ISSN: 2395-7639

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

(A Monthly, Peer Reviewed Online Journal)

Visit: www.ijmrsetm.com

Volume 4, Issue 4, April 2017

A Study on Prevailing Data Mining Techniques

Umang

Assistant Professor, Department of Information Technology, KU, SSJ Campus Almora, Uttarakhand, India

ABSTRACT: Massive or voluminous data is available on several online platforms; an enormous amount of information can be retrieved from this massive data. So as to unveil hidden knowledge, there is a need to process this huge voluminous data to capture underlying information or hidden patterns; this process of unveiling hidden information is leading to the evolution of data mining. In this study, we have analyzed the data mining process, which comprises several steps to mine the data; further, we have investigated the different data mining techniques.

KEYWORDS: Analysis, Association, Classification, Clustering, Data mining process.

I.INTRODUCTION

Data mining is extracting helpful information or pattern from a large amount of data. Data mining is also referred to as knowledge data discovery [2]. The need for data mining arises due to the evolution of information technology. Data mining is an interdisciplinary approach comprising an integrated database, artificial intelligence, machine learning, statistics, etc. Data mining is a multi-step process, which requires accessing and preparing data for Mining, using data mining algorithms, analyzing results and taking appropriate action.[4] The data, which is accessed, can be stored in one or more operational databases. In data mining, the data can be mined by carrying out various processes.

Data mining involves the following steps.

1. Data Cleaning: It is the first step in the data mining process, where we clean the data, i.e. we remove noise & outliers so that the desired data can be used for the mining process.
2. Data Integration: Data from multiple heterogeneous sources such as flat files, transactional data, and data warehouse are being combined & integrated for mining purpose
3. Data selection: Data relevant to the mining task are being selected
4. Data Transformation: Data is transformed & consolidated appropriately for the mining process.
5. Mining Methods: Different data mining methods are being selected & applied to the data sets.
6. Pattern Evaluation: The criteria for pattern evaluation are being examined
7. Knowledge Presentation: The appropriate knowledge presentation technique, such as bar graphs, is selected for representing exciting patterns.

II. ISSUES IN DATA MINING

The need for data mining rises due to the availability of an enormous amount of data, so a technique is required to retrieve valuable patterns from a large amount of data. Furthermore, 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 [8].
The main challenges to data mining and the corresponding considerations in designing the algorithms are as follows:

1. Enormous datasets and high dimensionality.
2. Assessing the statistical significance.
3. Understandability of the discovered patterns.
4. Non-standard incomplete data and data integration.
5. Handling redundant data.

ISSN: 2395-7639

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

(A Monthly, Peer Reviewed Online Journal)

Visit: www.ijmrsetm.com

Volume 4, Issue 4, April 2017

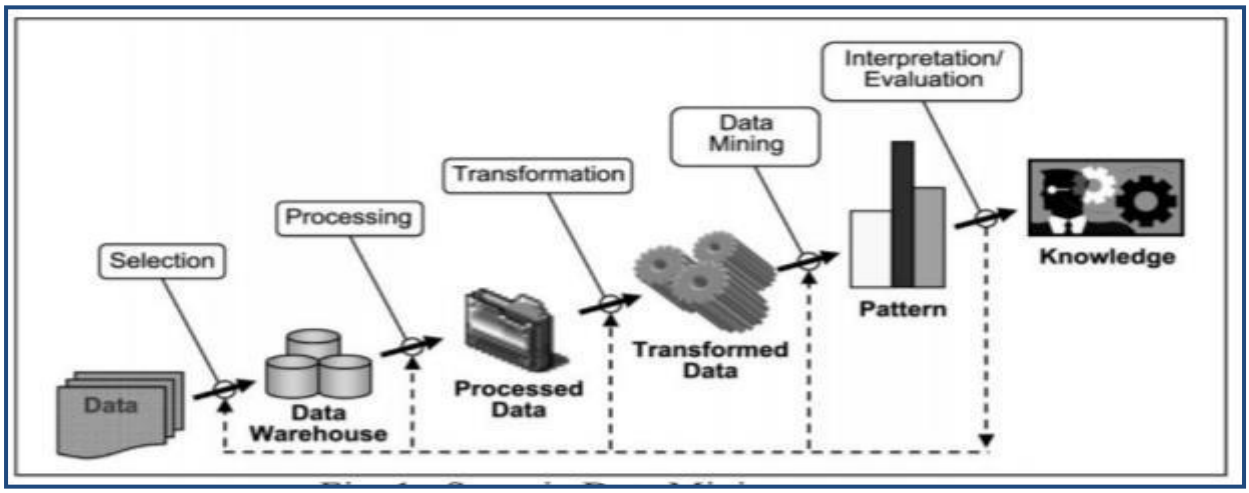

Fig. 1: Data Mining Process

III. TASKS OF DATA MINING

Data mining consists of the following activities:

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

The first three tasks - Classification, estimation and prediction rules are examples of directed data mining or supervised learning. In directed data mining, the goal is to use the available data to build a model that describes one or more particular attribute(s) of interest (target attributes or class attributes) in terms of the rest of the available features. The following three tasks – association rules, clustering and description are examples of undirected data mining, i.e. no attribute is singled out as the target, and the main goal is to establish some relationship among all details [6].

