Global Scientific JOURNALS

GSJ: Volume 11, Issue 4, April 2023, Online: ISSN 2320-9186
www.globalscientificjournal.com

CONCEPTS AND TECHNIQUES IN DATA MINING

[1]Nwosu P. C. and [2]Onuodu F. E.
[1,2]Department of Computer Science, Faculty of Computer Science
[1,2]University of Port Harcourt, Rivers State, Nigeria

Abstract

Data mining offers an efficient approach to obtain valuable information and knowledge from huge amount of dataset. There has been an enormous increment in data generation and collection capacity because of the new trends in technology that enables capturing and storing massive amount of data. The tremendous growth in data storage has resulted in the need for efficient methods that provides the basis for the transformation of these data into valuable information and knowledge. This paper reviews concept and various techniques that are applied in data mining along with the various processes involved. It also addressed one of the major issues of data mining¬over-fitting and ways to handle it.

Keywords: data mining, database, techniques, machine learning.

1. INTRODUCTION

Data mining is a branch of Artificial Intelligence (AI) that involves extracting of unknown information from huge amount of datasets. It employs complex mathematical algorithms to evaluate the probability of future occurrences; it is also referred to as Knowledge Discovery (KDD). It is an effective technology that can simplify problem-solving in organizations and enable them to pay more attention to information their storage (Deshpande and Thakare, 2010). It entails using various techniques such as statistics, clustering, database systems, machine learning etc to try and identify patterns and relationships in data in order to make informed decision about the data. Essentially, data mining can be used in conjunction with various other methods in different areas like statistics, machine learning, pattern recognition, database system, data warehouse system, visualization, information retrieval, etc. to obtain additional information on data and to used to extract unknown patterns, characteristics, future trends and promotes decision-making process in organizations. Precisely, it is a mathematical process applied in data analysis in different aspects, and in turns categorized and summarized into useful information. It comprises different processes, approaches and algorithms that are applied to determine and obtain meaningful patterns from huge data repositories. As a result of the importance in decision making, nowadays, data mining has gained a lot of attention and has become an important instrument in carrying out different kinds of tasks in organizations (Gupta and Chandra, 2020).

It is an inter-disciplinary field because it is applied in so many different fields. There has been an enormous increment in data generation and collection capacity due to recent trend in technology, which enables capturing and storing massive amount of data. The excessive growth in data storage has made it imperative to have efficient techniques that will assist in transforming these data into valuable information and knowledge (Sadiku et al., 2015). Data mining offers an efficient approach to obtain valuable information and knowledge from huge amount of dataset. It could as well be referred to as a process that involves analysis of data obtained from various sources and transforming them into meaningful information. It involves database systems, artificial intelligence, pattern recognition, visualization, data warehousing, mathematical and statistical processes.

Because there are various techniques applied in data mining as well as a lot of different kinds of information and data presentation methods, it becomes imperative to specify the application boundaries and importance of different data mining methods. It is also fundamental to identify different problem solving approaches applied to the data mining such as classification, regression, clustering etc (Vadim, 2018).

2. REVIEW OF RELATED WORKS

Rawat (2017) presented "Comparative Analysis of Data Mining Techniques, Tools and Machine Learning Algorithms for Efficient Data Analytics". They analyzed different kinds of open source data mining tools including Orange, Weka, RapidMiner, R, KNIME, DataMelt etc. Their main purpose was to discover the most suitable tools and methods for data classification processes that could be applied to data mining. Their analysis showed that better performance could be achieved by combing several tools as well as the techniques.

Kaur et al. (2020) proposed “Systematic Literature Review on Mining Software Repositories”. The review was carried out using three hundred (300) related journal papers. The major purpose was to discover major areas of application, appropriate tools, the required data set, software project types and the applicable Software Development Life Cycle phases and recent trend in Mining Software Repositories.

Madni et al. (2017) presented “Data Mining Techniques and Applications: A Decade Review”. They reviewed various techniques in data mining and their applications. They grouped the available techniques to discover the different fields where data mining could be utilized. In conclusion, they came up with a journal and brief overview of various tools that are applicable in data mining.

