Data Mining Techniques and Applications - A study

Koyel Datta Gupta
Department of Computer Science & Engineering
Maharaja Surajmal Institute of Technology

Abstract- With the advent of new technologies, storage of large amount of data has become simple. To handle this enormous data efficiently for analysis is a however a big task. For years researchers around the world, are devoted to improve data mining techniques for better analysis and management of data. This involves evolution of different mining techniques and deploys them for varied application. This paper presents a study on the diverse techniques of data mining and its applications.

Keywords – ***Data mining, data mining applications, knowledge discovery database***

I. INTRODUCTION

Exponential growth in the amount of data storage calls for efficient techniques to retrieve them from the database for improved decision making .This procedure is referred as Knowledge Discovery Process. Knowledge Discovery in Databases (KDD) otherwise known as Data mining is non-trivial retrieval of inherent, unknown and potentially valuable information from raw data stored in data repositories. The primary objective of data mining is to retrieve appropriate information and ascertain the underlying patterns in the raw data for better analysis. Data mining can be defined as a technique to retrieve unknown analytical information from enormous data repositories [1-2].

The KDD process consists of a few steps starting with raw data collection leading to some form of pattern discovery.

- **Data cleaning** phase removes irrelevant data from the stored data.
- **Data integration** phase combines multiple data sources into one unified source.
- **Data selection** involves collection of relevant data from the storage.
- **Data transformation** converts collected data into appropriate formats for the mining procedure.
- **Data mining** is applied to retrieve useful information for data analysis.
- **Pattern evaluation** unearths underlying patterns from the retrieved information according to the specified measures.
- **Knowledge representation** involves visual representation of discovered pattern.

Data mining tools are essential to envisage upcoming development and perform business analysis in order to assist organizations to formulate knowledge-driven decisions for future. The data mining techniques can be deployed to analyze the current scenario of business and predict the future course of action for better performance in a stipulated time.

Data mining can be supervised or unsupervised in nature. When supervised learning is adapted, the input patterns and the desired output patterns are specified. The parameters of the data mining tool adapt in a way such that the application of an input pattern results in the desired pattern at the output. In case of unsupervised learning, the data mining technique extracts the features from the input data based on a specified performance measure.

The rest of the paper is organized as follows. Section 2 presents the data mining task. Study of recent data mining techniques and their respective applications is presented in section 3. Section 4 concludes this paper.

II. DATA MINING TASKS

The data mining task can be divided as one of the following two types [3].

i. Generation of descriptive models
   The descriptive data mining tasks identifies hidden patterns describing the data that can be analysed by subjects.
ii. Generation of predictive models

Predictive data mining uses data set to envisage prospective values of items of interest by prediction, estimation or classification.

The tasks can be supervised or unsupervised in nature. Table - 1 mention some of the data mining tasks and give a brief description about them

Table - 1 Data Mining Tasks

Nature of Learning	Data Mining Task	Brief Description	Existing Algorithms
Supervised	Classification/Prediction	Classification is the unearthing of a predictive learning function that categorizes a data item into one of several predefined classes[4] Prediction is used to evaluate the probable value of an item in a class.	Decision Trees Naïve Bayes Support Vector Machine
Supervised	Regression	Regression is discovering function with minimal error to model data.[4]	Multiple Regression
Supervised	Estimation	Estimation involves discrete outcomes such as yes or no with continuously valued outcomes.	Auto Regression
Unsupervised	Clustering	Clustering is placing a set of similar items together and placing dissimilar items in separate groups.	K-Means
Unsupervised	Association Mining	Association mining is finding relationship between multiple items in a set of data.	Apriori
Unsupervised	Feature Extraction	Feature Extraction uses statistical information about attributes in a data set and categorizes attributes into general characteristics and reduces the number of features.	Non-Negative Matrix Factorization

II. Study of Recent Data Mining Techniques and Applications

In [5], the authors propose a technique to incorporate diverse data mining models into multidimensional models for achieving the design of data warehouses with association rules. For conceptual design of data warehouse, the paper provides an UML profile to indicate Association Rules mining. The association rules are based on the request pattern of user.

