National Conference on Recent Research in Engineering and Technology (NCRRET-2015) International Journal of Advance Engineering and Research Development (IJAERD)
e-ISSN: 2348 - 4470 , print-ISSN:2348-6406

# A Review on Data mining Techniques

**Jimesh Rana**
*Asst.Proff, Computer Science & Engineering Department*
*DJMIT,Mogar,Anand*
*Gujarat Technological University*
*Jimesh.rana@djmit.ac.in*

**Anu Thomas**
*Asst.Proff, Computer Science & Engineering Department*
*DJMIT,Mogar,Anand*
*Gujarat Technological University*
*Anu.thomas@djmit.ac.in*

***Abstract— Presently, a very large amount of data stored in databases is increasing at a tremendous speed. This requires a need for new techniques and tools to aid humans in automatically and intelligently analyzing large data sets to acquire useful information. This growing need gives a view for a new research field called Knowledge Discovery in Databases (KDD) or Data Mining, which attract a attention from researchers in many different fields including database design, statistics, pattern recognition, machine learning, and data visualization. Data mining is the process of discovering insightful, interesting, and novel patterns, as well as descriptive, understandable and predictive models from large-scale data. In this paper we overviewed different tasks includes in Data mining. Data mining involves the tasks like anomaly detection, classification, regression, association rule learning, summarization and clustering.***

***Keywords— Data Mining, classification, clustering, association rules***

## I. INTRODUCTION

The last decade has experienced a revolution in information availability and exchange of it through internet. In the same strength more business as well as organizations began to collect data related to their own operations, while the database technologist have been seeking efficient mean of storing, retrieving and manipulating data, the machine learning community focused on techniques which used for developing, learning and acquiring knowledge from the data. Data Mining is the process of analysing data from different perspectives and summarizing it into useful information.

Data mining consists of extract, transform, and load transaction data onto the data warehouse system, store and manage the data in a multidimensional database system, by using application software analyse the data, provide data access to business analysts and information technology professionals, present the data in a useful format, like a graph or table. Data mining involves the anomaly detection, association, classification, regression, rule learning, summarization and clustering.

## II. DATA MINING

Data mining is the exploration and analysis of large data sets, in order to discover meaningful pattern and rules . The key idea is to find effective way to combine the computer's power to process the data with the human eye's ability to detect patterns. The objective of data mining is to design and work efficiently with large data sets. Data mining is the component of wider process called knowledge discovery from database. [4]. Data Mining is the process of analysing data from different perspectives and summarizing the results as useful information. It has been defined as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data"

The definition of data mining is closely related to another commonly used term knowledge discovery [2]. Data mining is an interdisciplinary, integrated database, artificial intelligence, machine learning, statistics, etc. Many areas of theory and technology in current era are databases, artificial intelligence, data mining and statistics is a study of three strong large technology pillars. Data mining is a multi-step process, requires accessing and preparing data for a mining the data, data mining algorithm, analysing results and taking appropriate action. The data, which is accessed can be stored in one or more operational databases. In data mining the data can be mined by passing various process.

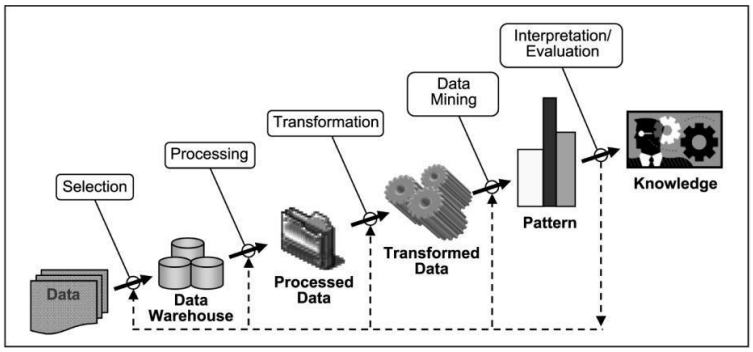


Fig. 1 : Steps in Data Mining process

In data mining the data is mined using two learning approaches i.e. supervised learning or unsupervised learning [5].

