International Journal of Electrical, Electronics and Computer Engineering **13**(1&2): 40-42(2024)

ISSN No. (Online): 2277-2626

A Review of Different Data Mining Techniques While Storing Data

Sonia*, Sangeeta Negi and Shalini
Department of School of Computer Science and Engineering,
Govt. PG College, Dharamshala, Himachal Pradesh Technical University (HPTU) India.

(Corresponding author: Sonia)*

(Received: 30 January 2024, Accepted: 13 April 2024)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Data Mining is a technique of extracting the useful data or pattern from a large amount of unstructured data. Data Mining is a technique in which we will analyse the raw data to identify the useful pattern which is then use for decision making. Here we will discuss the different data mining techniques and finding the best technique among these. There are various types of data mining techniques some of these are Classification, Clustering, Prediction, Genetic Algorithm, Nural Network, Decision Trees, Text Minning etc.

Keywords: Data mining, clustering, classification, genetic algorithm, prediction, neural network, regression, decision trees.

INTRODUCTION

Data Mining is the process of extracting and discovering patterns in large sets involving methods at the interaction of machine learning, statistics, and database systems [Data Mining Curriculum. ACM SIGKDD]. Data Mining is useful technique which is use to finding the useful data from the large amount of data by analyzing the data. The main goal of the data mining is to find the hidden patterns and also find the relation in that data. Data Mining are widely use in the healthcare, finance, marketing, scientific analysis, education sectors, research, etc.

Data Mining is a gradual process which contain several steps (Fig. 1) which includes the following steps:

Data Cleaning: In this step, the data is cleaned by removing the noise and conflict.

Data Integration: In this step, the data is integrated from different sources.

Data Selection: In this step, the analysis of data is done so that we can extract relevant data.

Data Transformation: In this step, gathering of data to transform it into suitable shapes of the mining process.

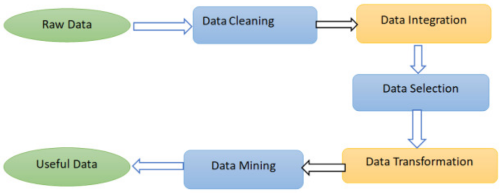

Fig. 1.

Data Mining: In this step, the main goal is extracting the hidden patterns by applying the data mining techniques.

Knowledge representation: In this step, the extracted knowledge is visually represented to the users to understand the results.

DATA MINING TECHNIQUES

Data Mining techniques are used to extract the useful data pattern from the unstructured data. There are various data mining techniques some of these are as follows:

1. Classification: Classification is a data mining technique which is used to identify the patterns from the data. This technique classifies the data items in different classes for the purpose of predict and identify the classes of unknown objects. Let's take example of students, we can classify students according to their marks and predict the grades of a student. This technique is mainly used in the risk assessment, spam filtering, fraud detection etc.

2. Clustering: Clustering is another data mining technique in which elements having same features are grouped together and is called cluster. For example, in shopping websites grocery products, shoes, watches, clothing have different groups. It is mainly used in image processing, pattern recognition and data analysis. There are different types of clustering methods some are these are as follows:

a. Partitioning Clustering Method: In this method the data is divided into non-hierarchical groups. Here the k-value is defined before clustering where k is the number of clusters.

b. Hierarchical Clustering Method: In this method the data is grouped as a set of nested clusters (or as a hierarchical tree).

c. Density-based Clustering Method: In this method the grouping of data is according to the dense region. Highly dense region is combined together and lower dense region and combined separately.

d. Executive Clustering Method: In this method there is no overlapping of data element. Each data element must present only in a single group.

e. K means Clustering Method: In this method K-means an iterative procedure that partition N objects clusters. K-means is perhaps the most widely used clustering method, and especially the best-known of the partitioning-based clustering methods that uses centroids for cluster presentation [Estivil-Castro, 2002].

3. Prediction: Prediction is also a data mining technique which is the combination of clustering, classification and other data mining techniques. It is similar to classification but the data item does not divide into different classes. Prediction also analyzes the past data or event and the predict the future event.

4. Regression: Regression is a data mining technique which is used to establish a relation between dependent and one or more independent variable. Dependent variable is also known as response variable and independent variables are also known as predictor variables. Regression technique is generally used in demand forecasting, price optimization, etc.

There are various types of regression techniques.

a. Single Linear Regression: In single regression technique there is only one independent variable and the relation between independent and dependent variables are linear.

b. Multiple Linear Regression: In multiple regression technique there is more than one independent variable and the relation between dependent and independent variables are linear.

For example: The crop yield is a dependent variable which depends on some dependent variables like temperature, rainfall and many more.

c. Logistic Regression: In logistic regression technique there is more than one independent variables and the dependent variable is categorial or binary (0, 1).

5. Association Rule Mining: This data mining technique is used to identify pattern or association among variables in a data set. In this technique we analyze the frequency of co-occurrence of variables in dataset and then identifying the patterns that occur most frequently. This technique is mostly used in the market based analyzes like the shopkeeper analyzes that the customers were buying butter with the bread. This technique also known as associative classification. Association rule analysis has three main techniques which are lift, support and confidence.

