatuses can answer the questions that have generally been excessively tedious, making it impossible to

determine. They set up databases for finding shrouded designs, finding prescient data that specialists may miss since it lies outside their desires. [1,6]

2. Data Mining Overview

The growth of Information Technology has generated a large amount of databases and huge data in various areas. The research in databases and information technology has given boost to an approach to store and manipulate this precious data for further decision making. Data mining is a procedure of extraction of useful information and patterns from huge data. It is additionally called as an information disclosure process, learning mining from information, information extraction or information/shape investigation. Data mining is a consistent process that is used to search through large amounts of data in order to find useful data. The goal of this technique is to determine patterns that were previously unknown. Once these figures are discovered they can additionally be utilized to settle on specific choices for the advancement of their organizations. Three steps involved are Exploration, Pattern identification and Deployment.

Exploration: In the first measure of data exploration data is cleaned and transformed into another form, and important variables and then nature of data based on the problem are determined.

Pattern Identification: Once data are explored, refined and defined for the specific variables in the second stride is to form pattern identification. Identify and choose the patterns which make the best prediction.

Deployment: Patterns are deployed for desired effect. [2]

3. Scope of Data Mining

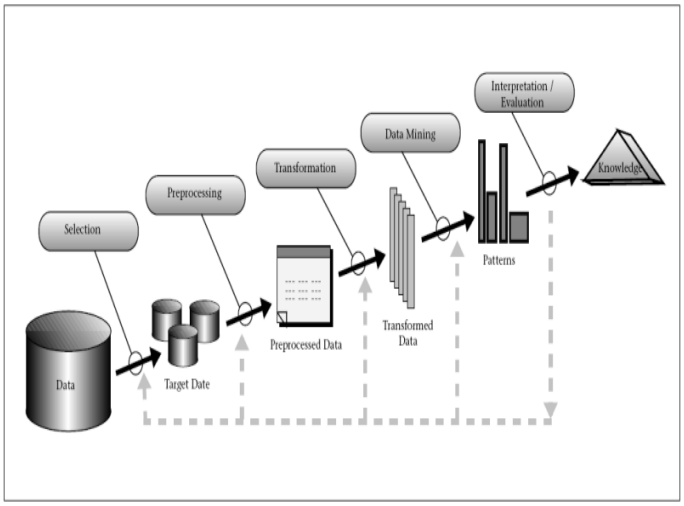
Data mining gets its name from the likenesses between looking for significant employment data in a vast database for instance, finding connected items in gigabytes of store scanner information and digging a mountain for a vein of profitable metal. The two procedures require either filtering through a tremendous total of material, or keenly examining it to discover precisely where the esteem dwells. Collapsed the databases of adequate size and quality, information mining innovation can produce new business openings by giving these abilities:

3.1 Automated expectation of patterns and directs

Data mining computerizes the system of finding prescient data in huge databases. Inquiries that customarily called for broad hands-on investigation would now be able to be addressed straight forwardly from the information rapidly. An unmistakable case of a prescient issue is focused on advertising. Data mining utilizes information on past limited time mailings to key out the objectives destined to boost rate of return in future mailings. Other prescient issues incorporate anticipating insolvency and different types of default, and recognizing portions of a populace liable to answer correspondingly to given occasions.

3.2 Automated disclosure of already obscure figures

Data mining devices clear through databases and recognize beforehand shrouded designs in single step. A case of example revelation is the examination of retail deals information to recognize apparently inconsequential items that are regularly obtained together. Other example revelation issues incorporate recognizing false charge card exchanges and distinguishing bizarre information that could symbolize the information section scratching blunders.



The most generally honed systems in data mining are:

Artificial neural networks: Non-straight prescient models that learn through preparing and look like natural neural systems in social association.

Decision trees: Tree-formed structures that guide sets of decisions. These decisions create rules for the compartmentalization of a dataset. Particular decision tree strategies incorporate Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).

Genetic algorithms: Optimization methods that utilization operation, for example, genetic blend, changes, and common determination in a plan in view of the ideas of advancement.

