Review Paper on Data Mining Techniques and Applications

Mr. Sharan L Pais[1], Sujan P S[2], Suraj Ankolekar[3], Sushma K N[4], Swetha S[5]

Faculty, Department of Information Science and Engineering[1]

Students, Department of Information Science and Engineering[2,3,4,5]

Alva's Institute of Engineering and Technology, Mijar, Mangalore, Karnataka, India

Abstract: *Extraction of hidden and valuable patterns and information from data is known as data mining. A new technique called data mining aids firms in making proactive, knowledge-driven decisions by being able to forecast future trends and behaviors. The purpose of this study is to demonstrate the data mining process and how it can assist decision-makers in reaching better conclusions. Data mining is actually very beneficial for any firm that has a large volume of data. Data mining speeds up the performance of ordinary databases. Due to the wise choices made with the use of data mining, they also aid in boosting profits. This essay demonstrates the many procedures involved in data mining and how they can be applied by different data.*

Keywords: Data Mining

I. INTRODUCTION

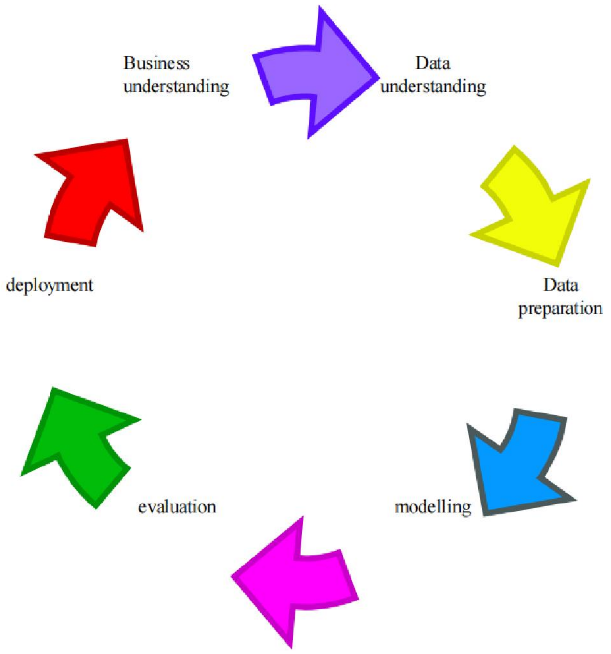
Data mining is the process of obtaining accurate, previously undiscovered, and useful information from enormous sets of data. Making important business decisions using the information that has been collected from the data is the goal of data mining. In order to extract meaningful business information from a big volume of data, data mining assists end users. This expression is frequently used to refer to any type of extensive data processing. The outputs of the mining process ought to be reliable, original, practical, and clear. Exploratory data analysis, a subfield of statistics, and knowledge discovery and machine learning, a subfield of artificial intelligence, are all related to data mining. In the first half of this essay, data mining is briefly introduced. The second section provides examples of the data mining method, and the third portion examines several data mining methods. The fourth portion focuses on various Data Mining application areas, and the fifth section examines the conclusion and potential future applications.

II. REVIEW OF LITERATURE

KDD was defined as "a nontrivial process of recognising legitimate, unique, potentially helpful, and finally intelligible patterns in data" by Fayyad et al. (1996) in their work "From data mining to knowledge discovery in databases." Any collection of true facts that are available in an electronic format were used to expand the definition data. Patterns are models that are stated in a language as a subset of data. The patterns must be true and able to be modeled for any new data in order to be valid. The process consists of several processes, ranging from data preparation through knowledge augmentation, all of which are repeated until the desired results are obtained. Nontrivial suggests that, in order to distinguish it from the conventional computation of values, there should be some form of inference computation. In their study published in 1997[4], Fayyad and Stolorz described According to KDD, mining is just one phase of a broader technique for extracting priceless knowledge from data. This process also uses several additional algorithms. [5] Charles et al1998 In the current modern age, where conventional marketing channels like mass marketing are showing a decline trend, data mining has been advocated as a useful technique for direct marketing in order to boost product marketing. By using data mining, we may identify buying trends from a client list and identify potential buyers. Data mining as a direct marketing tool has proven to be more profitable than conventional mass marketing strategies because it only targets potential customers. In their article "A survey of data mining and knowledge discovery tools," Michael Goebel et al. (1999) presented a broad overview of typical knowledge discovery tasks and several approaches to address these.

