International Journal of Research
Available at https://journals.pen2print.org/index.php/ijr/
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 09
August 2019

The Study Of Four Data Minng Techniques For Business Intelligence Concept

Ya Min[1], Khin Myat Nwe Win[2], Aye Mya Sandar[3]
[1]Faculty of Computer Science Department, University of Computer Studies (Lashio), Shan State, Myanmar
[2]Faculty of Computer Science Department, University of Computer Studies (Mandalay), Mandalay, Myanmar
Information Technology Supporting and Maintenance Department, University of Computer Studies (Mandalay)
Mandalay, Myanmar

yamin1977lso@gmail.com,khinmyatnwewin@googlemail.com, ayemyasandarpaing@gmail.com

Abstract

Data mining is the process of discovering potentially useful, interesting, and previously unknown patterns from a large collection of data. The objective of this paper is to present the study of four techniques on what are concept of Data Mining (DM) in Business Intelligence (BI). The paper highlights detain about the basic concept of data mining in introduction session. It also involves about the Knowledge Discovery in Databases and Five Major Elements in Data Mining. BI is the hot topic among all industries aiming for relevance. DM and BI work together to process and analyses data to lighten workload for the user and organization and hence in understanding disco.

Keywords: Data mining; Data Mining Techniques; Business Intelligence.

I. Introduction

Today, we have far more information than we can handle: from business transactions and scientific data, to satellite pictures, text reports and military intelligence. Information retrieval is simply not enough anymore for decision-making. Confronted with huge collections of data, we have now created new needs to help us make better managerial choices. These needs are automatic summarization of data, extraction of the essence of information stored, and the discovery of patterns in raw data. [8] Data mining is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers to develop more effective marketing strategies, increase sales and decrease costs. Data mining depends on effective data collection, warehousing, and computer processing.[10]

According to paper [9], Data mining is the extraction of useful patterns and relationships from data sources, such as databases, texts, the web. It has nothing to do however with SQL, OLAP, data warehousing or any of that kind of thing. It uses statistical and pattern matching techniques. The concern in data mining are noisy data, missing values, static data, sparse data, dynamic data, relevance, interestingness, heterogeneity, algorithm efficiency, size and complexity of data. Data mining has become a popular tool for analyzing large datasets. The

efficient database management systems have been very important assets for management of a large corpus of data and especially for effective and efficient retrieval of particular information from a large collection whenever needed. The increase of database management systems has also contributed to recent massive gathering of all sorts of information. Information retrieval is simply not enough anymore for decision-making. [9]

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making [11]. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. The term KDD (Knowledge Discovery in Databases) refers to the overall process of discovering useful knowledge from data, where data mining is a particular step in process. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. [9].

Five Major Elements in Data Mining 1) Extract, transform, and load transaction data onto the data warehouse system. 2) Store and manage the data in a multidimensional database system. 3) Provide data access to business analysts and information technology professionals. 4) Analyze the data by application software. 5) Present the data in a useful format, such as a graph or table. [9]

II. Four Data Mining Techniques

Galvanize (2016) proposed in the paper [1], four data mining techniques with their related examples,

1). Regression (predictive)

2). Association Rule Discovery (descriptive)

3). Classification (predictive)

4). Clustering (descriptive)

1). Regression (predictive)

Regression is the most straightforward, simple, version of what we call "predictive power." When we use a regression analysis we want to predict the value of a given (continuous) feature based on the values of other features in the data, assuming a linear or nonlinear model of dependency.

For examples:

- Predicting revenue of a new product based on complementary products.
- Predicting cancer based on the number of cigarettes consumed, food consumed, age, etc.
- Time series prediction of stock market and indexes.

Regression techniques are very useful in data science, and the term "logistic regression" will appear almost in every aspect of the field. This is especially the case due to the usefulness and strength of neural networks that use a regression-based technique to create complex functions that imitate the functionality of our brain. [1]

2). Association Rule Discovery (descriptive)

Association rule discovery is an important descriptive method in data mining. It's a very simple method, but you'd be surprised how much intelligence and insight it can provide the

kind of information many businesses use on a daily basis to improve efficiency and generate revenue.

Our goal is to find all rules (X —> Y) that satisfy user-specified minimum support and confidence constraints, given a set of transactions, each of which is a set of items. Given a set of record search of which contain some number of items from a given collection we want to find dependency rules which will discover occurrence of an item based on occurrences of other items.