Minku et al. (2016) worked on “Data Mining for Software Engineering and Humans in the Loop”. They discussed the responsibility of software engineers in adopting techniques in data mining. They pointed out that the inability of software engineer to get involved in developing models for data mining is a major setback to the acceptance of data mining techniques by software engineering practitioners. They argued that participation of professionals will enhance the performance of the models and as well improve their level of acceptance.

Padhy et al. (2012) presented “The Survey of Data Mining Applications and Feature Scope”. They carried out a concise review of different applications of data mining that would be beneficial to researchers to pay attention to different problems associated to data mining. The study revealed that none of the data mining tools referred to as generic is actually 100% generic.

They were able to determine that most of the customized data mining applications demonstrate high rate of accuracy while general purpose application have drawbacks. Their study revealed that creating and designing a data mining system that can operate dynamically across different fields is highly challenging.

Sharma et al. (2017) worked on "Explorative Study of Web Data Mining Techniques and Tools: A Review". They looked at several online data mining methods and resources used to gather useful information from the internet. Crucial issues related to online data mining were as well reviewed, which comprises the fundamentals of online data mining approaches, tools and their taxonomy. They suggested that their research will beneficial to scholars that are working in the area of online data mining.

3. VARIOUS TECHNIQUES USED IN DATA MINING

Singh and Yadav (2012) asserted that data mining assumes techniques from different research areas including clustering, regression analysis, machine learning and artificial neural network. These areas are illustrated in Figure 1.