Closest neighbor strategy: A method that orders each record in a dataset in light of a blend of the classes of the k record(s) most like it in an authentic dataset (where k ^{3} 1).

Rule induction: The extraction of helpful if-then principles from information in view of factual noteworthiness. [3,9]

4. TYPES OF DATA MINING

Data mining frameworks can be ordered by different criteria the arrangement is as per the following:

4.1 Classification of data mining frameworks as indicated by the kind of data source mined

This characterization is as indicated by the kind of information took care of, for example, spatial data, interactive media information, time-arrangement information, content information and World Wide Web.

4.2 Classification of data mining frameworks as indicated by the data model

This arrangement in light of the data display included, for example, social database, question situated database, information distribution center, value-based database, and so on.

4.3 Classification of data mining frameworks, as indicated by the kind of knowledge discovered

This grouping in light of the assortment of information found or data mining functionalities, for example, portrayal, segregation, affiliation, order, bunching, and so forth. A few frameworks have a tendency to be far reaching frameworks offering a few information mining functionalities together.

4.4 Classification of data mining frameworks, as per excavation techniques used

This characterization is as per the information investigation approach utilized, for example, machine learning, neural nets, hereditary calculations, insights, perception, database arranged or information distribution center situated, and so forth.

The classification can likewise consider the level of client connection engaged with the information mining procedure, for example, inquiry driven frameworks, intelligent exploratory frameworks, or self-sufficient frameworks. A thorough framework would offer a wide assortment of information mining procedures to fit distinctive circumstances and alternatives, and offer diverse degrees of client interaction.[4]

5. Data Mining Applications

In this segment, we have focused some of the applications of data mining and its techniques are analyzed respectively order.

5.1 Data Mining Applications in Healthcare

Data mining applications in health can have enormous potential and usefulness. Nevertheless, the success of healthcare data mining hinges on the availability of clean healthcare data. In this regard, it is critical that the healthcare industry look into how data can be better captured, stored, prepared and mined. Possible charges include the standardization of clinical vocabulary and the sharing of data across organizations to enhance the benefits of healthcare data mining applications.

5.2 Data Mining is used an Emerging Trends in the Education System in the Whole World

In Indian culture most of the parents are uneducated. The main aim of in Indian government is the quality education not for quantity. But the daylight by day the education schemes are changing and in the 21st century a huge number of universities are established. As the numbers of universities are shown side by side, each and every day a millennium of students are enrolls across the country. With enormous number of advanced education wannabes, we trust that information mining innovation can help connecting learning hole in higher instructive frameworks. The concealed examples, affiliations that are distinguished by information mining procedures from instructive information can enhance basic leadership forms in higher instructive frameworks. This approach can bring advantages such as maximizing educational system efficiency, decreasing student drop-out rate, and increasing student's promotion rate, increasing student retention rate, increasing student's transition rate, increasing educational improvement ratio, increasing success, increasing student learning outcome, and reducing the cost of system processes. In this current era we are using the KDD and the data mining tools for extracting the knowledge this knowledge can be employed for improving the quality of education. The decision tree classification is utilized in this type of applications.

5.3 Data Mining is now used in many different areas in Manufacturing Engineering

When we retrieve the data from manufacturing system, then the customer is to apply these data for different purposes like to find the errors in the data, to enhance the design methodology, to make the good quality of the data, how best the data can be supported in making the decision. Only most of the time the data can be first analyzed, then after finding the hidden patterns which will be control the manufacturing process which will further enhance the quality of the products. Since the significance of data mining in assembling has unmistakably expanded in the course of the most recent 20 years, it is currently proper to basically survey its history and Application.

5.4 Sports Data Mining

The data mining and its strategy is utilized for a use of Sports focus. Data mining is utilized as a part of the business purposes, as well as it utilized as a part of the plays. On the planet, a colossal number of amusements are accessible where every single day the national and universal diversions are to be booked, where an immense number of data are to be kept up.