For example: Assume you have a dataset of all your past purchases from your favorite grocery store, and I found a dependency rule (minimizing with respect to the constraints) between these items: {Diapers} —> {Beer}.

This creates dependencies, based on the specified minimum support and confidence, which are defined as such.

$$\boldsymbol{support} = \frac{\boldsymbol{number\ of\ transactions\ containing\ X\ and\ Y}}{\boldsymbol{toal\ number\ of\ transactions}}$$

$$\boldsymbol{confidence} = \frac{\boldsymbol{number\ of\ transactions\ containing\ X\ and\ Y}}{\boldsymbol{number\ of\ transactions\ containing\ X}}$$

The applications for associate roles are vast and can add lots of value to different industries and verticals within a business. Here are some examples: Cross-selling and up-selling of products, network analysis, physical organization of items, management, and marketing. This was an industry staple for decades in market basket analysis, but in recent years, recommendation engines have largely come to dominate these traditional methods. [1]

3). Classification (predictive)

Classification is another important task you should handle before digging into the hardcore modeling phase of your analysis. Assume you have a set of records, each record contains a set of attributes, where one of the attributes is our *class* (think about letter grades). Our goal is to find a model for the *class* that will be able to *predict* **unseen** or **unknown** records (from external similar data sources) *accurately* as if the label of the class was **seen** or **known**, given all values of other attributes. In order to train such a model, we usually divide the data set into two subsets: *training set* and *test set*. The training set will be used to build the model, while the test set used to validate it. The accuracy and performance of the model is determined on the test set. Classification has many applications in the industry, such as direct marketing campaigns and churn analysis:

Direct marketing campaigns are intended to reduce the cost of spreading marketing content (advertising, news, etc.) by *targeting* a set of consumers that are likely to be interested in the specific content (product, discount, etc.) based on their revealed past data and behavior. The method is simply to collect data for a similar product (for simplicity) introduced in the recent past and to *classify* the profiles of customers based upon whether they **did buy** or **didn't buy**. This target feature will become the *class attribute*. Now we need to enhance the data with additional demographic, lifestyle, and other relevant features in order to use this information as input attributes to train a classifier model. [1]

Churn is the measure of individuals losing interest in your offering (service, information, product, etc.). In business it's incredibly important to monitor churn and attempt to identify why subscribers (clients, etc.) decided to stop paying for the subscription. In other words, churn analysis tries to predict whether a customer is likely to be lost to a competitor. [1] To analyze churn, we need to collect a detailed record of transactions with each of the past and current customers, to find attributes that can explain or add value to the question in hand. Some of these attributes can be related to how

engaged the subscriber was with the services and features that the company offers. Then we simply need to *label* the customers as **churn** or **not churn** and find a model that will best fit the data to predict how likely each of our current subscribers is to churn. [1]

For example: A classification model could be used to identify loan applicants as low, medium, or high credit risks. A classification task begins with a data set in which the class assignments are known. For example, a classification model that predicts credit risk could be developed based on observed data for many loan applicants over a period of time. In addition to the historical credit rating, the data might track employment history, home ownership or rental, years of residence, number and type of investments, and so on. Credit rating would be the target, the other attributes would be the predictors, and the data for each customer would constitute a case. The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible values: for example, high credit rating or low credit rating. Multiclass targets have more than two values: for example, low, medium, high, or unknown credit rating. [2]

4). Clustering (descriptive)

Clustering is an important technique that aims to determine object groupings (think about different groups of consumers) such that objects within the same cluster are similar to each other, while objects in different groups are not. The Clustering problem in this sense is reduced to the following:

Given a set of data points, each having a set of attributes, and a similarity measure, find clusters such that:

- Data points in one cluster are more similar to one another.
- Data points in separate clusters are less similar to one another.

In order to find how close or far each cluster is from one another, you can use the Euclidean distance (if attributes are continuous) or any other similarity measure that is relevant to the specific problem.