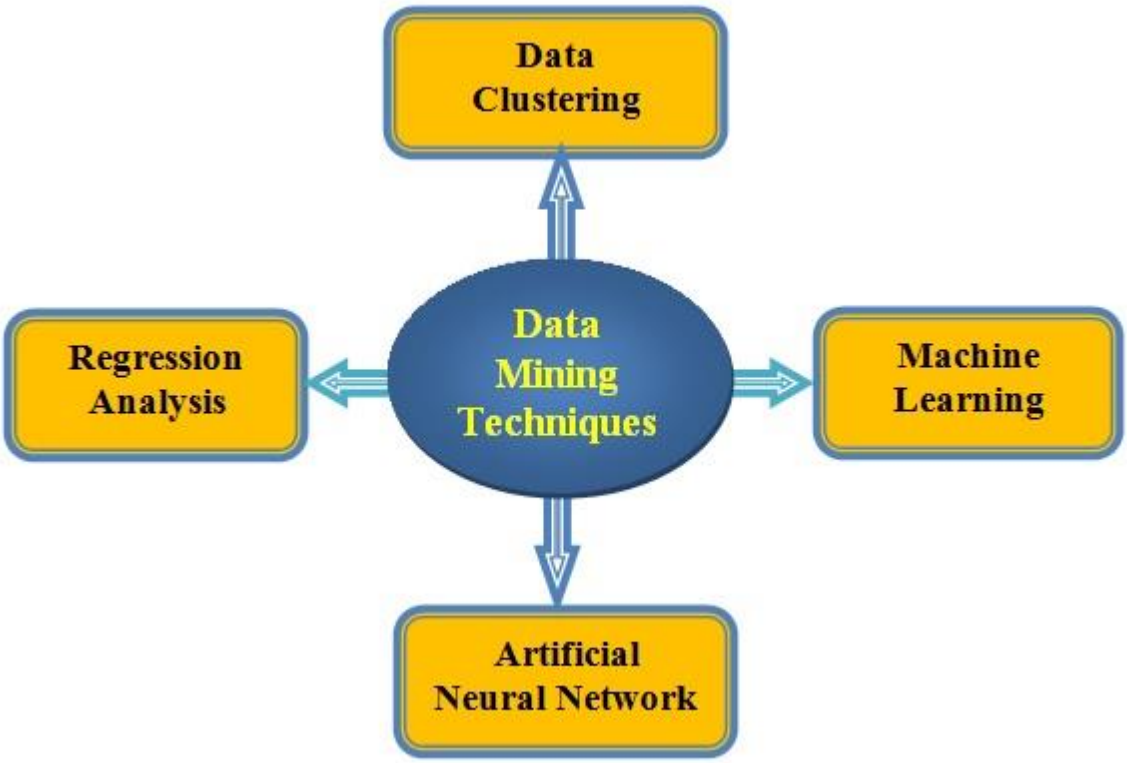

Figure 1: Techniques of Data Mining

a) Data Clustering

Clustering is a procedure that involves categorizing a series of various data depending on their attributes. This enables data miners to easily split data into subsets in order to extract valuable information regarding general data samples and attributes. It involves different methods such as Partitioning method, Hierarchical method, Density-based method, Grid-based method etc.

b) Regression Analysis

Regression analysis is a technique employed to determine the relationship that exist among different variables based on the existence of other features. It is primarily used to specify the likelihood of occurrence of a particular variable. It is a statistical technique mainly used for classification, but the data comprises of both dependent and independent variables. These variables assume any value in a series of real numbers. Regression is of two major types; namely: Simple Linear Regression and Multiple Linear Regression.

c) Machine Learning

Machine learning is a process used in data mining which enables systems to employ algorithms to learn using existing patterns. It is a part of Artificial Intelligence (AI) that helps to make computer system more intelligent by training them to carry out tasks depending on the data they gather. Fundamental machine learning approaches include Supervised Learning, Unsupervised Learning, Semi-supervised Learning and Reinforcement Learning.

d) Artificial Neural Networks

Artificial Neural Network mimics the human brain. It tries to replicate the ability of the human brain to learn new things and adapt to our ever changing environment. It comprises processing units referred to as neurons just like the human brain, these neurons try to replicate the structure and behaviour of the natural neuron. These neurons have two main functions which are collecting inputs and generating outputs (Dongare et al. 2012). Artificial Neural Network algorithm are powerful and flexible, they have the ability to learn by examples. They can be used in analyzing spatial environmental data.

4. DATA MINING PROCESSES

a) Data Collection

This involves gathering and identifying relevant data for the task. Data can be collected from various sources such as database, data warehouse, repositories etc and could be available in different forms. Data could also be obtained from external sources in their raw form and stored for other subsequent steps in the data mining processes.

b) Data Cleaning

Data cleaning involves structuring data to eliminate corrupted data, duplicate values, noisy data, null values etc. Completing this process enables essential information to be gathered for analysis. It involves different processes including data verification, eliminating duplicate data, removing irrelevant data, adding missing values, unifying data types and removing errors.

c) Data Association

Data Association is used to find unusual or relevant correlations among variables in databases. It involves two approaches: the Single-dimensional Association and the Multi-dimensional Association. It is basically used to determine marketing plan and survey in organizations.

d) Data Transformation

Data transformation describes the procedure for transforming unprocessed data into a format appropriate for use. It provides a means of ensuring that data is accurate and consistent. It also helps in enhancing the performance of data mining algorithms through reducing the volume of data and as well scaling the data to a common range of values. Its main goal is to facilitate the mining of valuable information. It involves various processes that includes smoothing, summary, generalization, integration, normalization, reduction, aggregation and attribute construction.

e) Data Classification

Classification is a basic data mining process that categorizes data sets into different classes depending on their attributes. It is a procedure that allows data points from huge datasets to be classified according their characteristics and usage. It is also used to identify sizable groups within a demographic or population sample to enable organization to obtain deeper insights from data. It can be carried out using various algorithms such as Decision Trees, K-Nearest Neighbors, Naive Bayes, Support Vector Machine and Regression.

f) Data Visualization

Data Visualization deals with ways of representing data. It involves the basic methods by which results of data analysis could be presented. It is used to pass the result across to users and executives in organizations. This could be done using charts, scatter diagrams, network diagrams, histograms, maps, smart arts etc.