DOI: 10.48175/568

III. PROCESS OF DATA MINING



The data mining process is a multi-step process that cannot be finished in one go. In other words, it is not that easy to extract the necessary information from the vast amounts of data. It is not confined to any one industry. In essence, the method has developed from knowledge discovery procedures that are commonly utilized in industry. The main goal of the data mining process is to increase the effectiveness of massive data projects. Data cleansing, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge representation are among the operations that must be finished in the specified order.

3.1 Business Understanding

The business comprehension phase focuses on comprehending the project's goals and requirements, evaluating the current environment, and setting data mining objectives from a business perspective. We create the project's initial plan during this phase. This phase involves a number of activities, including establishing company goals, assessing the present environment, choosing a data mining aim, and creating a project strategy.

3.2 Data Understanding

Initial data collection, data description, data exploration, and the assurance of data quality are all included in this phase. It primarily focuses on discovering the fundamental properties of the data, such as its structures, its quality, and any particularly intriguing subsets. The primary duties involved in this phase are gathering preliminary data, describing data, examining data, and verifying data.

3.3 Data Preparation

All the tasks necessary to create the final data collection in the correct form are included in this step. This phase's primary tasks include selecting data, cleaning data, integrating data, and transforming data. Data is prepared for production at this phase. This step results in a data set that can be utilized for modeling.

3.4 Modeling

We choose modeling methodologies, modeling parameters, and evaluate the developed model in accordance with the business objectives during the data modeling process. More intricate models that are appropriate for the data can be used once a deeper comprehension of the data has been achieved (typically through pattern identification sparked by examining model output). The different tasks carried out during this phase include choosing the modeling technique, creating the test design, creating the model, and evaluating the model. The actions below are done in order to create a satisfactory model.

3.5 Evaluation

In terms of data analysis, this step validates the model. This step involves verifying the model and the modeling process while keeping the business goals in mind. The different tasks carried out at this phase include reviewing the process and analyzing the results. The company objectives should be taken into account while evaluating the evaluation outcomes. The decision to move the model during the deployment phase is either go or no-go.

3.6 Deployment

The knowledge that was acquired in the form of a model must now be arranged and presented in a way that business users can use it. This procedure can be as straightforward as producing a report or as sophisticated as repeatedly putting the repeatable data mining technique into practise. The execution stage is now. Plan deployment, plan implementation, and other duties are all part of this phase.

IV. DATA MINING TECHNIQUES

4.1 Association

One of the well-known data mining approaches is association, which looks for patterns based on the connections between variables in a single transaction. Because it makes advantage of the relationships between things to identify the different objects that occur most frequently in the data collection, it is also known as the relation technique. In order to display the likelihood of associations between data items or variables inside big data sets in different types of databases, association rules employ if-then expressions. In order to find sales correlations in transactional data or in medical datasets, association rules are frequently utilized.

4.2 Classification

In order to provide effective prediction and analysis in large data sets, classification technique is used to categorize a collection of data into distinct groups or classes.

Classification is a technique for identifying a specific class of customers, objects, or items in a data set by defining a variety of criteria. For instance, by identifying several qualities, it is simple to classify buildings into various classes (depending on occupancy or type of construction) (structure, height, or unit). By contrasting the defined qualities in the database with a new building, you can apply it to a certain class. These principles can be used to categorize customers according to their age, gender, and socioeconomic class.

