Data Mining Techniques: Review

Abdullahi Sidow Osman[1]

[1]Faculty of Computer and Information Technology, Al-Madinah International University (MEDIU), Malaysia, sedo1414@gmail.com

Abstract

Data mining defines as the practice of examining a large pre-existing database in order to generate new information. A robust technology has a large potential to help organizations to focus on the most significant information in stored data warehouses. Data mining tools and techniques will predict future trends by making the business more proactive, and better knowledge-driven decisions. Data mining techniques could be answer questions that related to the business which traditionally were too time-consuming to resolve. The goal of this paper is to introduce data mining techniques. The focus will be on techniques that include Clustering, Decision, tree Prediction, and Neural Networks.

Keywords: Data mining, Classification, Decision, Association, Neural Networks

I. INTRODUCTION

Big data Defines as one of the most important technologies in an aspect of the growth of any brand now. In modern times, organizations are using big data analytics techniques to attain major goals in their business, especially in terms of organization growth and customer satisfaction.

At the same time, it is important to understand that comprehending and analyzing big data is important for the successful growth and expansion of an organization.

For this reason, data mining techniques are very useful as they can assist organizations to analyze big data in an efficient manner. Although there are various data mining techniques available which can used to feed into different problems and provide insights into those special subsequent business problems [1].

Every sample of data mining techniques will effect in a different result. This shows that identifying the business problems will go a long way in assisting brands to execute the right data mining techniques and whereby get the best results as well.

Data mining techniques involve collection, extraction, analysis and statistical methods. To describe the entire range of data analytics used buzzword.

At the same moment, it is necessary to keep in mind that data-mining techniques further indicates the discovery of unusual or dependencies records that were undetected, unknown interesting patterns.

For this reason, it is very important to develop a significant strategy in this way, which influence of data mining techniques that understood by the organization simply.

The term of data mining was beginning in the 1990s to define analysis of data without and a pre-concluded hypothesis Data fishing or data dredging which has used before.

The main goals of the data mining tools and techniques are to collect certain information that could be clearly applying to large datasets [2].

Here are some significant data mining techniques, which could take the business forward in a comprehensive and strong manner.

II. DATA MINING TECHNIQUES

Data mining techniques are the process of identifying data patterns and trends in order to get useful information in a huge data set so that you can judge or decide.

There are many data mining techniques have been developing and using in data mining projects; including association, classification, clustering, decision tree, prediction, and Neural Networks etc. Each technique has its rules and methods, which determine the kind of problem it solves. In the following sections, we will examine those data mining techniques briefly.

A. Association

Association is among the well-known data mining techniques, it discovers patterns based on the relationship between variables in the same transaction. It is also known as *relation technique* because it uses the relationship between items and discovers the frequent occurrence of different items that appear with the highest frequencies within the data set.

Association rules use the if-then statements in order to show the probability of relationships between data items or variables within large data sets in various types of databases. Association rule have a number of applications and is widely used to help discover sales correlations in transactional data or in medical datasets [3].

Association is widely used by retailers because it helps to understand the customer purchase behaviors.

Based on historical sale data, retailers may be able to find out that customers always buy crisps when they buy beers, and, therefore, they can put beers and crisps next to each other to save time for the customer and increase sales [4].

Because of its retail origins, association rule often referred to as market basket analysis [3].

B. Classification

Classification technique used to classify a collection of data into deferent groups or classes in order to get accurate prediction and analysis in huge data set.

Classification can used to build up an idea of the category of customer, object, or item in a data set by describing multiple attributes in order to identify a particular class. For example, you can easily categorize buildings into different types (based on the occupancy, or based on the type of construction) by identifying different attributes (structure, height, or unit). Given a new building, you may apply it to a particular class by comparing the defined attributes in the database. These principles can applied to customers, to classifying them by age, gender, and social group.

Additionally, classification can used as a feeder to the result of other techniques, like, decision trees to determine a classification, or clustering to use common attributes in different classifications to identify clusters.

C. Clustering

Clustering is one among the first techniques used in data mining. Clustering process involves an analysis of one or more attributes to identify data that are similar to each other in order to understand the differences and similarities between the data set. The process of clustering sometimes called segmentation because it segments the data into categories in order to identify a cluster of correlating results. For example, we can use clustering technique for book management in a library to keep books that have some kinds of similarities in one shelf and give it a meaningful name for readers to grab books in that topic easily, instead of looking for the whole library.

D. Decision tree

Decision tree techniques could applied as part selection criteria. In addition, to help the use and selection of specific data within the overall structure.
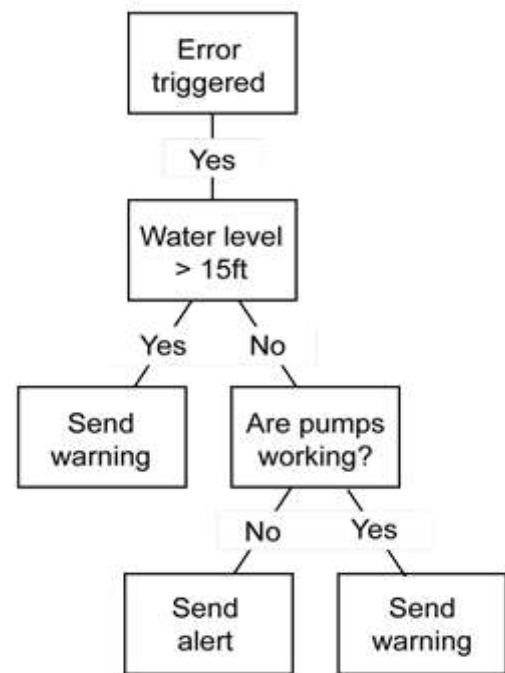
.