A useful application of clustering is marketing segmentation, which aims to subdivide a market into distinct subsets of customers where each subset can be targeted with a distinct marketing strategy. This is done by collecting different attributes of customers based on their geographical- and lifestyle-related information in order to find clusters of similar customers. Then we can measure the clustering quality by observing the buying patterns of customers in the same cluster vs. those from different clusters. [1]

For example: Data mining K means algorithm is the best example that falls under this category. In this model the number of clusters required at the end is known in prior. Therefore, it is important to have knowledge of the data set. These are iterative data mining algorithms in which the data points closer to the centroid in the data space will be aggregated to the single cluster. Number of centroid is always equal to the number of clusters. [3]

III. Business Intelligence Using Data Mining

As described by the author [4] Business Intelligence (BI) is a concept of applying a set of technologies to convert data into meaningful information. Basically, the term business intelligence has two different meanings when related to intelligence. The first is the human intelligence or the capacity of a common brain applied to business affairs. Business Intelligence has become a novelty, the applications of human intellect and new technologies like artificial intelligence is used for management and decision making in different business related problems. The second is the information which helps raise currency in business. The intelligent knowledge gained by experts and efficient

technology in managing organizational and individual business. [5]

Emergence of business intelligence has thrown a light upon the new dimensions of the data collected over a business. In this paper [6] the author said that risk management and enterprise decision-making are inseparable from mining tools. Business Intelligence (BI) can only be acquired by using mining of data in different ways. Use of data warehousing and Information Systems (IS) have made it possible for enterprise datasets to grow rapidly.

With the prescient knowledge the author in paper [7] has said that the demand for more sophisticated and intelligent BI solutions is constantly growing due to the fact that storage capacity grows with twice the speed of processor power. This unbalanced growth relationship will over time make data processing tasks more time consuming when using traditional BI solutions.

There are a variety of advanced data processing techniques that can help BI processes to run efficiently which are offered by DM. The comprehensive process of applying BI for a business problem is referred to as the Knowledge Discovery in Databases (KDD) process and is vital for successful DM implementations with BI in mind. [5]

IV.CONCLUSION

Nowadays, Data Mining has great importance in competitive business environment. Data mining has importance concerning result the patterns, forecasting, discovery of knowledge etc., in different business domains. Data mining has wide-ranging application domain almost in every industry where the data is generated that's why data mining is considered one of the most important limits in database and information systems and one of the most promising interdisciplinary developments in Information Technology. A new concept of Business Intelligence data mining has evolved now, which is widely used by leading corporate houses to stay ahead of their competitors. Business Intelligence (BI) can help in providing latest information and used for competition analysis, market research, economic trends, consume behavior, industry research, geographical information analysis and so on. Business Intelligence Data Mining helps in decision-making.

References

[1]. Galvanize (2016), "4 Data Mining Techniques for Businesses (That Everyone Should Know)",

https://blog.galvanize.com/four-data-mining-techniques-for-businesses-that-everyone-should-know/

[2].

https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#i1005746

[3]. http://dwgeek.com/various-data-mining-clustering-algorithms-examples.html/

[4]. Arti J. Ugale, P. S. Mohod, "Business Intelligence Using Data Mining Techniques on Very Large Datasets", International Journal of Science and Research (IJSR), Volume 4 Issue 6, June 2015 , pp-2932-2937

[5]. Brojo Kishore Mishra, Deepannita Hazra, Kahkashan Tarannum and Manas Kumar,""Business Intelligence using Data Mining Techniques and Business Analytics" ,5th International Conference on System Modeling & Advancement in Research Trends, 25th_27'h November, 2016 College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad, India

[6]. Prachiagarwal , "Benefits and Issues Surrounding Data Mining and its Application in th e Retail Indu stry", International Journal of Scientific and Research Publications, Volume 4 , Issue 7, July 2014.

[7]. NielsArnth-Jensen, "Applied Data Mining for Business Intelligence

[8]. Osmar R. Zaïane, 1999," CMPUT690 Principles of Knowledge Discovery in Databases".

[9]. Ms. Aruna J. Chamatkar* , Dr. P.K. Butey," Importance of Data Mining with Different Types of Data Applications and Challenging Areas".Int. Journal of Engineering Research and Applications www.ijera.com ISSN : 2248-9622, Vol. 4, Issue 5(Version 3), May 2014, pp.38-4

[10].

https://www.investopedia.com/terms/d/datamining.asp

[11]. M. Shiga, I. Takigawa, and H. Mamitsuka, "A spectral clustering approach to optimally combining numericalvectors with a modular network," in KDD, 2007, pp. 647–656