ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

Review Article

Available Online at www.ijarcs.info

Role of Data Mining: A Survey and its Implications

Amit Singh
Assistant Professor, Computer Application
Government Degree College
Ramnagar, India

Santosh Kumar Agray
TGT Computer Science
Army Public School Ratnuchak
Jammu, India

Abstract: Data mining is a procedure of mining information from huge set of heterogeneous data which is scattered over different storage repositories, smart devices, internet of things etc. In this paper, we have discussed about various Data Mining techniques viz classification, clustering, regression and decision tree. We have also discussed use of various data mining applications in financial analysis, health care, fraud detection and prevention, telecommunication industries, science and engineering.

Keywords: Data, Data Mining Techniques, application, decision trees, classification, clustering, pre-processing

I. INTRODUCTION

The explosive growth of information technology has generated tremendous amount of data from smart devices, search engines, medical data, etc., which accumulate data in different formats like text, hypertext, imaging data, geo spatial formats and to analysis this amount and type of data a new field has been emerged called data mining which enable us to fetch knowledge and discovers hidden patterns from those data repositories.

Data Mining is the process of extraction of hidden patterns from the huge data. Data mining is a powerful tool for future trends based on the current and historical data which helps organisation for better decision making.

Data Mining or knowledge discovery in databases, as it is known, is the non-trivial extraction of implicit, previously unknown and potentially useful information from the data. This encompasses a number of technical approaches, such as clustering, data summarization, classification, finding dependency networks, analysing changes, and detecting anomalies [1]

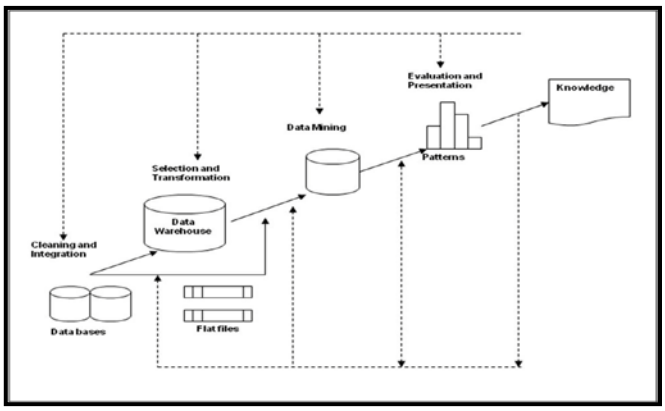

Figure 1. Knowledge Discovery Process

II. PHASES OF KNOWLEDGE DISCOVERY PROCESS(KDD)

The phases of KDD, starting with the raw data and finishing with the extracted knowledge are given below:

Selection

In this stage, only significant data gets selected which are relevant to some criteria.

A. *Pre-processing*

Pre-processing is the data cleaning stage where noise and inconsistent information is removed.

B. *Transformation*

In this step aggregation or summary operations are performed to transform into consolidated forms appropriate for data mining tasks.

C. *Data Mining*

This stage is concerned with the searching for interested patterns of data in a particular form.

D. *Interpretation and Evaluation*

After data mining stages patterns are converted into knowledge, which helps decision-making.

E. *Data Visualization*

Data visualization helps users to interpret large sets of data and detect the interested patterns visually [1].

III. DATA MINING TECHNIQUES

A. *Classification*

In Classification data samples are divided into target classes. For each data points the classification technique predicts the target class. For example, patient class can be classified into “high risk” or “low risk” patient which is based on their disease pattern using classification approach. This is called as supervised learning approach having previously known class categories. Two methods of classification are Binary and multilevel. In multiclass classification approach, has more than two target classes while binary classification has only two possible classes. Data set is divides into two sets called training and testing dataset. We trained the classifier by

using training dataset. The classifier correctness could be tested by using testing dataset [9].

B. Clustering

Cluster analysis is the collection of patterns (which is represented as a vector of measurements in a multidimensional space) into clusters based on similarity [10]. The difference between clustering (unsupervised classification) and discriminant analysis (supervised classification) is very important to understand. In supervised classification, we are provided with a collection of labeled (preclassified) patterns [5]; the problem is to label a newly encountered, yet unlabelled, pattern. Typically, the given labeled (training) patterns are used to learn the accounts of classes which are used to label a new pattern. In the case of clustering, the job is to group a given collection of unlabeled patterns into meaningful clusters. In a sense, labels are connected with clusters also, but these category labels are data driven; that is, they are obtained exclusively from the data. Clustering is useful in several exploratory pattern-analysis, grouping, decision-making, and machine-learning conditions, including data mining, document retrieval, image segmentation, and pattern classification. However, in many such problems, there is little prior information (e.g., statistical models) available about the data, and the decision-maker must make as few assumptions about the data as possible. It is under these restrictions that clustering methodology is particularly appropriate for the exploration of interrelationships among the data points to make an assessment (perhaps preliminary) of their structure. The term "clustering" is used in several research communities to describe methods for grouping of unlabeled data [6]. These communities have different terminologies and assumptions for the components of the clustering process and the context in which clustering is used. Thus, we face a dilemma regarding the scope of this survey. The production of a truly comprehensive survey would be a monumental task given the sheer mass of literature in this area. The accessibility of the survey might also be questionable given the need to reconcile very different vocabularies and assumptions regarding clustering in the various communities [7].