5. DATA MINING PRIMITIVES

To assume that data mining applications can operate automatically and independently in order to extract hidden information in a specific huge set of data without human intervention is significantly not right. It would be very good to have such autonomous system in data mining, nevertheless, in actual practice; it will be inefficient, may return irrelevant result, may be complex and lacking validity. A more likely possibility is that people can interact with a data mining application using a set of primitives that allow users to flexibly interact with data mining systems in order to direct the mining process. Data mining query languages are designed to incorporate these primitives, enabling efficiency and effectiveness in knowledge discovery. The following primitives are used in defining data mining queries:

a) Task-relevant Data

The first primitive is to specify the data on which to carry out the data mining task. If a user is only concerned about a subset of data in a database, then it will be meaningless to comprehensively try to mine the whole database. It is important to specify the relevant part of the database to be worked on, instead of trying to mine a whole database. This is known as essential attributes.

However, specifying the essential attributes for mining could be a difficult task for users, due to lack of full understanding of the interesting attributes for mining. This could result in relevant data with strong semantic links to them being ignored. With this type of issue, certain procedures can be applied to aid in providing a more accurate characterization of the task-relevant data. It can comprise of operations that have the capability to assess and rate attributes in accordance with their applicability to the defined operation. Furthermore, to

improve the user-specified initial dataset, methods that look for attributes with powerful semantic links can be applied.

b) The Type of Knowledge to be Mined

Specifying the exact type of information to be extracted from a given set of data is very essential as it helps to ascertain the type of task that will be performed on the data, which could include clustering, characterization, regression, classification, association or analysis.

c) Background Knowledge

It is also important for users to indicate the information and the knowledge already at their disposal about the area to be mined. This can focus on basic information at different levels of abstraction. It helps to guide the data mining procedure and assessing the correctness of the information extracted.

d) Interestingness Measures

Interestingness measures relate to operations that are employed in separating irrelevant patterns from information. These could be useful in guiding process of data mining or to assess the mined patterns. Different interestingness measures may be associated with different kinds of knowledge and indicating them goes a long way in the evaluation of the patterns in terms of simplicity, originality, efficiency and applicability. This is important because it helps to additionally limit the generation of irrelevant patterns by the process.

e) Presentation and Visualization of Discovered Patterns

This refers to the formats that are used to present the results of the data mining process. There exist different kinds of data presentation method such as charts, scatter diagrams, network diagrams, histograms, maps etc. To ensure that data mining application is efficient, the system should have the ability to show the output in more than one data representation format and then allow users to determine their preferred method from the available alternatives. This is because; some kinds of data presentation method may be more suitable in certain scenario than others.

6. DATA INTEGRATION IN DATA MINING

Data Integration is the practice of combining data from various locations or sources while putting into consideration the ACID property which are atomicity, consistency, isolation, and durability. It involves the combination data from a few different sources into comprehensible form and still maintains and presents a consistent viewpoint of the data.

It is important that a data mining system is connected to a database or data warehouse system because without such integration, there will be no any platform for interaction. The major emphasis on data mining application design is centred on building efficient and effective systems for mining large sets of data. As such, a data mining system functions well in environments where it is required to interact with systems that can act as a data source such as a database or data warehouse system.

6.1 Data Mining Integration Schemes

Some of the available integration schemes are as follows:

a) No Coupling: This integration scheme requires that the functions of the data mining system do not include any of the functions of the integrated system such as database or data warehouse system. Data is obtained from a specific source and processed using some data mining algorithms and the result is stored in another file.

b) Loose Coupling: In this scheme, the data mining system is allowed to use some of the functions of the integrated database or data warehouse system. Data is obtained from a repository controlled by these systems and data mining operations are carried out on them and thereafter the generated result is stored in a file or in an assigned location in the database or data warehouse systems. It has more advantages because it can access any part of data in the storage by utilizing query processing, indexing, and other system functions. Nevertheless, most loosely coupled data mining systems are memory-intensive.

c) Semi−tight Coupling: This integration scheme allows data mining systems to be integrated with a database or a data warehouse system and further provides effective implementations of a couple of data mining primitives in the database. Some of the primitives are sorting, indexing, aggregation, analysis and pre-computing of some necessary statistical measures like sum, count, max, min etc.

d) Tight Coupling: This integration scheme requires that the data mining system is efficiently combined with a database or data warehouse system. The data mining system is considered as an operational module of an information system.

