© 2018 IJSRSET | Volume 4 | Issue 11 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology
DOI : https://doi.org/10.32628/IJSRSET1841079

An Overview of Data Mining Techniques and Applications and its Future Scope

Nithya C , Saravanan V
Department of Computer Science, Hindusthan College of Arts And Science, Coimbatore, India

ABSTRACT

Data mining is a process which finds useful patterns from large amount of data. The paper discusses few of the data mining techniques, algorithms and some of the organizations which have adapted data mining technology to improve their businesses and found excellent results. The greater part of data mining methods can manage distinctive information sorts.Data mining may be defined as the science of extracting useful information from databases. It also called knowledge discovery. Using a combination of machine learning, statistical analysis, modeling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow the prediction of future.

Keywords : Data Mining Techniques; Data Mining Algorithms; Data Mining Applications.

I. INTRODUCTION

The primary aim of data mining is to extract the useful information for users from a large amount of data. 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. 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. Data mining is frequently recognized to be "A mix of detail, Ai and Database research.

The aim of this study is covered:

1) Provide an overview of existing techniques that can be used for extracting of useful information from databases.

2) Provide a feature classification technique that identifies important aspects to study knowledge discovery.

3) Investigate existing knowledge discovery and data mining software tools using the proposed feature classification scheme.

OVERVIEW OF DATA MINING

The development of Information Technology has generated large amount of databases and huge data in various areas. The research in databases and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data pattern analysis, typically deals with data that have already been collected for some purpose rather than the data mining analysis. This means that the objectives of data mining exercise play no role in the data collection strategy. The data sets examined in data mining are often large.

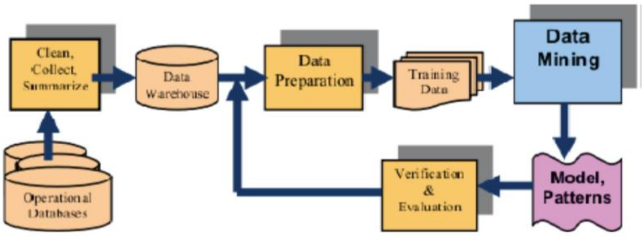
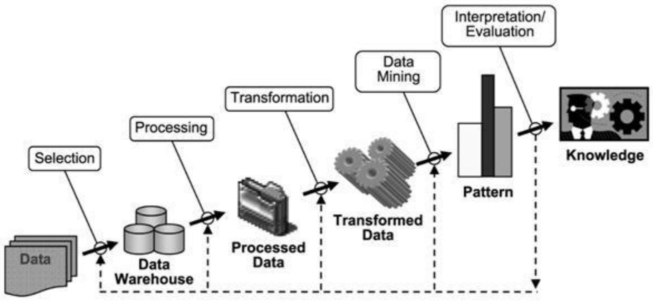


Figure 1: The KDD (Knowledge Discovery Process) and data mining process (Han & Kamber, 2002)

II. METHODS AND MATERIAL

DATA MINING PROCESS

Data mining is also known as Knowledge Discovery in Database, refers to finding or “mining" knowledge from large amounts of data. 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 [1]. In Data mining, Knowledge extraction or discovery is done in seven sequential steps as in Fig 1.



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

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.

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.

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.

DATA MINING, MACHINE LEARNING AND STATISTICS :

Data mining takes advantage of advances in the fields of artificial intelligence (AI) and statistics.

Both disciplines have been working on problems of pattern recognition and classification. Both communities have made great contributions to the understanding and application of neural nets and decision trees.

Data mining does not replace traditional statistical techniques. Rather, it is an extension of statistical methods that is in part the result of a major change in the statistics community. The development of most statistical techniques was, until recently, based on elegant theory and analytical methods that worked quite well on the modest amounts of data being analyzed. The increased power of computers and their lower cost, coupled with the need to analyze enormous data sets with millions of rows, have allowed the development of new techniques based on a brute-force exploration of possible solutions.

New techniques include relatively recent algorithms like neural nets and decision trees, and new approaches to older algorithms such as discriminant analysis. By virtue of bringing to bear the increased computer power on the huge volumes of available data, these techniques can approximate almost any functional form or interaction on their own. Traditional statistical techniques rely on the modeler to specify the functional form and interactions. The key point is that data mining is the application of these and other AI and statistical techniques to common business problems in a fashion that makes these techniques available to the skilled knowledge worker as well as the trained statistics professional. Data mining is a tool for increasing the productivity of people trying to build predictive models.