Typical pattern clustering activity involves the following steps [Jain and Dubes 1988]:

- pattern representation (optionally including feature extraction and/or selection),
- definition of a pattern proximity measures appropriates to the data domain,
- clustering or grouping,
- data abstraction (if needed), and
- assessment of output (if needed).

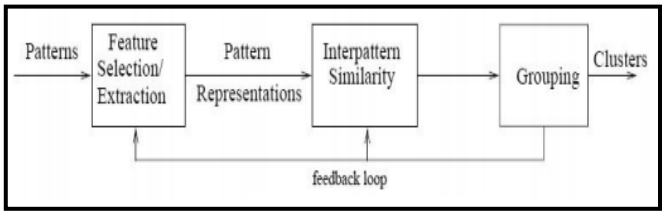

Figure 2. Stages in Clustering

C. Regression

Regression is a technique which is used to find out functions that explain the association among different variables. In statistical modeling two types of variables are used where one is called independent variable and another one is called dependent variable and which is usually represented by using 'Y' and 'X' [11]. There is only one dependent variable while independent variable may be one or more than one. By using Regression dependencies of one variable upon others may be established. There are two type of regression which is based on number of independent variables, first is linear and second one is Non-Linear. Linear regression classifies relation of a dependent variable and one or more independent variables. It utilizes linear function for its construction. It discovers a line and computes vertical distances of points from the line and minimize sum of square of vertical distance. In this method, dependent and independent variables are previously known and purpose is to spot a line that relates between these variables. But, linear regression is restricted to numerical data only and cannot be used for categorical data. Logistic regression is a type of non-linear regression can take categorical data and predicts the probability of occurrence using logic function. Logistic regression is of two types Binomial and multinomial. Binomial regression forecasts the result for a dependent variable when there occur only two likely outcomes such as either a person is dead or alive while the multinomial handles the situation when dependent variable has three or more results. For example, either a patient is at 'low risk', 'medium risk' and 'high risk' [3].

D. Decision Tree (DT)

DT is like flowchart in which each non-leaf nodes denotes a test on a specific attribute and every branch means an outcome of that test and every leaf node have a class label. The node at the highest labels in the tree is called root node [12]. For example, we have a financial institution decision tree which is used to decide that a person must grant the loan or not. Building a decision for any problem doesn't want any type of field knowledge. Decision Trees is a classifier that use tree-like graph. In operations research analysis Decision tree is most widely used for calculating conditional probabilities. Using Decision Tree, decision makers can select best alternative and traversal from root to leaf indicates unique class separation based on maximum information gain. Decision Tree is widely used in healthcare field [3].

IV. DATA MINING APPLICATIONS

A. Data mining for Financial Data Analysis

Data collected by the financial and banking industry are relatively reliable, complete and of high quality, which enables organized data analysis and Data Mining.

Data mining has been used widely in the banking and financial markets. Data mining is used to predict credit fraud, to estimate risk, to perform trend analysis, and to analyse profitability, stock-price forecasting, direct marketing campaigns, in portfolio management, in commodity price prediction, in mergers and acquisitions, as well as in forecasting financial disasters.

Data warehousing and data mining are being used at American Express to cut spending. American Express has merged its worldwide purchasing system, corporate purchasing cards and corporate-cards and created a single Microsoft SQL Server database. This allows American Express to find exceptions and patterns to target for cost cutting. One of the main applications is loan application screening. Statistical methods were used by American Express to categorized loan applications into three classes:

- Those that should be accepted,
- Those that should definitely be denied.
- Those which required a human expert to judge.

In about 50% of the cases the human experts could correctly forecast if an applicant would default on the loan or not whereas rules produced by Machine learning were much more precise—correctly forecasting default in 70% of the cases—and that were instantly put into use [2].

B. Data Mining for Telecommunication Industries

The telecommunications industry generates and stores a tremendous amount of data. These data include call detail data, which describes the calls that traverse the telecommunication networks, network data, which describes the state of the hardware and software components in the network, and customer data, which describes the telecommunication customers. The amount of data is so great that manual analysis of the data is difficult, if not impossible. The need to handle such large volumes of data led to use of data mining.

Data mining is being used by companies like Air Touch Communications, AT&T, and AMS Mobile Communication Industry Group to improve their marketing activities. There are many companies including Lightbridge and Verizon that use data-mining to detect cellular fraud for the telecommunications industry. Advanced visualization techniques are another trend that are used to model and analyze wireless-telecommunication networks [2].