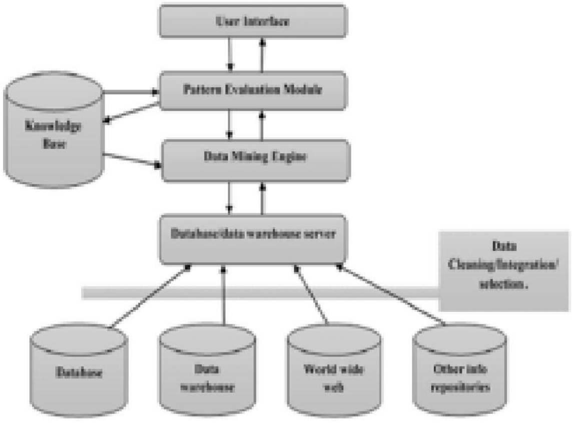
4.3 Clustering

One of the earliest methods used in data mining is clustering. In order to comprehend the differences and similarities between the data collection, the clustering method analyses one or more qualities to discover data that are similar to one another. In order to find a cluster of findings that correlate, the clustering process—also known segmentation—segments the data into groups. For instance, we can use the clustering technique to organize books in a library so that readers can easily find books on a particular topic without having to search through the entire collection by grouping comparable volumes together on one shelf and giving it a descriptive name.

4.4 Decision tree

Techniques for decision trees could be used as part of the selection criteria. Additionally, to facilitate the use and choice of particular data within the overall structure.

V. ARCHITECTURE OF DATA MINING



5.1 Knowledge Base

It functions as the start of the entire data mining process. It serves as a guide for looking for information or determining how fascinating the patterns that emerge are. Concept hierarchies, which group qualities or their values into different degrees of abstraction, may be a part of this form of knowledge.

5.2 Data Mining Engine

It is a fundamental part of the mining system and includes all the modules required for carrying out data mining operations, including characterization, prediction, cluster analysis, outlier analysis, and evolution analysis.

5.3 Pattern Evaluation Module

This module typically has interestingness measures attached to it. It maintains constant communication with the data mining engine to keep its attention on looking for intriguing patterns. Depending on the data mining technique employed, it frequently employs thresholds to separate out identified patterns or may use a pattern evaluation module combined with a mining module.

5.4 User Interface

The module serves as a conduit between the data mining system and the users. It makes it simple and effective for users to interact with the system without worrying about the complexities of the process itself. Data sources (www, data warehouse, database, other repositories): These are the actual sources of data, and data mining success necessitates a vast amount of historical data. Data is generally kept by businesses in databases or data warehouses. Data warehouses occasionally contain many databases, text files, or spreadsheets. Another enormous source of data is the web. Database or data warehouse server: It contains actual data that is scheduled for retrieval. Its main duty is to retrieve data when users request it.

VI. DATA MINING TECHNIQUES

6.1 Classification

Data records are classified using classification algorithms into one of a number of predetermined classes. They operate by building a model from a training dataset made up of example records with predetermined class labels. Classification is a technique for supervised learning [3]. The classification of data involves two steps. A model is created in the first stage by examining data tuples from training data that have a specific set of attributes. The class label attribute value for each tuple in the training data is known. The model can be used to categorize the unknown tuples if its accuracy is deemed satisfactory [4]. Different classification models, such as Bayesian Classification, Neural Networks, Support Vector Machines (SVM), and Classification by Decision Tree Induction, can be utilized.

DOI: 10.48175/568

6.2 Clustering

Clustering is the process of grouping data into clusters so that the data objects inside each cluster are comparable to one another are grouped together. There may be numerous ways to categorize data objects; there is no one right method for clustering. Unsupervised learning is a type of clustering in which no class labels are given. Data records should instead be classified according to how similar they are to other records. For instance, clustering can be used to create profiles of persons who replied to past mailing campaigns. These profiles can then be used to anticipate response and to filter mailing lists to get the best response.

