IJCAT - International Journal of Computing and Technology, Volume 5, Issue 4, April 2018
ISSN (Online) : 2348-6090
www.IJCAT.org
Impact Factor: 0.835

# Basics of Data Mining Techniques and its Application

**[1] S. Sathiyapriya; [2] Dr. A. Kanagaraj**

[1]Ph.D Research Scholar
NGM College
Pollachi – 642001

[2]Assistant Professor
PG Department of Computer Science
NGM College, Pollachi-642001

**Abstract -** Data mining is the process of analyzing data from different views and summarizing it into useful data. "Data mining, also popularly referred to as knowledge discovery from data (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories or data streams.". This paper provides a survey on various data mining techniques such as classification, clustering, regression, summarization and so on. This paper also discusses some of the data mining applications.

***Keywords –*** *KDD, Data Mining, Data Mining Techniques, Data Mining Application.*

## 1. INTRODUCTION

Data mining, discovering of hidden predictive information from large data sets and it is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses.[11] Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

## 2. DATA MINING PROCESS

Data mining is also known as Knowledge Discovery in Database, refers to finding or "mining" knowledge from large amounts of data.[1] Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. So, many people use the term "knowledge discovery in data" or KDD for data mining.

In Data mining, Knowledge extraction or discovery is done in seven sequential steps as in Figure 1.

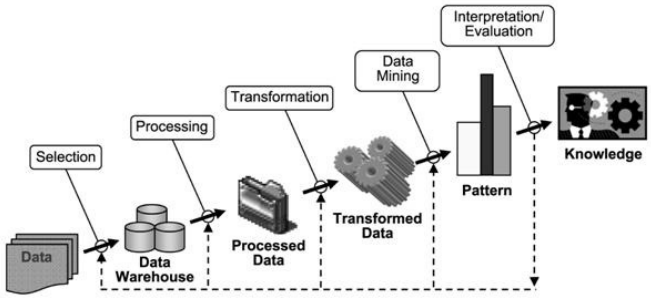


Figure 1: KDD Process

**i) Data cleaning:** This is the first step to eliminate noise data and irrelevant data from collected raw data.[2]

**ii) Data integration:** At this step, various data sources are combined into meaningful and useful data.

**iii) Data Selection:** Here, data relevant to the analysis are retrieved from various resources.

**iv) Data transformation**: In this step, data is converted or consolidated into required forms for mining by performing different operations such as smoothing, normalization or aggregation.

**v) Data Mining**: At this step, various clever techniques and tools are applied in order to extract data pattern or rules.[2]

**vi) Pattern evaluation:** At this step, Attractive patterns representing knowledge are identified based on given measures.

**vii) Knowledge representation:** This is the last stage in which, visualization and knowledge representation techniques are used to help users to understand and interpret the data mining knowledge or result.[2]

The goal of knowledge discovery and data mining process is to discover the patterns that are unknown among the huge set of data and interpret useful knowledge and information.

## 3. DATA MINING TECHNIQUES

Data mining process is extraction of information from large data sets and transforms it into some understandable form for further uses. So it helps to achieve the specific objectives. The goal of a data mining effort is normally either to create a descriptive model or a predictive model. A **Descriptive model** presents the data in concise form which is essentially a summary of the data points, finds patterns in the data and understands the relationships between attributes represented by the data. The Descriptive model includes tasks such as Clustering, Association Rules, Summarizations, and Sequence Discovery. The **predictive model** works by making a prediction about values of data, which uses known results found from different data sets .

### 3.1 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. This approach frequently employs decision tree or neural network-based classification algorithms.[3] The common characteristics of classification tasks are as supervised learning, categories dependent variable and assigning new data to one of a set of well-defined classes. Classification technique is used in customer segmentation, modeling businesses, credit analysis, and many other applications. E.g., classify countries based on population, or classify bikes based on mileage.

### 3.2 Regression:

Regression is another Predictive data-mining model is also known as supervised learning technique. Basically, regression takes a numerical dataset and develops a mathematical formula that fits the data.[5] This technique analyzes the dependency of some attribute values, which is dependent upon the values of other attributes mainly, present in same item. In the regression techniques target value are known. For example, you can predict the child's behavior based on family history. [4]

### 3.3 Time Series data analysis:

Time-series database uses sequences of values or events obtained over repeated measurements of time. The values are typically measured at equal time interval such as hourly, daily, weekly. A sequence database is any database that consists sequence of ordered events, sometimes having concrete notions of time. For example, Web page traversal sequences and customer shopping transaction sequences are sequence data, but they may not be time-series data.

### 3.4 Prediction:

This technique discovers the relationship between independent variables and the relationship between dependent and independent variables.[6] The prediction is to predict a future state, rather than a current one. Its applications include obtaining forewarning of natural disasters (flooding, hurricane, snowstorm, etc), epidemics, stock crashes, etc. As another example, the sales volume of computers accessories can be forecasted based on the number of computers sold in the past few months.

### 3.5 Clustering:

Clustering is a collection of similar data objects. Dissimilar object is another cluster. It is way finding similarities between data according to their characteristic.[4] Clustering can be considered 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 and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but this method is expensive so clustering can be used as preprocessing approach for attribute subset selection and classification.

For example, image processing, pattern recognition, city planning. Astronomy - aggregation of stars, galaxies, or super galaxies.

**3.6 Summarization**:

Summarization is referred as the abstraction or generalization of data.[12] The summarization technique maps data into subsets with simple descriptions. The summarized data set gives general overview of the data with aggregated information. Simple summarization methods such as tabulating the mean and standard deviations are often applied for data analysis, data visualization and automated report generation.

For example: length can be summarized as meters, centimeters or millimeters.

