An Analysis on Different Methods of Data Mining Techniques with Its Purposes and Issues

Rupinder Kaur
ResearchScholar
GuruGobindSinghCollegeofModernTechnology
,Kharar, Punjab
itsrupinderkaur@gmail.com

Harpreet Kaur
AssistantProfessor,CSEDepartment
GuruGobindSinghCollegeofModernTechnology,
Kharar, Punjab
Billing02harpreet@gmail.com

Abstract— It is not wrong to say that the advancement in the technology has a great impact on the human society. It has been altered the way of doing business, providing and receiving the services, managing the organizations etc. The most direct effect is the completed change of 3 information collection, conveying, and exchange. Data mining is an emerging domain that is specifically applied to extract the meaningful data from the large amount of available contents. Various authors have been developed the various prominent methods to for data mining.

This study generates a review to the data mining and along with this a brief introduction to the knowledge discovery data base is also given in this work. Difference data extraction processes are cover under this by author.

Keywords—*Data Mining, Data Extraction, Pattern Matching.*

*****

I. INTRODUCTION

Data Mining is a technique which is used for extracting the meaningful data. The records are kept in a huge database for the purpose of future use**.** Data Warehouse can assemble all diversity of data. Frequently the group of data depends ahead the variety of industries for which it is being used [1]. The mainstream of the industries preserves indication of each type of data. Although, some companies only accumulates that information which is precious and significant for them. The data stored in a warehouse are helpful in decision support system [2]. On the basis of historical data the decision regarding the future schemes can be taken easily or much effectively. Data mining is an automatic progression that removes meaningful information from data storage and uses this deleted information for different principles. Extracting meaningful data can be performed by matching patterns and is accomplished by cluster analysis [3,4], anomaly analysis, and dependency analysis. Spatial indexes are used to perform all the above functions or processes. The matched pattern is a form of a brief summary of the data stored in the data warehouse and these patterns is used for various decision making systems to make future predictions and make the right decisions [5]. For example, in the case of a machine learning system, this extracted information can be used for predictive analysis. In another illustration, data mining is the procedure of perceiving or inspecting a variety of groups of correlation data in a database that can be further used for predictive analysis in the near future [6]. Data examination, data gathering, and data anthology are not associated with data mining, but are still incorporated in the KDD process, knowledge discovery [7]. However, they all run as options or additional steps of KDD. Terms such as data dredging, data snooping, etc. use data mining to reveal pattern matching of large amounts of data stored in the warehouse [8].

II. KNOWLEDGE DISCOVERY DATABASE

The data mining process is based upon the KDD. It is used for Knowledge Discovery in Database. The KDD process includes various stages. The number of stages included in this is 5. These are as follows [9]:

(1) Selection
(2) Pre-processing
(3) Transformation
(4) Data Mining
(5) Interpretation/Evaluation.

Data mining has various problems, it is the most desirable research field. There are many problems related to data mining, such as hypothetical problems [10]. There are some problems interrelated to the sensible accomplishment of mining, such as exploration of interesting knowledge and unknown knowledge obtained from the actual database. Below is a list of key issues or problems related to data mining with corresponding solution [11].

1. Vast data set and high dimensional.
2. Fitting and evaluation of statistical significance.
3. Understanding of patterns.
4. Substandard imperfect data and data incorporation.
5. Changes and redundant data are mixed.

Data mining is the process of extracting interesting information from a large amount of available data sets. Data mining specifically performs the following tasks [12]:

For Directed data mining, classification, estimation and prediction are performed. Directed data mining is a term that illustrates the procedure of using specific data in a database to produce a sample or representation that defines one or more significant characteristic compared to the rest of the features [13]. Association, clustering, and description define the process of undirected data mining. Undirected data mining is the process by which relationships between attributes are developed.

a) Classification

Classification is a process performed in order to evaluate the characteristics of a given object. These properties are assigned to an existing object. Classification is completed with classes, with training sets including classes or reclassified examples [14]. The main purpose of the classification is to classify the pattern or data that have not been classified. Illustrations of classification are as follows.

- Classification of credit candidates based on squat, intermediate, or soaring threat.
- Branching and dangerous classification of vegetables and fruits like mushrooms.
- Identify the telephone line connected to the Internet.

b) Estimation

Quotation is the subsequent progression or assignment to sprint after classification. In this case, the output is estimated supports on the input pattern. The estimation is achieved based on a given input constraint [15]. These variables or constraints are unidentified. An example of the assessment procedure is as follows.