6. Text Mining: Text Mining is a data mining technique which is used to analyzing and extracting useful information from unstructured textual data like emails, customer reviews news articles and social media reviews.

7. Artificial Nural Network: Neural network is a type of data mining technique which is taken from the human nervous system that contain billion of neurons. Neural network filters the raw information from a data warehouse which helps user to take an effective decision. A neural network consists of interconnected nodes that is used to process information. It consists of three layers which are input layer, hidden layer and output layer each layer has its own functions. Nural Network is based on the concept of back propagation which is used to minimizes the error between the predicted output and the actual output. It has an ability to handle noise and missing data. It is one of the best data mining algorithm.

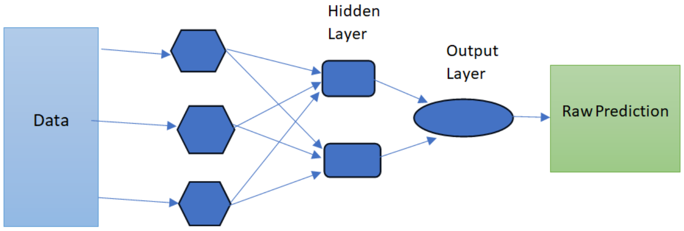

Fig. 2 (Nural Network Model).

Some methods of neural network in data mining are:
a. Feed-Forward Neural Network
b. Feedback Neural Network
c. Self-Organization Network

8. Generic Algorithm: This algorithm is similar to neural network which is also based on the concept of biology. Over generations, the genes transfer one generation to next generation. Genetic Algorithm equates the principle of natural evolution that is survival of the fittest. The real-life example of the genetic algorithm is like some of our features comes from our parents or grandparents. In general Genetic algorithms is a search algorithm based on the natural selection and genetics (Goldberg, 1989).

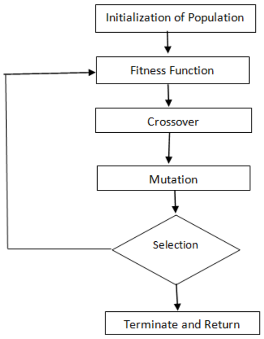

Fig. 3 (Process of Genetic Algorithm).

9. Decision Tree: Decision Tree is one of the simplest and easiest techniques of the data mining because the trees are can be easily understood to everyone. Trees are drawn from top (that is root) to bottom (that is leaves). A decision tree contains the nodes and edges.

Nodes represents any decision or event and edges represents the possible outcomes. Complex situation can be easily solved by the decision trees. The best example of the decision tree is the binary tree. In the Fig. 4 (decision making) given below define the decision tree to find the age of a person or we can say to find the category of a person (*i.e*. baby, child, young adults or old adults) according to their age.

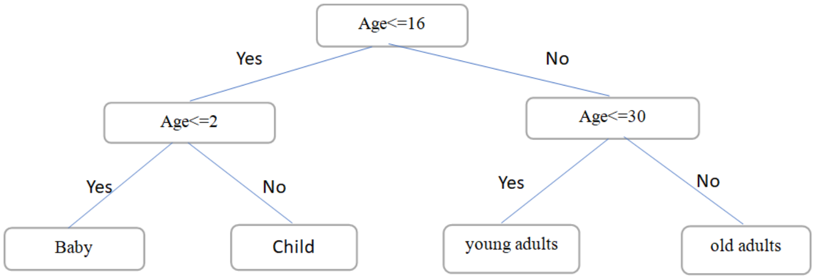

Fig. 4 (Decision Tree).

10. Memory-Based Reasoning: Memory-Based Reasoning is a data mining technique which uses the known instances or objects of any model to assume or predict the unknown instances by maintain the datasets of the instances. The distance function and the combination function are the two main components of Memory Based Reasoning. The distance function is used to find the distance between the newly arrived record and the records stored in datasets and then the combination function is used, which is used to combine the results of various distance function to produce the final result.

Memory-Based Reasoning is implement using the following steps:

1. The first step is to select the historical data from the datasets.
2. Then the next step is to compose the historical data in the best way.
3. After that the two important functions that is the distance function and the combination function is used to produce the final results.

CONCLUSION

Data Mining is the process of finding the different patterns from the raw data so we will use that for the decision making. There are various Data Mining Techniques some of the important techniques are classification, clustering, regression, neural network, genetic algorithm, decision tree etc. All the techniques have their different features. Classification technique is used to categorize the available entities. Clustering is opposite to the classification which makes the group of similar entities. Regression establishes the relation between independent and dependent variable. Genetic Algorithms are one of the important technique of data mining which is based on the natural selection and genetics idea. Neural Network is the technique in the data is passed through three layers to filter the data. At last, these algorithms are used for same purpose that is for data mining but use different methods or techniques and can have different uses or application.

REFERENCES

Data Mining Curriculum. ACM SIGKDD. 2006-04-30. Achieved from the original on 2013-10-14.

V. Estivill-Castro (2002). Why so many clustering algorithms: a position paper. *ACM SIGKDD Explorations Newsletter*, *4*(1), 65-75.

Goldberg, D. E. (1989). Genetic Algorithms in Search, Optimization, and Machine Learning. Reading. MA: Addision-Wesley.