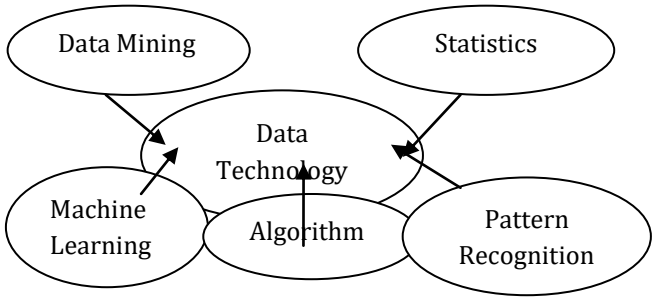


Fig: Data mining confluence in many disciplines

The data mining tools are used to give the information as and when we required. The open source data mining tools like WEKA is used for athletics. This means that users can run their data through one of the built-in algorithms, see what results come out, and then run it through a different algorithm to look if anything different stands out. As these programs are available in the form of open source in nature, that's why the users are frequent to modify the source code, so that others can get the updated data. In the sports world the vast measures of statistics are collected for each player, team, game, and season. In the game sports the data are usable in the form of statistical form where data mining can be used and discover the patterns, these patterns are often used to predict the future forecast. Data mining can be connected for forecast of execution, choice of players, instructing and preparing and for the key arranging. The information mining systems are used to manage the best or the most ideal squad to speak to a group in a group activity in a season, visit or amusement. [5,8]

6. DATA MINING LIFE CYCLE

The sequence of the stages is not rigid. Going back and forth between different phases is always required. It depends on the issue of each phase. There are six main phases to the process:

A. Business Understanding

This stage concentrates on understanding the venture targets and necessities from a business viewpoint, at that point changing over this learning into a data mining issue definition and a preparatory arrangement intended to accomplish the goals.

B. Data Understanding

It starts with an underlying information gathering, to get comfortable with the information, to distinguish information quality issues, to find first bits of knowledge into the information or to identify fascinating subsets to frame theories for shrouded data.

C. Data Preparation

In this phase, it collects all the different data sets and constructs the varieties of the activities basing on the initial raw data.

D. Modeling

In this stage, different displaying procedures are chosen and connected and their parameters are aligned to ideal esteems.

E. Evaluation

At this point the model is thoroughly evaluated and reviewed. The steps executed to build the model to be certain it properly achieves the business targets. At the terminal of this phase, a decision on the use of the data mining results should be reached.

F. Deployment

The aim of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. The organization stage can be as basic as creating a report or as mind boggling as experiencing a repeatable data mining process over the enterprise.[5]

7. DATA MINING TECHNIQUES

Data mining embraces its procedure from many research fields, including statics machine learning, database frameworks unpleasant sets, representation and neural systems.

A. Statistical Approach

Statistical models are built from a lot of training data. Numerous factual apparatuses have been utilized for data mining including, Bayesian system, relationship examination, relapse investigation and group examination. In the Bayesian system hubs speaks to states or variable while edges speak to conditions between customers. From the figure we can see that surge hour, extreme climate or mishap influences the movement which thus causes activity mess.

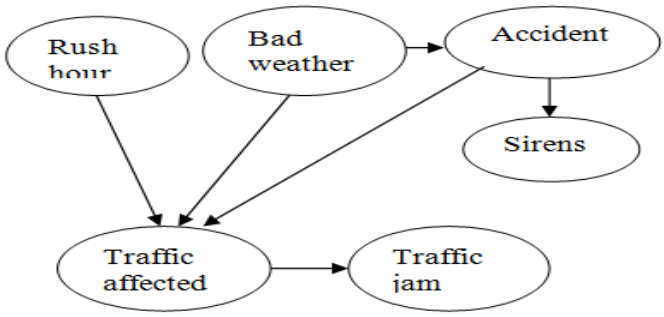


Figure 2

B. Machine Learning Approach

The most common machine learning methods applied for data mining include conceptual learning, inductive concept learning and decision tree induction. By adopting the path from root to leaf node an object class can be determined by a decision tree. Decision trees are induced from the training set and decision trees give classification rules. A simple decision tree is given in Figure 3, it determines the cars mileage from its size, transmission type and weight.
The leaf nodes are in solid boxes.