- Commencing the mother's prerequisite as an input inconsistent to guesstimate the number of family children.
- Establishment on the number of vehicles in the house approximation unpleasant proceeds of that house.

c) Prediction

A prediction is a process that can be regarded as either classification or estimation. This is a process that predicts results based on historical [16] behavior and future values. An example of the forecasting process is as follows.

- Based on buying customer behavior to predict whether to leave in the near future.
- On the basis of user’s actions and stimulated plans predicting that which customer would like to have value added services to his association.

d) Association Rules

Association rules are used to describe the relationship between a series of objects. It defines how various objects are linked or associated with each other in the database. To understand related rules, let's consider an example. Each transaction of a particular transaction is made up of items and X and Y are two data items. These relationships can be described because X contained in the database tends to contain Y [17].

e) Clustering

Clustering is a technique of dividing data into subclasses. These subclasses are called clusters. These clusters contain only relevant data sets. This makes it easy to select data as necessary [18]. Thus, the data are divided into category clusters according to the nature of the data set. There are many algorithms used to divide the data into various clusters. Among the algorithms used for clustering are separating techniques, hierarchical system, and grid-based clustering methods. There is a discrepancy between classification and clustering, but clustering is not based on predefined data sets or variables, but the classification of the data is a predefined value.

f) Regression

Regression is a statistical measure which is used to fit an equation to the dataset. The equation used for calculating the linear regression is as follows:

$$y = mx + b \ldots\ldots(1)$$

The equation determines the value of m and b with respect to the given value of x and output is the value of y. The Multiple regression calculation is an enhanced or advanced technique which can take more than one input and lead to the more complex equation such as quadratic equation [19].

g) Time Series Analysis

Time series analysis is a method for time series based data for extracting the meaningful statistical facts or data. This model produces the future events on the basis of past events.

h) Summarization

Summarization is process in which the data of the database. In this the data is summarized into smaller groups as the overview of whole database. The summarized data contains the aggregate data or information.

i) Sequence Discovery

Sequence discovery is a data mining model in which statistically relevant patterns are recognized. It is implemented in a case where the deliveries of the values are in a sequence [20].

III. RELATED WORK

Dr. T. Karthikeyan(2012), [2] This paper describes a technique based on Ant colony Optimization (ACO) known as Ant Miner. To discover rules in the database MAX-MIN ant system was optimize through Ant-Miner+. As a result, soil classification had been done based on different characteristics of different category of soil. In this paper, proposed technique was used to compare Ant miner and Ant miner+ algorithm, where evaluation had been performed on training and soil dataset to association rule. Results shows that Ant-Miner + was a better approach than Ant miner.

NeelamadhabPadhy (2012), [3]Data mining had been used in various fields for several purposes due to its extraction of useful knowledge property. Thus in this paper, prediction of diseases in health care industry had been described where medical data was extracted through data analysis tools so that useful knowledge can be accessed and used. As medical data was flourishing source of information due to which it was mandatory to keep it up to date and extraction of data at the right time was also an important part of medical industry. Thus in this project the main focus was on the medical decision based upon the symptoms of the diseases that look similar and rare so that a good decision can take by the doctors. Diagnosis Decision Support System that can take patients data and can result into an appropriate prediction. This system was used to extract veiled knowledge from a heart disease database recorded earlier. This model was also helpful in answering the complex queries based on its own strength.

U. K. Pandey, and S. Pal (2011), [4]Due to the advancement in technology, there is a large number data types or databasewhich can be used efficiently or in meaningful way. Student's related database is a huge dataset and can be used to explore the status of the students and level of the universities. This data is processed by using the data mining. In this paper author introduced a classification technique i.e. Bayes classification.

Michael J. Shaw (2001),[5] had provided the systematic approach to use the data mining and information management methods in order to manage the marketing information and support to take the decisions related to marketing. This approach which had been proposed in the paper can form the basis for improving the CRM. Now the businesses have increased level of ability to store the large amount of data in large databases because of the development and advancement in Information system and technology. On the other hand, most of the functional marketing insights for the customer characteristics and their trend of purchase were mostly not visible and unidentified. Present emphasis of CRM makes the function of marketing an ideal area of application to highly benefit from the implementation of data mining techniques to support the decision making.

Hongiun Lu (1996), [6]nowadays main focus of data base community was on classification technique which is one of the major problem of data mining. In this paper a new approach had been described to identify the symbolic classification rules by implementing the neural networks. Earlier neural networks were not used for data mining and it was because the criteria of classification were not stated explicitly as symbolic rules. The symbolic rules are suitable for authentication or interpretation by human. By implementing the proposed technique precise symbolic rules with high level of accuracy can be obtained from the Neural Networks. Firstly it was required to train the network to obtain the required precision rate. After this the next step was to eliminate the redundant connections in the network by implementing pruning paradigm. Then it was required to determine the activation values of the hidden units in the network, and finally the rules for the classification were developed by implementing the results of analysis. The results obtained by implementing the proposed technique it was observed that the technique was optimum.