**3.7 Association**:

The Association technique is used to extract the relationships between attributes and items. In this technique, the presence of one model implies the presence of another model i.e. item is related to another in terms of cause-and-effect. This is common in establishing a form of statistical relationships among different interdependent variables of data mining; association rules are useful for analyzing and predicting customer behavior. They also play an important role in shopping basket data analysis, product clustering, catalog design and store layout. The association rules are also build by programmers can be used to build programs capable of machine learning.

**3.8 Sequence Discovery**: Uncovers correlation among data. It is set of object each associated with its own timeline of events. For example, scientific experiment, natural disaster and analysis of DNA sequence.

## 4. DATA MINING APPLICATIONS

The Data mining applications are widely used in diverse areas such as retail stores, hospitals, banks, and insurance companies. Many domains like health care, finance insurance, retail stores combines the data mining applications with statistics, pattern recognition, and other important tools to perform data analytics. Data mining is used primarily for decision making.

**i. Medicare and health care:** Data mining in medicine enables to characterize patient activities to see incoming office visits. Data mining helps identify the patterns of successful medical therapies for different illnesses. Data mining is used to predict the volume of patients in future and for preventing from disease we will aware about it.[7]

**ii. Education*:*** Educational Data Mining is a blooming field which provides knowledge from educational Environment data. The goals of EDM are identified as predicting students' learning behavior, emotions and skills.[8] This study improves the educating methods by understanding the ward and to take accurate decisions respectively

**iii. Market Basket Analysis:** Market basket analysis is a technique that uses association rule mining to understand the purchasing behavior of the customer. It also allows the seller to understand his business, customer's needs and to make profitable change accordingly. The ultimate goal of market basket analysis is finding the products that customers frequently purchase together.[11]

**iv. Financial Banking:** Data mining can contribute to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information and market prices. The managers may find this information for better segmenting, targeting, acquiring, retaining and maintaining a profitable customer.

**v. Research Analysis:** Data mining is very useful in data pre-processing and integration of databases. Data mining allows the researchers to identify co-occurring sequences and the correlation between any activities .Data visualization and visual data mining help the researcher with a clear view of the data.

**vi. Fraud Detection:** The traditional fraud detection methods are expensive, time consuming and complex. Data mining aids in providing meaningful patterns and turning data into information. Valid and useful information is called as knowledge. The results are categorized into fraudulent or non-fraudulent.

**vii. Transportation:** Data mining helps determine the distribution schedules among warehouses and outlets and analyze loading patterns.

**viii. Agriculture:** Data mining is emerging technology in agriculture field for crop yield analysis 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).[9]

**ix. 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.[10] Cloud computing uses the Internet services that rely on clouds of servers to handle tasks. The data mining technique in Cloud Computing helps to perform efficient, reliable and secure services for their users.

## 5. CONCLUSION

According to the techniques of data mining listed above, it is learned that this a powerful and essential technique for performing manipulation of data that is data mining gives proper and targeted outcome from large and vastly growing data worldwide. This paper discusses the idea of data mining, the process of KDD, different techniques such as clustering, association, classification, prediction and so on. We also discussed some insights of the data mining applications.

## REFERENCES

[1] Tipawan Silwattananusarn, Dr.Kulthida Tuamsak "Data Mining and its Application for Knowledge Management: A Literature review from 2007 to 2012", IJDKP: International Journal of Data mining and Knowledge Management Process, Vol.2, Issue.5, PP.13-24, Sep-2012.

[2] Aarti Sharma, Rahul Sharma, Vivek Kr.Sharma, Vishal Shrivatava, "Application of Data Mining – A survey Paper" IJCSIT : Internationl Journal of Computer Science & Information Teachers, Vol.5, Issue.2, PP.2023-2025, 2014.

[3] K.Prema, A.Kumar Kombaiya "A survey on use of Data Mining Methods Techniques and Application, IJARSE: International Journal of Advanced Research in Science and Engineering, Vol.6, Issue.12, PP.532-539, Dec-2017.

[4] Smita, Priti Sharma, "Use of Data Mining in Various Field: A survey Paper", IOSR-JCE: IOSR Journal of Computer Engineering, Vol.16, Issue.3, PP.18-21, May-Jun 2014.

[5] Hemlata Sahu, Shalini Shrma, Seema Gondhalakar, "A Brief Overview on Data Mining Survey",IJCTTE: International Journal of Computer Technology and Electronics Engineering, Vol.1, Issue.3, PP.114-121.

[6] Aakanksha Bhatnagar, Shweta P.jadye, Madan Mohan Nagar, "Data Mining Techniques & Distinct Applications: A Literature Review" IJERT: International Journal of Engineering Research & Technology, Vol.1, Issue.9, PP.1-3, Nov-2012.

[8] Brijesh Kumar Baradwaj, Saurakh Pal, "Mining Educational Data to Analyze Students Performance", IJACSA: International Journal of Advanced Computer Science and Applications, Vol.2, Issue.6, PP.63-69, 2011.

[9] D Ramesh, B Vishnu Vardhan, "Data Mining Techniques and Applications to Agricultural Yield Data", IJARCCE: International Journal of Advanced Research in Computer and Communication Engineering, Vol.2, Issue.9, PP.3477-3480, Sep-2013.

[10] Ruxandra – Stefania PETRE, "Data Mining in Cloud Computing", Database Systems Journal, Vol.III, Issue.3, PP.67-71, 2012.

[11] Arun K Pujari, "Data Mining Techniques", Universities Press, India, PP.98-120, 2017.

[12] David L Olson, Dursun Delen, "Advanced Data Mining Techniques", Springer, PP.55-57, 2017.