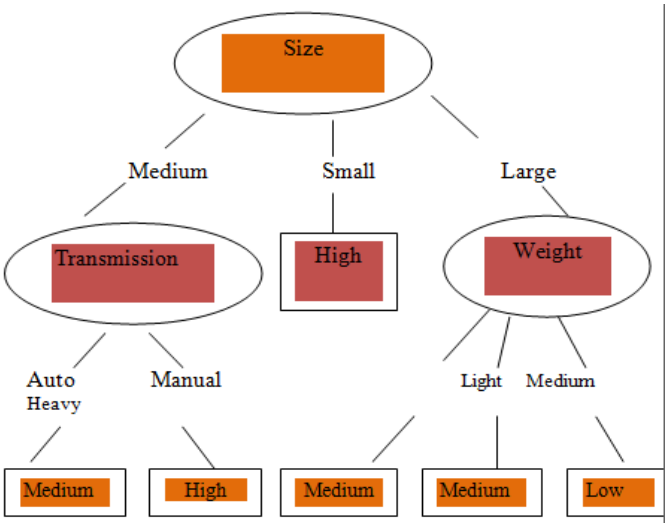


Figure 3. A Simple Decision Tree

From decision tree we can conclude, for instance, large size; the heavy weight car will have low mileage. Nodes represent three classes of mileage. [6]

8. Data Mining Methods

Popular data mining methods are as follows:
i. Decision Trees and Rules
ii. Nonlinear Regression and Classification Methods
iii. Example-Based Methods
iv. Probabilistic Graphical Dependency Models
v. Relational Learning Models

We found, these are some famous data mining methods are generally classified as: On-Line Analytical Processing, (OLAP), Classification, Clustering, Association Rule Mining, Temporal, Data Mining, Time Series Analysis, Spatial Mining, Web Mining etc. These methods employ different types of algorithms and data. The data source can be data warehouse, database, flat file or text file. The algorithms are Statistical based Algorithms, Decision based Tree, Nearest Neighbor, Neural Network, Genetic Algorithms, Rules based algorithm, Support Vector Machine, and so on. By and large the data mining calculations are completely snared on the two factors these are

(i) Which character of data sets is used?
(ii) What type of prerequisites of the user?
Establishing upon the above two factors, the data mining algorithms are used. A knowledge discovery
(KD) The process involves preprocessing data, selecting a data-mining algorithm, and post processing the mining results.

The Intelligent Discovery Assistants (IDA), helps users in applying valid knowledge discovery operations. The IDA can give clients three advantages:

- A deliberate count of legitimate learning disclosure forms
- Effective rankings of legitimate procedures by various measures, which help to pick between the alternatives.
- A foundation for sharing information, which advisers for organizes externalities.

Several other attempts have been made to automate this process and design of a generalized data mining tool that possesses intelligence to select the data and data mining algorithms and up to some extent the knowledge discovery.[7]

9. Visualizing Data Mining Model

The chief objective of visualization of data is the overall idea about the data mining model. In data mining most of the times we are extracting the data from the data repositories which are in the hidden form. It is really difficult task for an end user. So this visualization of the data mining model helps us to provide topmost levels of understanding and confidence. Because the user does not recognize what the data mining process has discovered.
The data mining models are categorized in two cases:

Predictive and Descriptive Models

The predictive model makes prediction about unknown or missing data values by applying the known values. Ex: Classification, Regression, Prediction and Time series analysis.
The descriptive model distinguishes the examples or connections in information and investigates the properties of

the information examined. Ex. Bunching, Association administers Summarization and Sequence disclosure.
Many of the data mining applications are aimed to anticipate the future state of the data.