Quang yang (2006),[6] identified ten challenging problems in the field of data mining. These problems have been finding out by initializing survey and consulting active research in the domain of data mining as well as machine learning. Based on their opinions and according to the worthy topics for future research 10 problems have identified. The primary factor of introducing these topics to provide high level guidelines for future researchers. This paper concluded ten most challenging problems in the domain of data mining and 14 received responses from this survey has also identified.

IV. CONCLUSION

The appearance of data mining locates for just mining of data from the data warehouse while it also achieves pattern matching and also realizes awareness of the vast amount of data. Data mining is a term repeatedly used in large-scale data processing and information processing sites, and data collection includes processes such as data collection, information extraction, and data warehouse analysis. in future more research can be organized to perform pattern matching and data extracting in an effective manner by using advanced techniques.

REFERENCES

[1] Anand V. Saurkar, "A Review Paper on Various Data Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 4,Pp 98-101, 2014.

[2] Dr. T. Karthikeyan,"A Study on Ant Colony Optimization with Association Rule" International Journal of Advanced Research in Computer Science and Software Engineering, 2012.

[3] NeelamadhabPadhy "The Survey of Data Mining Applications And Feature Scope" International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3,Pp 43-58, 2012.

[4] U . K. Pandey, and S. Pal, "Data Mining: A prediction of performer or underperformer using classification", (IJCSIT) Vol. 2(2), Pp.686-690, 2011.

[5] Michael J. Shaw, "Knowledge management and data mining for marketing", Elsevier, V. 32, Pp. 127-137, 2001.

[6] Hongiun Lu, "Effective Data Mining Using Neural Networks", IEEE, V. 8, N. 6, Pp. 957-962, 1996.

[7] Quang yang, "10 Challenging problems in Data Mining research", world Scientific, V. 5, N. 4, Pp. 597-604, 2006.

[8] D.A. Adeniyi , "Automated Web usage data mining and recommendation system using K-Nearest Neighbor classification method ", Applied Computing and Information, Vol 12, Pp 90-108, 2016.

[9] AnithaTalakokkula, "A Survey on Web Usage Mining, Applications and Tools", Computer Engineering and Intelligent System, Vol 6, Issue 2, Pp 22-30, 2015.

[10] Bo Cheng, Shuai Zhao, Changbao Li, Junliang Chen, "A Web Services Discovery Approach Based on Mining

Underlying Interface Semantics”, IEEE, Vol 29, Pp 950-962, 2017.

[11] SatyaPrakash Singh , Meenu, “Analysis of web site using web log expert tool based on web data mining”, IEEE, 2017,

[12] Yeqing Li, “ Research on Technology, Algorithm and Application of Web Mining”, IEEE, Vol 1, Pp 772-775, 2017.

[13] Z. A. Usmani, Saiqa Khan, Mustafa Kazi, AadilBhatkar, ShuaibShaikh, “ZAIMUS: A department automation system using data mining and web technology”, IEEE, Pp 1-6, 2017.

[14] Martin Lnenicka , Jan Hovad , JitkaKomarkova , MiroslavPasler, “A proposal of web data mining application for mapping crime areas in the Czech Republic”, IEEE, 2016.

[15] Viktor Medvedev, Olga Kurasova, GintautasDzemyda, “A new web-based solution for modelling data mining processes”, ELSEVIER, Vol 76, Pp 34-46, 2016.

[16] PetarRistoski, HeikoPaulheim, “Semantic Web in data mining and knowledge discovery: A comprehensive survey”, ELSEVIER, Vol 36, Pp 1-22, 2016.

[17] VenkataSubba Reddy Poli, “Fuzzy data mining and web intelligence”, IEEE, 2016.

[18] ZoltánBalogh, “Data-mining behavioural data from the web”, IEEE, Pp 122-127, 2016.

[19] Suvarn Sharma, AmitBhagat, “Data preprocessing algorithm for Web Structure Mining”, IEEE, Pp 94-98, 2016.

[20] Wang Lei , Liu Chong, “Implementation and Application of Web Data Mining Based on Cloud Computing”, IEEE, 2016.

[21] D. BavarvaBhaskar , Dheeraj Kumar Singh, “Multimedia questions and answering using web data mining”, IEEE, 2015.