6.3 Prediction

This method demonstrates how specific data properties will behave in the future. For instance, based on analysis of customer purchase transactions. A data item is mapped to a real valued prediction variable using regression. The relationship between one or more independent variables and dependent variables can be modeled using regression analysis. Prediction models are essentially functions with continuous values that are used to forecast missing or unavailable numerical data values rather than class labels. Identification of distribution trends based on the facts at hand is also included in prediction. The statistical technique of regression analysis is most frequently applied to numerical prediction. Regression techniques of many kinds, including linear regression, multivariate linear regression, nonlinear regression, and multivariate regression, are employed.

6.4 Association Rule

To extract the commonly used items from the vast data set, association and correlation are performed. According to association rules, a set of objects is associated with a different range of values for a different set of variables. The goal of association is to find patterns in data that are based on connections between items from the same transaction.

Due to its characteristics, association is also known as "relation technique." This data mining technique is used in market-based analysis to pinpoint a set, or sets of sets, of goods that customers frequently buy at the same time [18]. Business decisions like catalog design, cross-marketing, and customer shopping behavior analysis are all aided by this kind of technique [17]. For instance, each time a customer purchases audio

6.5 Neural Networks

The nonlinear predictive model known as a neural network is similar to biological structure and learns through training. Neural Networks offer predictions based on newly interesting situations and generate "what if" scenarios. These are ideal for inputs and outputs with continuous values. For instance, a neural network can be taught to determine the likelihood of any disease based on a variety of parameters. For prediction or forecasting purposes, neural networks are excellent at spotting patterns or trends in data [14].

6.6 Time Series Analysis

A time series is a collection of data that is collected over time at regular intervals, such as daily sales, and then analyzed using statistical methods to find patterns within the data. Using a model to produce predictions (forecasts) for future events based on previously recorded data is known as time series forecasting [19]. Consider the stock market. This approach

VII. APPLICATIONS OF DATA MINING IN VARIOUS FIELDS

7.1 Application of Data Mining in Health Care:

Data mining can be very helpful for the healthcare system, but it depends on having access to quality data. It is used in healthcare to determine a disease's diagnosis and prognosis, as well as the relationships between various diseases. In order to provide patients with better and more affordable care, doctors can find effective and best practises. Data mining offers a methodology and tools for transforming data into information for effective decision making because there is a tremendous amount of healthcare data that must be handled and examined.

7.2 Application of Data Mining in Educational Systems

Data mining in the educational system is an area that is still developing, but scholars are deeply interested in it. Due to the annual enrollment of millions of students in various institutions, there is an enormous volume of data. By identifying hidden patterns, connotations, and variances, data mining techniques can aid in bridging the knowledge gap in the educational system. This allows decision-making by stakeholders to be more effective.

7.3 Application of Data Mining in CRM

In order to provide a research synopsis on the use of data mining techniques in the CRM domain, data mining in CRM is now the most discussed research topic in industry and academia.

7.4 Application of Data Mining in Market Basket Analysis(MBA)

For market basket analysis, often known as MBA, different data mining techniques are applied. This method aids in determining the relationship between the numerous goods that a consumer places in his shopping cart while out shopping and tracks client buying behaviours. Businesses can give customers with a variety of options based on their buying habits by using data mining tools to detect the customer's buying patterns and behaviour.

7.5 Application of Data Mining in Sports Data:

Sports have also been impacted by data mining techniques. There are a tonne of sports being played, and each sport produces a tonne of statistical data. With relation to event scheduling and participant statistics, this vast amount of data must be kept up to date. Data mining can be used for strategy development, performance analysis, and forecasting.

7.6 Data Mining in Science and Engineering

bioinformatics, genetics, medical, education, and electrical engineering are all examples of engineering. Data mining is considered an interdisciplinary practise because of this. In understanding the mapping relationship between inter-individual variation in human DNA sequences and variability in disease susceptibility is a key objective in the field of research on human genetics. It is particularly beneficial in the detection, avoidance, and treatment of ailments.