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[5]. 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 datasets [3]. The Predictive data mining model includes classification, prediction, regression and analysis of time series as in figure 2

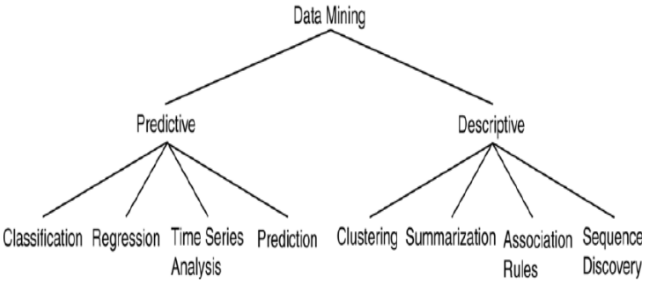


Figure 2 Data Mining Techniques

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 [2]. This approach frequently employs decision tree or neural network-based classification algorithms. 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.

REGRESSION : Regression is another Predictive data-mining model is also known as supervised learning technique. 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.

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[4]. For example, Web page traversal sequences and customer shopping transaction sequences are sequence data, but they may not be time-series data.

PREDICTION : This technique discovers the relationship between independent variables and the relationship between dependent and independent variables. The prediction is to predict a future state, rather than a current one [4]. 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.

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 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,

SUMMARIZATION : Summarization is referred as the abstraction or generalization of data. 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.

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.

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.

DATA MINING TYPES

Predictive Data Mining: It produces the model of the 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, non trivial information based on the available data set. It focuses on finding patterns describing the data that can be interpreted by humans.

DATA MINING TASKS

- Data processing [descriptive]
- Prediction [predictive]
- Regression [predictive]
- Clustering [descriptive]
- Classification [predictive]
- Link analysis/ associations [descriptive]
- Evolution and deviation analysis [predictive]

A. Classification

It predicts categorical class labels (nominal or discrete). Learning a function that maps an item into one of a set of predefined classes. It classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data.

B. Regression

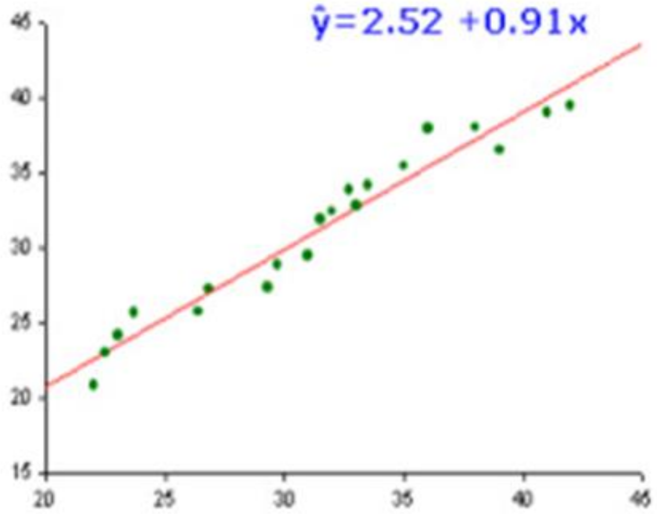
It is a statistical process for estimating or predicting the relationships among items or variables. It includes many techniques for analyzing and modeling of several variables, when focuses on the relationship between a dependent variable and one more independent variables.

Linear Regression

 w0 + w1 x + w2 y >= 0

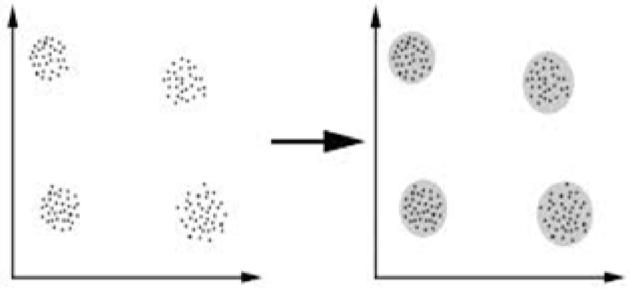
 It computes wi from data item to minimize squared error to 'fit' the data

 Not flexible enough



C. Clustering

The process of identify or grouping a set of physical objects or items into classes of similar objects or we can say that identify a set of groups of similar items.



D. Dependencies And Associations

Identify the significant dependencies between data attributes.