The work [6] uses a technique for maintaining the confidentiality during sequential pattern mining on network traffic data. Another technique referred as retention replacement is also applied for probabilistically changing query system. Meta tables are used to identify the occurrence of candidate patterns in a site.

The authors in [7] presented a concept of multicarrier communication for achieving high data rate and reduce delay stretched over fading channel. By deploying a distributed proportional integrative plus derivative (PID) controller along with data mining, the proposed ODMCA (online data mining control algorithm) controls the transfer rate of mined data, which causes high storage in main memory of end nodes.

A multidimensional association mining model is presented in [8] where users can construct efficient data mining models through acumen support of the ontologies that can determine concept extended rules, avert useless pattern generation and offer a dynamic knowledge unearthing mechanism.

In the supervised machine learning domain, a collection of techniques have been applied in [9] to increase the correctness and solidity of base learner. In this paper, the cluster combination problem is solved by formulating it as an optimization problem and a linear algebra based cluster combination algorithm is proposed.

In [10], an improved K-means algorithm is presented for clustering to determine unseen patterns that will offer supplementary resource of knowledge to healthcare professionals for better decision making.

IV.CONCLUSION

The paper made a brief review of some of the recent data mining techniques and their applications. The review is intended for researchers to determine the existing techniques and their challenges. Data mining has a deep impact on the society and has wide range applications. The diverse data mining techniques are applied to extract the unseen patterns and deduce knowledge from different databases. Feature selection and techniques for data mining is a crucial task and requires an insight understanding of the domain. Most of the domain specific data mining applications show accuracy above 90%. Both supervised and unsupervised techniques have deployed to retrieve knowledge from a set of data. However, the results obtained from supervised technique have been found to be more precise.

REFERENCES

[1] Han, J. and Kamber, M., Data mining: concepts and techniques, Morgan Kaufmann, 2000.
[2] Larose, D.T., Discovering knowledge in data: an introduction to data mining, JohnWiley and Sons, 2005.
[3] Chan, C. and Lewis, B. "A basic primer on data mining," *Information Systems Management*, Vol. 19, No. 4,2002, pp.56-60.
[4] Pang-Ning,T., Steinbach, M. and Kumar, V. "Introduction to Data Mining", (First Edition), Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 2005
[5] Zubcoff,J., Trujillo,J. " A UML 2.0 profile to design Association Rule mining models in the multidimensional conceptual modeling of data warehouses, " *Data & Knowledge Engineering*, Vol. 63,No. 1, 2007,pp. 44-62.
[6] Kim, S., Park, S., Won, J. and Kim, S. "Privacy preserving data mining of sequential patterns for network traffic data**,"** *Information Sciences*, Vol. 178, No. 3, 2008, pp. 694-713.
[7] Xiong,N., Yang,L.T. and Li,Y. " ODMCA: An adaptive data mining control algorithm in multicarrier networks," *Computer Communications,* Vol. 32, No. 3, 2009, pp. 560-567.
[8] Wu, C.-A., Lin, W.-Y., Jiang,C.-L. and Wu,C. "Toward intelligent data warehouse mining: An ontology-integrated approach for multi-dimensional association mining," *Expert Systems with Applications,* Vol. 38, No. 9, 2011, pp. 11011-11023.
[9] Xu ,S., Wang,Z. , Li,X. and Cao, R. "A novel cluster combination algorithm for document clustering," *in Proc. Sino-foreign-interchange conference on Intelligent Science and Intelligent Data Engineering,* 2012, p.189-195.
[10] Nguyen,D.T., Nguyen,G.T. Nguyen Lam V.T., "An Approach to Data Mining in Healthcare: Improved K-means Algorithm," *Journal of Industrial and Intelligent Information* Vol. 1, No. 1, 2013, pp.14-18.