C. Data Mining in Healthcare

Today is the age of data mining where prediction of variety of disease is enduring into procedure. Data mining has proved with flourished results in medical. Data mining has plenty of techniques and tools available and has been implemented in the industries. Data mining can be used to make successful decisions that will improve success of healthcare organization and health of the patients. Health care organizations and agencies could come across into these applications to find ideas on how to extract interested patterns from their own database systems.

Example: To enhance the company's capability to recognize high-risk patients, American Health ways uses predictive modeling technology from SAS. Basically, predictive modeling helps us identify patients who are moving toward a high-risk condition, which gives our nurse care coordinators a head start in identifying high-risk individuals, so that they can take steps to improve their quality of healthcare and prevent health problems in the future. By merging high tech solutions like predictive modeling with our clinical information system in support of our nurses, risk scores for each patient is generated. This results in better knowledge about the likely course of the patient's disease.

D. Data Mining in Science and Engineering

Huge amounts of data have been generated in science and engineering, for example, in astronomy, molecular biology, and chemical engineering. There are a lot of massive natural, technical, social information in science and engineering applications. This information includes gene, protein, and microarray information in biology; highway transportation information in civil engineering; topic- or theme-author-publication-citation data in library science; and wireless telecommunication data that is being shared among commanders, soldiers, and supply lines in a battle field. In such information networks, each and every link or node in a network has valuable, multidimensional information, such as textual contents, geographic information, traffic flow, and other properties. To process this data is very complex and difficult task [4].

Data-mining algorithms such as classification, cluster analysis and market basket analysis usually attempt to discover patterns in a data set containing identically distributed and independent samples. It can also be used in prediction which plays an important role in decision making.

Boeing has effectively applied machine - learning algorithms to the discovery of informative and useful rules from its plant data to improve its manufacturing process. In particular, it has been found that it is more beneficial to seek concise predictive rules that cover small subsets of the data, rather than generate general decision trees. A number of rules were discovered to predict events such as when a manufactured part is likely to fail or when a fault will occur at a particular machine. Due to these rules, potentially important anomalies were identified before failure.

E. Data Mining for Fraud Detection and Prevention

Another popular area where data mining can be implied in the banking industry is in fraud detection. To detect fraudulent actions is an alarming concern for many businesses organizations; and with the help of data mining more fraudulent actions can be detected and reported. Two different actions have been developed by financial institutions to detect fraud patterns. In the first approach, a bank bangs the data warehouse of a third party (potentially containing transaction information from many companies) and uses data mining programs to identify fraud patterns [8]. The bank can then cross-check those patterns with its own database for signs of internal trouble. In the second approach, fraud pattern identification is based strictly on the bank's own internal information. Most of the banks use a 'hybrid' approach. One system which is successful in detecting fraud is Falcon's 'fraud assessment system'. It is exploited by nine of the top ten credit card issuing banks, where it examines the transactions of 80 per cent of cards held within the US. Mellon Bank also uses data mining for fraud detection and is able to create an improved image for itself and its customers' funds remain protected from potential credit card fraud.

V. REFERENCES

[1] Arun K Pujari, "Data Mining Techniques", 2001.

[2] Jiawei Han, Micheline Kamber, and Jian Pei, "Data Mining Concepts and Techniques", 2011.

[3] Hillol Kargupta, Jiawei Han, Philip S. Yu, Rajeev Motwani, and Vipin Kumar, "Next Generation of Data Mining", 2008.

[4] Mehmed Kantardzic, "Data Mining: Concepts, Models, Methods, and Algorithms", 2011.

[5] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: A review", ACM SIGKDDExplorations Newsletter, vol. 6, pp. 90-105, 2004.

[6] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for projected clustering of high dimensional datastreams," in Proceedings of the Thirtieth international conference on Very large databases-Volume 30, p.863, 2004.

[7] R. Agrawal, J. E. Gehrke, D. Gunopulos, and P. Raghavan,"Automatic subspace clustering of high dimensional data fordata mining applications," Google Patents 1999.

[8] M. P. Thapliyal, "Data Mining: A Tool for Banking Industry", Vol-4, Issue-4, 2015.

[9] S.Yamini, Dr.V.Khanaa, Dr.Krishna Mohantha , A State of the Art Review on Various Data Mining Techniques 2016.

[10] Rakesh teki, Hari narayana P, Improving The Predictive Performance by Integrating Decision tree based Attribute Selection in Clustering, Vol-4 Issue 7, July 2013.

[11] Dan Campbell, Sherlock Campbell, Introduction to Regression and Data Analysis, Statlab Workshop, 2008.

[12] Himani Sharma, Sunil Kumar, A Survey on Decision Tree Algorithms of Classification in Data Mining, Vol-5 Issue 4, 2016.