Dogo Rangsang Research Journal
ISSN : 2347-7180
UGC Care Group I Journal
Vol-08 Issue-14 No. 01 February : 2021

A Brief Overview on Data Mining

Mr. Narottam sahu, Mrs.PRAGYAN PARAMITA PANDA*
Dept. OF Computer Science and Engineering, NIT , BBSR
chinmayaranjan@thenalanda.com, pragyanparamita@thenalanda.com*

Abstract: *In this paper, the idea of data mining was summarized and its significance towards its methodologies was illustrated. In the information Technology era info plays important role in each sphere of the human life. It's important to collect knowledge from completely different data sources, store and maintain the data, generate info, generate knowledge and circularize knowledge, info and information to each stakeholder. Because of large use of computers and electronics devices and tremendous growth in computing power and storage capability, there's explosive growth in data collection. Data mining is the notion of all strategies and techniques which permit analyzing very massive data sets to extract and find out previously unknown structures and relations out of such vast heaps of details. This paper provides the complete review regarding the data mining and its techniques.*

I. INTRODUCTION

Data mining refers to extracting or mining the data from great amount of data. The term data mining is suitably named as „Knowledge mining from data‟ or "Knowledge mining". Data collection and storage technology has created it attainable for organizations to accumulate vast amounts of data at lower price. Exploiting this hold on knowledge, so as to extract useful and actionable data, is the overall goal of the generic activity termed as data mining. The subsequent definition is given:

Data mining is that the method of exploration and analysis, by automatic or semiautomatic means, of huge quantities of data so as to find meaningful patterns and rules.

The data mining tasks are of various varieties depending on the usage of data mining result the data mining tasks can be classified as[1,2]:

1. Exploratory data Analysis: it's merely exploring the data without any clear concepts of what we are searching for. These techniques are interactive and visual.
2. Descriptive Modeling: It describe all the data, It includes models for overall change of distribution of the data, partitioning of the p-dimensional space into groups and models describing the relationships between the variables.
3. Predictive Modeling: This model permits the value of 1 variable to be foreseen from the best-known values of other variables.
4. Discovering Patterns and Rules: It concern with pattern detection, the aim is recognizing fallacious behavior by detecting regions of the space defining various kinds of transactions where the data points considerably different from the remaining.
5. Retrieval by Content: it's finding pattern almost like the pattern of interest in the data set. This task is most typically used for text and image data sets.

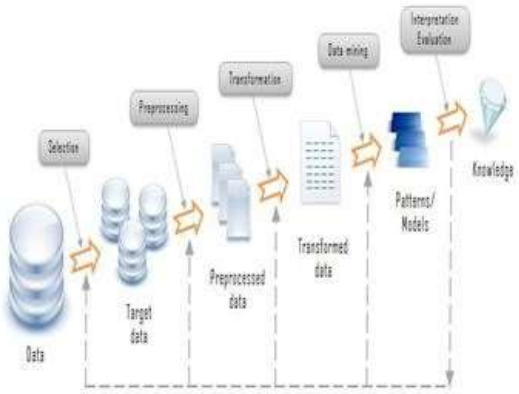

Fig 1. Data Mining

II. DATA MINING LIFE CYCLE

The life cycle of a data mining project consists of six phases [2,4]. The sequence of the phases isn't rigid. Moving back and forth between completely different phases is usually required. It depends on the result of every phase. The main

phases are:

1. Business Understanding: This phase focuses on understanding the project objectives and necessities from a business perspective, then changing this knowledge into a data mining problem definition and a preliminary plan designed to attain the objectives.
2. Data Understanding: It starts with an initial data collection, to get familiar with the data, to spot data quality issues, to discover 1st insights into the data} or to discover interesting subsets to make hypotheses for hidden information.
3. Knowledge Preparation: It covers all activities to construct the ultimate dataset from the initial information.
4. Modeling: in this phase, numerous modeling techniques are chosen and applied and their parameters are mark to optimal values.
5. Evaluation: in this stage the model is completely evaluated and reviewed. The steps executed to construct the model to be sure it properly achieves the business objectives. At the end of this phase, a choice on the employment of the data mining results should be reached.
6. Deployment: the aim of the model is to increase knowledge of the data, the knowledge gained ought to be organized and presented in a manner that the customer can use it. The deployment phase will be as easy as generating a report or as complicated as implementing a repeatable data mining process across the enterprise.