Figure 1 Decision tree Structure

In the decision tree, starting with a simple question which has two (or sometimes more) answers Every answer drives to extra question to support classification or identification the data thus, it could be categorized, or a prediction could be made based on each answer.

The decision tree used with classification systems often to attach type information, and with predictive systems, wherever various predictions might be based on previous historical experience that will help propel the structure of the decision tree and the output [5].

E. Prediction

Prediction is a comprehensive topic and goes from predicting the failures of components to understanding the fraud and even the prediction of company profits. It used in combination with the other data mining techniques, prediction includes analyzing trends, classification, pattern matching, and relation. Prediction made by analyzing past events or instances.

Applying the credit card authorization, for instance, combining decision tree analysis of past transactions with classification historical pattern matches to identify whether a transaction is fraudulent. Applying a Match between purchased of flights to the UK and transactions in the UK, it is probable that the transaction is valid [5].

F. Neural Networks

Neural Network is an important technique used by people nowadays. The technique often used in the early stages of the data mining technology. The artificial neural network formed out of the community of Artificial intelligence.

According to [6] neural networks are very easy to use, because they automated to a particular extent and because of this; the user does not expected to have much knowledge about the work or database. It is essential to know the following in order to make the neural network work efficiently.

- How the nodes are connected.
- Use the number of processing units.
- When the training process should stopped.

Neural Networks consists of two main parts – the node and the link

- **The node** – the node is freely matched to the neuron in the human brain.
- **The link** – the link is also freely matched to the connections between the neurons in the human brain

The neural network works as a collection of interconnected neurons that could form a single layer or multiple layers. This kind of formation in the neurons and their interconnections called the architecture of the network.

Neural networks considered as strong predictive modeling technique. However, it is very complex to understand even by experts. It creates very complex, which are very hard to understand fully

The neural network used in various kinds of applications. This used in the business to detect frauds taking place in the business [6].

Data compression technique is gain many methods have used several technique such as discrete cosine transform, discrete wavelet transform included neural network and deep learning methods to enhance compression ratio [7, 8] the main popular neural network in area of data compression is unsupervised learning model such as self-organization feature map SOFM [8, 9] the author presented SOFM and principle of vector quantization and entropy coding.

In data compression also presented special filter band-pass filter [7] before transformation and entropy coding in order to preprocessing data as first stage of image compression model [8].

Bit rate optimization and energy rate reduction or data compression in signal processing utilizes encoding techniques by bits representation compare to original form in several data transmission networks such as in such wireless, mobile, [10, 11] and powerline communications. The techniques of compression are used for the benefit of resources optimization and efficiency required to store and transfer bits [12].

III. CONCLUSION

This paper shows a detailed representation of data mining techniques. Big data is a term of concerning large volumes of complex data sets.
Data mining comprises extracting useful rules or interesting patterns from historical data. The high-performance computing paradigm is required to solve the problem of big data using some techniques that list in this paper for example Clustering, Decision, tree Prediction, and Neural Networks. The mentioned data mining techniques have been a guide to lead the business in a comprehensive and successful manner.

References

[1]. M. Rouse, "association rules (in data mining)," Techtarget, [Online]. Available: https://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining. [Accessed 3 December 2018].

[2]. Zentut, "Data Mining Techniques," Zentut.com, [Online]. Available: http://www.zentut.com/data-mining/data-mining-techniques/. [Accessed 3 December 2018].

[3]. EDUCBA, "ntroduction to data mining," Educba.com, [Online]. Available: https://www.educba.com/course/introduction-and-applications-of-data-mining. [Accessed 3 December 2018].

[4]. EDUCBA, "Data Mining Techniques for Successful Business (Tools, Software)," Educba.com, [Online]. Available: https://www.educba.com/data-mining-techniques/. [Accessed 3 December 2018].

[5]. M. Brown, "Data mining techniques," [Online]. Available: https://www.ibm.com/developerworks/library/ba-data-mining-techniques/index.html. [Accessed 3 December 2018].

[6]. EDUCBA, "7 Important Data Mining Techniques for Best results," [Online]. Available: https://www.educba.com/7-data-mining-techniques-for-best-results/. [Accessed 5 December 2018].

[7]. Hilles, S., & Maidanuk, V. P. (2014). Self-organization feature map based on VQ components to solve image coding problem. ARPN Journal of Engineering and Applied Sciences. Vol. 9,№ 9: 1469-1475

[8]. Hilles, S. M. (2018, July). Sofm And Vector Quantization For Image Compression By Component. In 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE) (pp. 1-6). IEEE.

[9]. Ibrahim, R., Hilles, S. M., Adam, S. M., & El-Ebiary, Y. (2016). Methodological Process for Evaluation of E-government Services base on the Federal Republic of Nigerian Citizen's E-government Services usage. Indian Journal of Science and Technology, 9(28).

[10]. Altrad, A. M. M., Osman, W. R. S., & Nisar, K. (2012). Modelling of Remote Area Broadband Technology over Low Voltage Power Line Channel. International Journal of Computer Networks & Communications, 4(5), 187.

[11]. Osman, W. R. S., Nisar, K., & Altrad, A. M. (2014, August). Evaluation of broadband PLC technology over Malaysia's indoor power line network. In 2014 2nd International Conference on Electronic Design (ICED) (pp. 275-280). IEEE.

[12]. Altrad, A. M., Amphwan, A., & Hilles, S. M. (2018, July). Adaptive Shuffled Frog Leaping Algorithm For Optimal Power Rate Allocation: Power Line. In 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE) (pp. 1-5). IEEE.