1. **Classification:** Classification is examining the features of a newly presented object and assigning to it a predefined class. Well-defined courses and a training set of reclassified examples characterize it. The task is to build a model that can be applied to unclassified data to classify it.[7] Examples of classification tasks include the Classification of credit applicants as low, medium or high risk.
2. **Estimation:** Estimation deals with continuously valued outcomes. Given some input data, we use analysis to come up with a value for some unknown continuous variables such as income, height or credit card balance [1] Examples of estimation tasks include: Estimating the number of students in a class from the input data of children's background knowledge.
3. **Prediction:** Predictive modelling uses statistics to predict outcomes [3]. Most often, the event one wants to predict is in the future, but predictive modelling can be applied to any unknown event, regardless of when it occurred. Any prediction can be thought of as Classification or estimation.
4. **Clustering:** Clustering is finding groups of objects such that the objects in one group will be similar to one another and different from the objects in another group. Cluster analysis can be used as a standalone data mining tool to gain insight into the data distribution or as a preprocessing step for other data mining algorithms operating on the detected clusters [3]. Many clustering algorithms have been developed and categorized from several aspects, such as partitioning, hierarchical, density-based, and grid-based methods.

Different types of Clusters are as follows:

1. Well-separated clusters [8]: 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.

ISSN: 2395-7639

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

(A Monthly, Peer Reviewed Online Journal)

Visit: www.ijmrsetm.com

Volume 4, Issue 4, April 2017

2. Centre-based clusters: A cluster is a set of objects such that a thing in a cluster is nearest (more similar) to the "centre" of a cluster than to the centre of any other cluster. The centre of a cluster is often a centroid.
3. 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.
4. Density-based clusters: A cluster is a dense region of points, separated according to the low-density areas from other parts of high density.
5. Shared Property or Conceptual Clusters: Finds clusters that share some common property or represent a particular concept.
6. Association Analysis: Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules found in databases using some measures of interestingness. Based on the concept of strict rules. An example of an association rule would be, "If a customer buys bread, he is 80% likely also to purchase butter.
   Several types of data mining techniques available are-
   Classification, clustering, regression, outer, sequential patterns, prediction, association rules and trackers. Several data visualization tools are also available; we can also visualize data in MS. Excel in form of graphs etc

IV.CONCLUSIONS

Data mining involves extracting valid rules or exciting patterns from massive data. It is an interdisciplinary discipline integrating multiple fields, such as artificial intelligence, machine learning, and Statistics. Data mining involves various techniques such as Classification, clustering, and association. Data mining is to discover knowledge that is of interest from large amounts of data stored in data repositories. Mining refers to the extraction of something which is of value; hence, data mining is to pull out valuable information from data. Numerous data mining tools are available in the market to predict future trends and assist decision-making, which further helps organizations make proactive decisions by looking into past and present data. The varied application areas of data mining are marketing/sales, customer relationship management, banking, insurance, fraud detection, bioinformatics and many more.

REFERENCES

[1] Leonid Churilov. Adil Bagirov, Daniel Schwartz, Kate Smith, Michael Dally, Journal of management information system : 2005, Data mining with combined use of optimization techniques and self organizing maps for improving risk grouping rules : application to prostate cancer patients
[2] Anthony Danna, Oscar H. Gandy, Journal of business ethics : 2002, All that glitters is not gold : Digging Beneath the surface of data mining.
[3] AC Yeo, KA Smith, RJ Willis and M Brooks, Journal of the operation research society : 2002 , A mathematical programming approach to optimize insurance premium pricing within a data minning framework.
[4] Shakil Ahmed, Frans Coenen, Paul Leng, Knowledge Information System : 2006, Tree based partitioning of data for association rule mining
[5] Timothy T. Rogers, James L. Mcclelland, Behavioral & Brain Sciences : 2008, Precis of Semantic Cognition : A Parallel Distributed Processing Approach.
[6] Ana Cristina, Bicharra Garcia, Inhauma Ferraz and Adriana S. Vivacqua, Artificial Intelligence for engineering design, analysis and manufacturing: 2009, From data to Knowledge Mining
[7] Rachid Anane, Computer and the humanities : 2001, Data mining and serial documents.
[8] Balaji Padmanabhan and Alexander Tuzhilin, Institute for Operation Research and Management Science : 2011, On the use of optimization for data mining : Theorotical Interaction and eCRM opportunities.