III. LITERATURE IN DATA MINING

Han et al. provided a comprehensive survey, in database perspective, on the data mining techniques developed recently [7]. Several major kinds of data mining methods, including generalization, characterization, classification, clustering, association, evolution, pattern matching, data visualization, and meta-rule guided mining, was reviewed by them. Techniques for mining knowledge in different kinds of databases, included relational, transaction, object-oriented, spatial, and active databases, as well as global information systems, was examined by them [7]. Clustering is the most commonly used technique of data mining under which patterns are discovered in the underlying data.[8] Sidhu et al. presented that how clustering was carried out and the applications of clustering. They also provided us with a framework for the mixed attributes clustering problem and also showed us that how the customer data can be clustered identifying the high-profit, high-value and low-risk customer [8]. Huang presented an algorithm, called k-modes, to extend the k-means paradigm to categorical domains. He introduced new dissimilarity measures to deal with categorical objects, replace means of clusters with modes, and use a frequency based method to update modes in the clustering process to minimize the clustering cost function.. Experimented on a very large health insurance data set consisted of half a million records and 34 categorical at-tributes showed that the algorithm was scalable in terms of both the number of clusters and the number of records[9] Berkhin surveyed that concentrated on clustering algorithms from a data mining perspective. [10] In [11] k-prototypes algorithm was proposed, which was based on the k-means paradigm but removed the numeric data limitation whilst preserved its efficiency. In that algorithm, objects were clustered against k prototypes. A method was developed to dynamically update the k prototypes in order to maximize the intra cluster similarity of objects. In [12] the efficiency and scalability issues were addressed by proposing a data classification method which integrated attribute oriented induction, relevance analysis, and the induction of decision trees. Such an integration lead to efficient, high quality, multiple level classification of large amounts of data, the relaxation of the requirement of perfect training sets, and the elegant handling of continuous and noisy data.

IV. DATA MINING TECHNIQUES

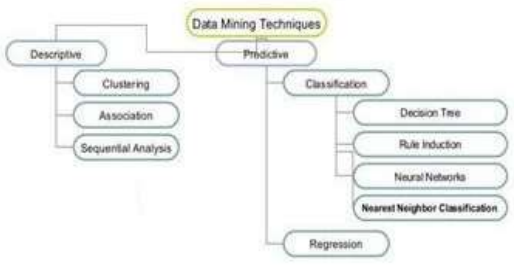

Fig 2. Data Mining Techniques

Descriptive approach includes models for overall chance distribution of the data, partitioning of whole data into sections and models describing the relationships between the variables. prophetic approach permits the value of 1 attribute/variable is simply too predicted from the best- known values of different attribute/variable. This paper studies the one descriptive technique i.e. clustering and one predictive technique i.e. classification.

A) Classification Approach

Classification could be a supervised learning technique [3]. Data classification is two-step method. In the beginning, a model is made by analyzing the data tuples from training data having a group of attributes. For every tuple in the training data, the value of class label attribute is understood. Classification algorithmic rule is applied on data training data to make the model. In the second step of classification, test data is employed to examine the accuracy of the model. If the accuracy of the model is appropriate then the model are often used to classify the unknown tuples [4].Classification techniques were developed as a crucial part of machine learning algorithms so as to extract rules and patterns from data that would be used for prediction. Classification techniques are accustomed classify data records into one among a set of predefined categories. They work by constructing a model of training dataset consisting of example records with best-known category labels [5].

B) Clustering Approach

Clustering is finding groups of objects such that the objects in one group are just like one another and totally different from the objects in another group. Clustering are often thought of the most necessary unsupervised learning technique. Clustering are often considered the most important unsupervised learning technique thus as each other problem of this type. It deals with finding a structure in a collection of unlabeled data. Clustering is the method of organizing objects into groups whose members are similar in some way [9]. Cluster analysis has been widely employed in several applications like business intelligence image pattern recognition web search biology and security. In business intelligence clustering are often used to organize an outsized number of customers into groups wherever customers among a group share similar characteristics. [5].

V. DATA MINING APPLICATIONS

Data Mining in e-Commerce. Data mining enables the businesses to understand the patterns hidden inside past purchase transactions, thus helping in planning and launching new marketing campaigns in prompt and cost- effective way. e-commerce is one of the most prospective domains for data mining because data records, including customer data, product data, users" action log data, are plentiful; IT team has enriched data mining skill and return on investment can be measured.

Data Mining in Industry. Data mining can highly benefit industries such as retail, banking, and telecommunications; classification and clustering can be applied to this area.