Find those attributes in which dependency are occurred and associates them to each other.

E. Summarization

Find a summarized or compact description of the dataset or a subset of the dataset.

III. DISCUSSION

APPLICATION OF DATA MINING

A. Spatial Data Mining

Spatial data mining refers to the extraction of knowledge,spatial relationships, or other interesting patterns not explicitly stored in spatial databases.A spatial database stores a large amount of space-related data, such as maps, preprocessed remote sensing or medical imaging data, and VLSI chip layout data.

Spatial Data Cube Construction and Spatial OLAP

As with relational data, we can integrate spatial data to construct a data warehouse that facilitates spatial data mining. A spatial data warehouse is a subject-oriented, integrated, time variant and nonvolatile collection of both spatial and non spatial data in support of spatial data mining and spatial-data-related decision-making processes.

There are three types of dimensions in a spatial data cube:

a) A non spatial dimension

b) spatial-to-non spatial dimension

c) spatial-to-spatial dimension

B. Spatial Clustering Methods

Spatial data clustering identifies clusters, or densely populated regions, according to some distance measurement in a large, multidimensional data set. Spatial Classification and Spatial Trend Analysis Spatial classification analyzes spatial objects to derive classification schemes in relevance to certain spatial properties, such as the neighborhood of a district, highway, or river.

C. Mining Raster Databases

Spatial database systems usually handle vector data that consist of points, lines, polygons (regions), and their compositions, such as networks or partitions. Typical examples of such data include maps, design graphs, and 3-D representations of the arrangement of the chains of protein molecules.

D. Multimedia Data Mining

A multimedia database system stores and manages a large collection of multimedia data, such as audio, video, image, graphics, speech, text, document, and hypertext data, which contain text, text markups, and linkages Similarity Search in Multimedia Data When searching for similarities in multimedia data, we can search on either the data description or the data content approaches:

a) Color histogram–based signature
b) Multifeature composed signature
c) Wavelet-based signature

E. Multidimensional Analysis of Multimedia Data

To facilitate the multidimensional analysis of large multimedia databases, multimedia data cubes can be designed and constructed in a manner similar to that for traditional data cubes from relational data. A multimedia data cube can contain additional dimensions and measures for multimedia information, such as color, texture, and shape.

F. Classification and Prediction Analysis of Multimedia Data

Classification and predictive modeling can be used for mining multimedia data, especially in scientific research, such as astronomy, seismology, and geoscientific research.

G. Mining Associations in Multimedia Data

a) Associations between image content and nonimage content features.
b) Associations among image contents that are not related to spatial relationships.
c) Associations among image contents related to spatial relationships.

H. Audio and Video Data Mining

An incommensurable amount of audiovisual information is becoming available in digital form, in digital archives, on the World Wide Web, in broadcast data streams, and in personal and professional databases, and hence a need to mine them.

I. Text Mining

Text Data Analysis and Information Retrieval Information retrieval (IR) is a field that has been developing in parallel with database systems for many years. Basic Measures for
Text Retrieval: Precision and Recall.

J. Mining the World Wide Web

The World Wide Web serves as a huge, widely distributed,global information service center for news, advertisements,consumer information, financial management, education,government,
e-commerce, and many other information services. The Web also contains a rich and dynamic
collection of hyperlink information and Web page access and usage information, providing rich sources for data mining.

K. Scientific Applications

Data collection and storage technologies have recently improved, so that today, scientific data can be amassed at much higher speeds and lower costs. This has resulted in the accumulation of huge volumes of high-dimensional data, stream data, and heterogeneous data, containing rich spatial and temporal information. Consequently, scientific applications are shifting from the ―hypothesize-and-test paradigm toward a ―collect and store data, mine for new hypotheses, confirm with data or experimentation‖ process.This shift brings about new challenges for data mining

L. Data Mining for Intrusion Detection

The security of our computer systems and data is at continual risk. The extensive growth of the Internet and increasing availability of tools and tricks for intruding and attacking networks have prompted intrusion detection to become a critical component of network administration.

M. Visual and Audio Data Mining

Visual data mining discovers implicit and useful knowledge from large data sets using data and/or knowledge visualization techniques.

In general, data visualization and data mining can be integrated in the following ways:

a) Data visualization
b) Data mining result visualization
c) Data mining process visualization
d) Interactive visual data mining

N. Data Mining and Collaborative Filtering

A collaborative filtering approach is commonly used, in which products are recommended based on the opinions of other customers. Collaborative recommender systems may employ data mining or statistical techniques to search for similarities among customer preferences.