i. Prediction is the process of studying the current and past states of the variables or attribute and prediction of its future state.
ii. Classification is a data mining technique of representing the target data to the classes or predefined groups, this is a supervise learning because the classes are predefined before the examination of the target data.
iii. The regression technique involves the learning of function that map information item or data set to a real valued prediction variable.
iv. In the time series analysis technique the value of an attribute or item is analyzed as it varies over time.
v. The term clustering means analyzes the set of different data objects without consulting a known class levels. It is concerned to as unsupervised learning or segmentation. It is the segmentation or partitioning of the data set into similar type of groups or bunches. The clusters are determined by grouping of similar type of objects into one cluster. The term segmentation is a process of partitioning of database into disjoint grouping of similar tuples.
vi. Summarization is the technique of representing summarize or accurate information from the data.
The association rule finds the association between the different attributes.
vii. Association rule mining is a process in two-step:
Finding all frequent item sets.
Generate strong association rules from the frequent attributes or item sets.
viii. Sequence discovery is a process of finding the sequence rules in the data set. This sequence can be utilized to understand the trend.

A novel way to define the KDD process

We have ground the broader meaning of the followings Patterns, data, Process, Valid, Novel, and Useful Understandable of KDD. The Knowledge revelation in data vault or databases is the non-paltry procedure of distinguishing legitimate, valuable, crisp, and at last justifiable examples in data.

Data	A set of facts, F.
Patterns	An expression E in language L described facts in a subset FE of F
Process	It means different operations associated with the KDD. The operations involving preparation of the data, searching the different patterns, Judging the knowledge and evaluation etc.
Valid	Those patterns which are discovered that are completely new one and which can be used feature.
Novel	Derive the hidden patterns
Useful	Newly discovered patterns should be used for different actions.

Table to describe the new form of word [8]

10. Conclusion

In this paper, we briefly surveyed the various data mining concepts, its techniques and applications. Data Mining is not a new term, but in the recent years its growth day by day touches great horizons. It has spread in almost all areas nowadays. It is clear that Data Mining tools helps in extracting useful or meaningful knowledgeable attributes or data from the unimaginable massive data. This review would be helpful for the researchers focus on the various matters of data mining.
In future, we will review the popular classification algorithms and significance of their evolutionary computing approach in designing of efficient classification algorithms for data mining. [8]

11. References

[1] Neelamadhab Padhy, Dr. Pragnyaban Mishra , and Rasmita Panigrahi "The Survey of Data Mining Applications And Feature Scope", Vol.2, No.3, June 2012 DOI : 10.5121/ijcseit.2012.2303 43
[2] Nidhi Trivedi "DATAMINING TECHNIQUE APPLICATION AND FUTURE SCOPE" Vol-2 Issue-6 2016 IJARIIE-ISSN(O)-2395-4396
[3] Annan Naidu Paidi" Data Mining: Future Trends and Applications" Vol.2, Issue.6, Nov-Dec. 2012 pp-4657-4663 ISSN: 2249-6645
[4] Pragnyaban Mishra ,Neelamadhab Padhy, Rasmita Panigrahi" THE SURVEY OF DATA MINING APPLICATIONS AND FEATURE SCOPE" ISSN 2249-5126
[5] Poonam B. Patthe" The Survey of Data Mining Applications and Feature Scope" Volume 4 Issue IV, April 2016 IC Value: 13.98 ISSN: 2321-9653
[6] S.D.Gheware, A.S.Kejkar, S.M.Tondare" Data Mining: Task, Tools, Techniques and Applications" Vol. 3, Issue 10, October 2014 ISSN (Online): 2278-1021 ISSN (Print) : 2319-5940
[7] Neelamadhab Padhy1, Dr. Pragnyaban Mishra 2, and Rasmita Panigrahi3" The Survey of Data Mining Applications and Feature Scope", Vol.2, No.3, June 2012 DOI: 10.5121/ijcseit.2012.2303 43
[8] Dr. Zubair Khan1, Ashish Kumar2, Sunny Kumar3"A Survey of Data Mining: Concepts with Applications and its future scopes" Volume 2 Issue 3, May-Jun 2014, ISSN: 2347-8578
[9] Mrs. Bharati, M. Ramageri" DATA MINING TECHNIQUES AND APPLICATIONS" Vol. 1 No. 4 301-305, ISSN: 0976-5166