### *A. Supervised Learning*

In supervised learning (often also called directed data mining) the variables under investigation can be split into two groups: explanatory variables and one (or more) dependent

variables. The goal of the analysis is to specify a relationship between the dependent variable and explanatory variables the as it is done in regression analysis. To proceed with directed data mining techniques the values of the dependent variable must be known for a sufficiently large part of the data set.

*B. Unsupervised Learning:*

In unsupervised learning, all the variables are treated in same way, there is no distinction between dependent and explanatory variables. However, in contrast to the name undirected data mining, still there is some target to achieve. This target might be as data reduction as general or more specific like clustering. The dividing line between unsupervised learning and supervised learning is the same that distinguishes discriminate analysis from cluster analysis. Supervised learning requires, target variable should be well defined and that a sufficient number of its values are given. Unsupervised learning typically either the target variable has only been recorded for too small a number of cases or the target variable is unknown.

## III. ISSUES IN DATA MINING

Data mining has evolved into an important and active area of research because of the theoretical challenges and practical applications associated with the problem of discovering interesting and previously unknown knowledge from real-world databases. The main challenges to the data mining and the corresponding considerations in designing the algorithms are as follows:

1. Massive datasets and high dimensionality.
2. Over fitting and assessing the statistical significance.
3. Understandability of patterns.
4. Non-standard incomplete data and data integration.
5. Mixed changing and redundant data.

## IV. TASKS OF DATA MINING

Data mining as a term used for the specific classes of six activities or tasks as follows:

1. Classification
2. Estimation
3. Prediction
4. Affinity grouping or association rules
5. Clustering
6. Description and visualization

*A. Classification*

Classification consists of examining the features of a newly presented object and assigning to it a predefined class. The classification task is characterized by the well-defined classes, and a training set consisting of reclassified examples. The task is to build a model that can be applied to unclassified data in order to classify it. Examples of classification tasks include:

- Classification of credit applicants as low, medium or high risk
- Classification of mushrooms as edible or poisonous
- Determination of which home telephone lines are used for internet access

*B. Estimation*

Estimation deals with continuously valued outcomes. Given some input data, we use estimation to come up with a value for some unknown continuous variables such as income, height or credit card balance. Some examples of estimation tasks include:

- Estimating the number of children in a family from the input data of mothers' education
- Estimating total household income of a family from the data of vehicles in the family
- Estimating the value of a piece of a real estate from the data on proximity of that land from a major business centre of the city.

*C. Prediction*

Any prediction can be thought of as classification or estimation. The difference is one of emphasis. When data mining is used to classify a phone line as primarily used for internet access or a credit card transaction as fraudulent, we incorrect, but the uncertainty is due to incomplete knowledge only: out in the real world, the relevant actions have already taken place. The phone is or is not used primarily to dial the local ISP. The credit card transaction is or is not fraudulent. With enough efforts, it is possible to check.

Predictive tasks feel different because the records are classified according to some predicted future behaviour or estimated future value. With prediction, the only way to check the accuracy of the classification is to wait and see. Examples of prediction tasks include:

- Predicting the size of the balance that will be transferred if a credit card prospect accepts a balance transfer offer
- Predicting which customers will leave within next six months
- Predicting which telephone subscribers will order a value–added service such as three-way calling or voice mail.

Any of the techniques used for classification and estimation can be adopted for use in prediction by using training examples where the value of the variable to be predicted is already known, along with historical data for those examples. The historical data is used to build a model that explains the current observed behaviour. When this model is applied to current inputs, the result is a prediction of future behaviour [7].

*D. Association Rules*

An association rule is a rule which implies certain association relationships among a set of objects (such as "occur

together" or "one implies the other") in a database.