O. Security of Data Mining

Data security–enhancing techniques have been developed to help protect data. Databases can employ a multilevel security model to classify and restrict data according to various security levels, with users permitted access to only their authorized level. Privacy sensitive data mining deals with obtaining valid data mining results without learning the underlying data values.

LIMITATIONS OF DATA MINING

Data mining requires skilled technical and analytical specialists who can structure the analysis and interpret the output that is created. Consequently, the limitations of data

mining are primarily data or personal related, rather than technology-related.

Data mining can help reveal patterns and relationships, it does not tell the user the value or significance of these patterns.

Another limitation of data mining is that while it can identify connections between behaviors and/or variables, it does not necessarily identify a causal relationship.

FUTURE SCOPE

Data mining extract the useful information and provide accurate data for decision making. Data mining help in predict the future trends and its behavior for the business purpose. It extract valuable business information in a large database for example, finding linked products in gigabytes or terabytes of store scanner data .Given databases of sufficient size and good quality, data mining technology can generate new decision making business opportunities by providing these capabilities.

Data Mining automates the process of finding predictive information from large data bases. It uses the current or past promotional mailings data to identified the most likely to maximize the return on investment on future mailings. On the other hand Data Mining technologies detect the fraud detection and identifying segments of population likely to respond of similar events or task.

A) Artificial Neural Networks

It is Non-linear predictive models that learn through training set and resemble biological neural networks in structure.

B) Decision Trees

Tree-shaped type structures that represent sets of decisions. Each node is judgment and separately represent for the decision. These decisions generate rules for the classification of a dataset.

c) Genetic algorithms

Genetic algorithm is optimization techniques that use process such as genetic combination, natural selection and mutation in a design based on the concepts of evolution.

d) K-Nearest neighbor method

K-Nearest Neighbors algorithm is a non-parametric method used for classification and regression. It is technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset.In Future, It may be happed that the knowledge is not provided by the database, or any techniques used to retrieve a minimized data, The input itself taken by the technique. Monitoring the use of user or requested data by the system,it predicts and retrieves whatever the data is needed by the user. Here the kind of data wants to retrieve it predicted by techniques and executed in background.

IV. 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. Data mining applications can use a variety of parameters to examine the data as an application, compared to other data analysis applications, such as structured queries or statistical analysis Software, data mining represents a difference of kind rather than degree. Data mining involves the use of data analysis tools to discover previously unknown, valid patterns and relationships in large data sets Data mining is becoming increasingly common in both the private and public sectors. Data mining applications in various fields use the variety of data types. The different methods of data mining are used to extract the patterns and thus the knowledge from this variety databases. Efficient and effective data mining in large database poses numerous requirements and great challenges to researchers and developers. The dramatically increasing demand for better decision support is answered by an extending availability of knowledge discovery, and data mining is one step at the core of the knowledge discovery process.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 information from the unimaginable massive data. This review would be helpful for the researchers to focus on the various issues 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.

V. REFERENCES

[1]. Aarti Sharma et al, “Application of Data Mining – A Survey Paper”, International Journal of Computer Science and Information technologies’, Vol. 5 (2), 2014.

[2]. Brijesh Kumar Baradwaj, Saurabh Pal" Mining Educational Data to Analyze Students Performance" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011

[3]. Nikita Jain, Vishal Srivastava "DATA MINING TECHNIQUES: A SURVEY PAPER" IJRET: International Journal of Research in Engineering and Technology, Volume: 02 Issue: 11 | Nov-2013.

[4]. Prof. Dr. Wolfgang Karl Hardle," Time Series Data Mining Methods: A Review", Berlin, March 25, 2015.

[5]. Pradnya P. Sondwale, "Overview of Predictive and Descriptive Data Mining Techniques" IJARCSSE, Volume 5, Issue 4, April 2015

[6]. Neelamadhab Padhy, Dr. Pragnyaban Mishra "The Survey of Data Mining Applications And Feature Scope" International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012

[7]. Harshna, NavneetKaur, "Survey paper on Data Mining techniques of Intrusion Detection", International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013.

[8]. https://www.google.co.in/search?q=kdd+process

[9]. https://www.google.co.in/search?q=forms+of+data+preprocessing.