6.2 Techniques for Data Integration

There are various data integration techniques in data mining. Some of them are as follows:

a) Manual Integration: The integration process is carried out manually by the user without any form of automation. The user interacts directly with the system and the source of data. But this can only be suitable for a limited amount of data as applying it for a huge, sophisticated, and recurring integration will definitely be time-consuming.

b) Middleware Integration: Middleware is a type of software that provides functionalities to software applications beyond those available in the main system. Middleware provides a common platform for all external and internal interactions. It enhances services such as messaging, authentication, communication etc. The middleware obtains data from different locations such as database or information system, normalize and validate the data and before storing it in a central storage location. It is often essential when integrating old and new system.

c) Application-based Integration: This method makes use of software applications that are designed to extract, transform and store data from various data sources. It is an efficient

method that saves time and effort; however, developing such application is not an easy task because technical knowledge is required and it also has deficiency with huge amount of data and many data sources. It is best suited for small amount of data and few data sources.

d) Common Storage Integration: This method enables managing and securing data that are used by multiple programs. It obtains relevant data from different sources, combines them together logically before storing them in a central storage location. This is the method used for implementing data warehouse; thus, it is sometimes referred to as data warehouse integration.

e) Uniform Access Integration: This method allows data from multiple sources to still remain in their original locations and various users are able to have a unified view of the data. It does not require a central data storage location; its major advantage is that there is no movement of data from one location to another.

7. ISSUE OF OVER-FITTING IN DATA MINING

Over-fitting is regarded as a major issue in data mining. It occurs as a result of a statistical model fitting exactly against its training data and fails to fit its future state; as a result, the model is unable to generalize (Ragavi et al. 2018). Unfortunately, the inability of the system to generalize makes it difficult for the system to perform accurately against unknown data. This makes the system inefficient since generalization of a model to new data is actually what makes application of machine learning algorithms to problem-solving possible, enabling it to carry out data classification and predictions.

Over-fitting could be as a result of a number of reasons, so identifying the cause makes it easier to handle. Over-fitted data introduces bias into a model; it becomes relevant only to its dataset and affects the accuracy of making future predictions using new datasets thereby making the model inefficient.

7.1 Causes of Over-fitting

a) Noisy Data: Noisy data are irrelevant information that could be contain in data. If a training set contains a lot of noisy data, the model can easily pick the irrelevant data during training and this will affect the accuracy of the result.

b) Small Training Set: If a model is trained with a very small size data set, such model contains far less data samples to accurately represent all the likely values of the input data. The model will have problem with identifying all possible traits in the data and this lowers the accuracy of the model.

c) Complex System: Complexity of a model could also cause it to over-fit data, so model design should be made to be as simple as possible.

d) Training a Model for Too Long: If model training on sample data continues for a very long time, it begins to gain knowledge of the noise in the data or begins to pick irrelevant

information in the data sets. This makes the model to become closely fitted to the training data, thereby resulting to over-fitting.

7.2 How to Eliminate Over-fitting

It is important to understand how to completely eliminate over-fitting in data mining so ensure accuracy, efficiency and effectiveness in the process. Listed below are some of methods that can be applied to prevent over-fitting:

a) Increase the Training Sample: It is important to train an algorithm using more number of samples. This helps to increase the accuracy of the model by providing more chances to identify the prevailing traits that exist among the input and output variables. In actual sense, this method is very efficient if clean and less noisy data is used as input in the model.

b) Feature Selection: This refers to the process of determining the most relevant variable within the sample dataset for training the model and then discarding the non-relevant or noisy data samples. When developing a new model, there are parameters that are important in determining a specific outcome, feature selection helps to pinpoint such parameters. It helps to simplify a model to establish the dominant traits in the sample data.