7.7 Data Mining in Earthquake Prediction Maps

An earthquake is a quick shift of the Earth's crust brought on by the abrupt release of stress that has built up along an inner geologic fault. The two most common types of earthquake predictions are: projections (from months to years in the future) and near-term prognoses

7.8 Data Mining in Agriculture

a four-parameter analysis, including the year, rainfall, production, and sowing area. Based on the information now available, yield prediction is a significant agricultural challenge that has to be solved. Data mining techniques like K Means, K closest neighbour (KNN), artificial neural networks, and support vector machines can be used to tackle the yield prediction problem.

7.9 Data Mining in Cloud Computing

In cloud computing, data mining techniques are applied. Utilizing cloud computing to execute data mining techniques would enable users to access useful data from a practically integrated data warehouse while saving money on infrastructure and storage. Internet services that rely on clouds of servers are used by cloud computing to handle jobs. The data mining method is used in cloud computing to provide users with services that are effective, dependable, and secure.

7.10 Data Mining in Bioinformatics

Because it is data-rich, data mining is particularly suited for bioinformatics. Massive datasets amassed in biology and other allied fields of the life sciences, such as medicine and neurology, can be usefully mined to extract knowledge. Gene discovery, protein function inference, disease diagnosis, disease prognosis, disease therapy optimization, reconstruction of protein and gene interaction networks, data purification, and prediction of protein sub-cellular localization are examples of applications of data mining to bioinformatics.

VIII. CONCLUSION

The author's analysis of data mining is as follows: Data mining is the process of finding, examining, and sorting through massive amounts of data to find relationships, patterns, or scientific correlations. SDM is a method for extracting interesting, practical, and non-trivial patterns from huge geographic datasets (Geographic Data Mining). Due to the intricacy of many information types, spatial linkages, and spatial auto-correlation, obtaining comparable patterns from huge datasets must be more challenging than extracting patterns from conventional numeric or classified data. It was underlined how crucial it is to distinguish between geographical data mining and traditional data mining, as well as the significant contributions made by spatial data mining research. The neural network functions as a group of interconnected neurons that may form one layer or more. The architecture of the network is this type of formation in the neurons and their connections. Neural networks are regarded as a powerful predictive modeling method. Even specialists find it difficult to understand, though. It produces incredibly intricate things that are challenging to fully comprehend.

The neural network is employed in many different applications. This is employed in the business to find scams occurring there [6].

In order to improve data compression, a number of techniques, including discrete wavelet transform, discrete cosine transform, and neural network methods, have been used.

REFERENCES

[1]. R. Tamilselvi and S. Kalaiselvi, “An Overview of Data Mining Techniques and Applications Keywords: Data mining Techniques; Data mining algorithms; Data mining applications 1. Overview of Data Mining,” Int. J. Sci. Res., 2013.

[2]. R. S. J. d. Baker, “Data mining,” in International Encyclopedia of Education, 2010.

[3]. P. VIKRAMA, P and Radha Krishna, “Data Mining Data mining,” Min. Massive Datasets, 2005.

[4]. F. A. Hermawati, “Data Mining Data mining,” Min. Massive Datasets, 2005.

[5]. A. Twin, “Data Mining Data mining,” Min. Massive Datasets, 2005.

[6]. B. A. B. Ii, “Data Mining Data mining,” Min. Massive Datasets, 2005.

[7]. Y. Chen, D. Hu, and G. Zhang, “Data mining and critical success factors in data mining projects,” in IFIP International Federation for Information Processing, 2006, doi: 10.1007/0-387-34403-9_39.

[8]. K. M. Raval, “Data Mining Techniques | Data Mining Articles,” Int. J. Adv. Res. Comput. Sci. Softw. Eng., 2012.

[9]. https://sites.google.com/site/ijcsis/ ISSN 1947-5500

[10]. M. Rouse, "association rules (in data mining)," Techtarget, [Online]. Available: https://searchbusinessanalytics.techtarget.com/definition/assoc

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

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

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

DOI: 10.48175/568