One of the key success factors of insurance organizations and banks is the assessment of borrowers" credit worthiness in advance during the credit evaluation process. Credit scoring becomes more and more important and several data mining methods are applied for credit scoring problem.

Data Mining in Health Care. In health care, data mining is becoming increasingly popular, if not increasingly essential. Heterogeneous medical data have been generated in various health care organizations, including payers, medicine providers, pharmaceuticals information, prescription information, doctor"s notes, or clinical records produced day by day. These quantitative data can be used to do clinical text mining, predictive modeling, survival analysis, patient similarity analysis, and clustering, to improve care treatment and reduce waste. In health care area, association analysis, clustering, and outlier analysis can be applied.

Data Mining in City Governance. In public service area, data mining can be used to discover public needs and improve service performance, decision making with automated systems to decrease risks, classification, clustering, and time series analysis which can be developed to solve this area problem.

E-government improves quality of government service, cost savings, wider political participation, and more effective policies and programs , and it has also been proposed as a solution for increasing citizen communication with government agencies and, ultimately, political trust. City incident information management system can integrate data mining methods to provide a comprehensive assessment of the impact of natural disasters on the agricultural production and rank disaster affected areas objectively and assist governments in disaster preparation and resource allocation.

VI. CONCLUSION

Most of the previous studies on data processing applications in numerous fields use the range of data varieties range from text to pictures and stores in kind of databases and data structures. The various ways data mining are used to extract the patterns and therefore the knowledge from this selection databases. Selection of data and strategies for data mining is a vital task in this process and desires the knowledge of the domain. Several attempts are created to design and develop the generic data mining system however no system found completely generic. Thus, for each domain the domain expert"s assistant is necessary. The domain specialists shall be guided by the system to effectively apply their knowledge for the use of data mining systems to generate needed knowledge. The domain specialists are needed to work out the variety of data that should be collected in the specific problem domain, selection of specific data for data mining, cleaning and transformation of data, extracting patterns for knowledge generation and finally interpretation of the patterns and knowledge generation.

REFERENCES

[1]. Bhise, R. B., S. S. Thorat, and A. K. Supekar. "Importance of data mining in higher education system." IOSR Journal Of Humanities And Social Science (IOSR-JHSS) ISSN (2013): 2279-0837.
[2]. Padhy, Neelamadhab, Dr Mishra, and Rasmita Panigrahi. "The survey of data mining applications and feature scope." arXiv preprint arXiv:1211.5723 (2012).
[3]. Sumathi, N., R. Geetha, and S. Sathiya Bama. "Spatial data mining—techniques trends and its applications." Journal of Computer Applications 1, no. 4 (2008): 28-30
[4]. Shaikh,Yasmin ,and Sanjay Tanwani. "INTERACTIVE TEMPORAL MINING OF WORKFLOW LOGS." (2013).
[5]. http://en.wikipedia.org/wiki/Sequential_Pattern_Mi ning
[6]. http://en.wikipedia.org/wiki/Intention_mining
[7]. Han, Jiawei. "Data mining techniques." In ACM SIGMOD Record, vol. 25, no. 2, p. 545. ACM, 1996
[8]. Sidhu, Nimrat Kaur, and Rajneet Kaur. "Clustering In Data Mining."
[9]. Huang, Zhexue. "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining." In DMKD, p. 0. 1997.
[10]. Berkhin, Pavel. "A survey of clustering data mining techniques." In Grouping multidimensional data, pp. 25-71. Springer Berlin Heidelberg, 2006.
[11]. Huang, Zhexue. "Clustering large data sets with mixed numeric and categorical values." In Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining,(PAKDD), pp. 21-34. 1997.
[12]. Kamber, Micheline, Lara Winstone, Wan Gong, Shan Cheng, and Jiawei Han. "Generalization and decision tree induction: efficient classification in data mining." In Research Issues in Data Engineering, 1997. Proceedings. Seventh International Workshop on, pp. 111-120. IEEE, 1997.
[13]. Antonie, Maria-Luiza, Osmar R. Zaiane, and Alexandru Coman. "Application of Data Mining Techniques for Medical Image Classification." In Proceedings of the Second International Workshop on Multimedia Data Mining, MDM/KDD'2001, August 26th, 2001, San Francisco, CA, USA, pp. 94-101. 2001.
[14]. Kohavi, Ronny, and J. Ross Quinlan. "Data mining tasks and methods: Classification: decision-tree discovery." In Handbook of data mining and knowledge discovery, pp. 267-276. Oxford University Press, Inc., 2002.