c) Data Augmentation: This is a method that artificially increases the amount of data by generating new data points from existing data. Even though data cleaning is important to remove noisy or irrelevant data from the training data set, occasionally noisy data may be added to stabilize a model; however, using this method requires extreme caution.

d) Regularization: Regularization adds a form of constraint on the input parameters with the larger coefficients, which in turn restrains the variation in the model. This method is used if a model is too complex; the features in the model are simply reduced. Knowing the exact feature to eliminate that will not affect the functionality of the system could be difficult, so regularization becomes imperative. Common regularization approaches include L1 regularization, Lasso regularization, and Dropout regularization, they all serve the same purpose but in different ways.

e) Early Stopping: This approach tries to suspend the training process before the model begins to pick the irrelevant information within the sample data set. But this should be used with caution because pausing the process too early could lead to entirely different issue.

f) Ensemble Methods: This method comprises of a set of classifiers such as decision trees. Their predictions are aggregated to identify the most common output. This mainly reduces the variance in a noisy dataset. The most common approaches are bagging and boosting.

8. Conclusion

This paper presents the concepts and various techniques that are applied in data mining along with its applications and tools. A major issue of data mining was also addressed. Different

mining methods like clustering, classification, regression and association rule are efficiently used in organizations based on the kind of information that is required. Relevant information obtained can be utilized in taking decisions in solving software engineering problems.

References

Deshpande, S. P. and Thakare, V. M. (2010). Data Mining System and Applications: A Review. *International Journal of Distributed and Parallel Systems (IJDPS)*, 1(1), 32-44.

Dongare, A. D., Kharde, R. R. and Kachare, A. D. (2012). Introduction to Artificial Neural Network. International Journal of Engineering and Innovative Technology (IJEIT), 2(1), 189-194.

Gupta, M. K. and Chandra, P. (2020). A Comprehensive Survey of Data Mining. *International Journal of Information Technology (IJIT)*, 12(4), 1243–1257.

Madni, H. A., Anwar, Z., and Shah, M. A. (2017). Data Mining Techniques and Applications—A Decade Review. Proceedings of the 23rd International Conference on Automation and Computing, IEEE Publishers, 715-721.

Minku, L. L., Mendes, E. and Turhan, B. (2016). Data Mining for Software Egineering and Humans in the Loop. *Progress in Artificial Intelligence*, 5(4), 307–314.

Kaur, A., Kaur, K., Chopra, D. and Kaur, A. (2020). Systematic Literature Review on Mining Software Repositories. *International Journal of Innovative Science, Engineering and Technology (IJISET)*, 7(1), 2348-7968.

Kukreja, H., Bharath N., Siddesh C. S. and Kuldeep, S. (2016). An Introduction to Artificial Neural Network. International Journal of Advance Research and Innovative Ideas in Education, 1(5), 27-30.

Padhy, N., Mishra, P. and Panigrahi, R. (2012). The Survey of Data Mining Applications and Feature Scope. International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), 2(3), 43-58.

Ragavi, R., Srinithi, B. and Anitha Sofia, V. S. (2018). Data Mining Issues and Challenges: A Review. *International Journal of Advanced Research in Computer and Communication Engineering*, 7(11), 118-121.

Rawat, K. S. (2017). Comparative Analysis of Data Mining Techniques, Tools and Machine Learning Algorithms for Efficient Data Analytics. Journal of Computer Engineering (IOSR-JCE), 19(4), 56-61.

Singh, R. P. and Yadav, P. (2012). Intelligent Information Retrieval in Data Mining. *International Journal of Scientific and Engineering Research* 3(5), 1-6.

Sadiku, M. N. O., Shadare, A. E. and Musa, S. M. (2015). Data Mining: A Brief Introduction. *European Scientific Journal*, 11(21), 509-513.

Sharma, S., Soni, D. and Sharma, A. K. (2017). Explorative Study of Web Data Mining Techniques and Tools: A Review. *International Journal of Computer Science and Technology (IJCST)*, 8(1), 43-47.

Vadim, K. (2018). Overview of Different Approaches to Solving Problems of Data Mining. *Procedia Computer Science*, 123:234–239.