Given a set of transactions, where each transaction is a set of literals (called items), an association rule is an expression of the form X Y , where X and Y are sets of items. The intuitive meaning of such a rule is that transactions of the database which contain X tend to contain Y. An example of an

association rule is:

"30% of farmers that grow wheat also grow pulses; 2% of all farmers grow both of these items". Here 30% is called the confidence of the rule, and 2% the support of the rule. The problem is to find all association rules that satisfy user-specified minimum support and minimum confidence constraints.

*E. Clustering*

Cluster analysis can be used as a standalone data mining tool to gain insight into the data distribution, or as a pre-processing step for other data mining algorithms operating on the detected clusters. Many clustering algorithms have been developed and are categorized from several aspects such as partitioning methods, hierarchical methods, density-based methods, and grid-based methods .Further data set can be numeric or categorical.

Clustering is the task of segmenting a diverse group into a number of similar subgroups or clusters. What distinguishes clustering from classification is that clustering does not rely on predefined classes. In clustering, there are no predefined classes. The records are grouped together on the basis of self similarity. Clustering is often done as a prelude to some other form of data mining or modelling. For example, clustering might be the first step in a market segmentation effort, instead of trying to come up with a one-size-fits-all rule for determining what kind of promotion works best for each cluster[6].

*1) General Types Of Clusters:*

- *Well-separated clusters*: A cluster is a set of points so that any point in a cluster is nearest (or more similar) to every other point in the cluster as compared to any other point that is not in the cluster.
- *Center-based clusters* A cluster is a set of objects such that an object in a cluster is nearest (more similar) to the
  "center" of a cluster, than to the center of any other cluster. The center of a cluster is often a centroid.
- *Contiguous clusters* A cluster is a set of points so that a point in a cluster is nearest (or more similar) to one or more other points in the cluster as compared to any point that is not in the cluster.
- *Density-based clusters* A cluster is a dense region of points, which is separated by according to the low-density regions, from other regions that is of high density.
- *Shared Property or Conceptual Clusters* Finds clusters that share some common property or represent a particular concept.

*F. Description and Visualization*

Data visualization is a powerful form of descriptive data mining. It is not always easy to come up with meaningful visualizations, but the right picture really can be worth a thousand association rules since the human beings are extremely practiced at extracting meaning from visual scenes.

Knowledge discovery goals are defined by the intended use of the system.

There are two types of goals: (1) verification and (2) discovery. With verification, the system is limited to verifying the user's hypothesis. With discovery, the system autonomously finds new patterns. The discovery goal is further divided into prediction, where the system finds patterns for predicting the future behaviour of some entities and description, where the system finds patterns for presentation to a user in human understandable form.

## V. CONCLUSIONS

Data mining involves extracting useful rules or interesting patterns from huge historical data. Many data mining tasks are available and each of them further has many techniques. Data mining is an interdisciplinary, artificial intelligence, application of the data found in hidden, regularity which are not known by people in advance, but is potentially useful and ultimately understandable information and knowledge of non-trivial process In this paper we discusses some issues in Data Mining and activities used for Data mining task.

## REFERENCES


[1] Amandeep Kaur Mann, Navneet Kaur, *Survey Paper on Clustering Techniques*, *IJSETR,2278 – 779*.

[2] Pavel Berkhin, A *Survey of Clustering Data Mining Techniques*, pp.25-71, 2002.

[3] Oded Maimon, Lior Rokach, *Data Mining AND Knowlwdge Discovery Handbook*, Springer Science + Business Media.Inc, pp.321-352, 2005.

[4] Han, J., Kamber, M., *Data Mining Concepts and Techniques*, Morgan Kaufmann Publisher, 2001.

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

[6] Aastha Joshi, Rajneet Kaur, *A Review: Comparative Study of Various Clustering Techniques in Data Mining*, IJARCSSE, Vol. 3, 2013, 2277 128X

[7] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta, *A Comparative Study of Various Clustering Algorithms in Data Mining,, International Journal of Engineering Reserch and Applications (IJERA)*, Vol. 2, Issue 3, pp.1379-1384, 2012.

[8] Pradeep Rai, Shubha Singh, *A Survey of Clustering Techniques*, International Journal of Computer Applications, 2010.