A Review on Comparative Analysis of Data Mining-Algorithms and Techniques

Prerna Jain

Abstract: ***The subfield of technologies within the discipline of computer science is known as "data mining" Moreover, statistics are put to use in order to identify trends within the database. Essentially, the primary goal of the data mining methods is to extract from the database any information that will be valuable in the future or that will be needed at some point in the future and transform it into a usable format. Understandable framework for future usage. There are several approaches to choose from, each of which may be helpful in carrying out data mining operations in an effective manner. There is some information accessible in this document concerning comparative studies or analyses of various data mining approaches and certain data mining algorithms. This information may be found in this publication. This means that reading this paper will assist us in gaining knowledge regarding data mining technologies, data mining algorithms, as well as some data of corporations that have upgraded or enhanced their profits by modifying their data mining technologies and have obtained impressive output as a result of doing so.***

Keywords: Data mining techniques, data mining technology, computer science, statistics, algorithms

1. Introduction

The advancement of information technology has resulted in the production of a great deal of information and information in a variety of domains. The study of information as well as the development of information technology have produced a method that may be used to store and use this vital knowledge for future decision-making. Facts mining is the process of extracting meaningful data and patterns from large amounts of unstructured data. In certain contexts, it is also referred to as the process of collecting information, deriving information from data, extracting information, or doing information analysis. The intelligent technique of utilizing a vast quantity of information to acquire valuable data is known as data mining. This approach is utilized for searching [1]. The objective of this method is to recognize patterns that were not recognized to be there before. If these established patterns can be employed in the same way to make specific judgments, then the company owners will be able to enhance their operations. Data mining algorithms result in the development of methods that have been in use for at least ten years, but which are just now being put to use as resources that are mature, dependable, and intelligible in addition to being more sustainable than the traditional approaches [2].

Knowledge Discovery in Databases, abbreviated as KDD, is the process of extracting nontrivial facts from statistical information in databases that are likely to be of value. These details might be implicit, newly discovered, or both. Data mining is an integral aspect of the knowledge discovery process, despite the fact that KDD and data mining are often treated as interchangeable terms [3].

Three steps involved in KDD are:

1) Exploration:
    a) Records are scrubbed, converted to a different format, and basic information is extracted as the first stage in the statistics research process.
    b) Variables are identified, and then the character of the statistics is established depending on the problem.
2) Pattern Identification:
    a) The second phase, structural pattern identification, comes after the information has been investigated, improved, and explained with regard to the specific variables. Find and choose the patterns that have a good track record of predicting outcomes.
3) Deployment:
    a) The Patterns are used in order to get the desired results.

2. Literature Review

This article offers details on a variety of data mining techniques, as well as algorithms for comparative research. In this article, we will discuss the many different kinds of data that may be mined using the aforementioned method. In addition to this, we have gone through the documentation of certain data mining methods such as integration, organizational principles, and so on.

According to Han et al. (2011), data mining is a method that may be characterized as the process of examining archives and extracting information that is ambiguous but helpful.

Data mining has the ability to uncover previously undiscovered links and reveal previously concealed patterns and trends by delving through large amounts of data, as stated by Sumathi et al (2006).

According to Hui et al. (2000), the goal of data mining may be broken down into four distinct categories, depending on the kind of work that is carried out: integration, fragmentation, integration, and retrieval.

3. Related Work

There is a huge variety of data that may be obtained all across the globe. Within the scope of this work, we have discussed several varieties of data that may be extracted with the assistance of this method. In its most basic form, data mining involves more than one kind of media or data. The mining of data should always be done in tandem with some kind of data storage. However, algorithms and techniques may have a degree of adaptability when they are applied to certain categories of data. In point of fact, different kinds of information provide a vastly different set of obstacles. Information, including relational information, related data, and targeted information, as well as trade information, informal and well-structured repositories like the World Wide Web, high-profile information like location data, multimedia information, timeline information, and textual information, and even flat files can all be mined using data mining [3-5]. Here are some instances in further detail:

- Flat Files: The most typical source of data mining algorithms for data mining is in the form of flat files. It is a data file with the most fundamental level, stored in binary format, and includes a data mining method that follows a predetermined pattern. Details may include things like transactions, statistics on timelines, and scientific ratings.
- Relationship database: the connection database comprises a collection of tables that include the values of the trademark or the values of the symbols from the business relationship. This information may be retrieved using the relationship identifiers. Tables are made up of columns and rows, with the columns representing the qualities and the rows representing the multiples.
- Database: A database, also known as a database, is compiled from a large amount of data and analyzed so that it may be used to its maximum potential inside the same integrated system. The data repository provides users the chance to do analysis on records derived from a variety of sources while using the same framework.

4. Methodology

There is a wide variety of data mining software that may be downloaded or updated. Some of them are specialized structures that are constrained to a certain information source or that are restricted to particular data mining procedures [4]. It is possible to classify the framework of data mining according to a variety of criteria, including the following:

Classification of the several kinds of data sources that were mined: This section provides a classification of data mining systems based on the many types of records that are handled, including multimedia data, chronological data, text data, World Wide Web data, and geographical data, amongst others.

Classification in accordance with the data model that was drawn on: This part classifies data mining systems in accordance with the relevant reality model, such as the relational database, the object-oriented database, the database, transactions, and so on.

Classification based on the newly acquired information is as follows: This categorization organizes data mining systems according to the data kinds or operations that are performed on the data, such as demographics, discrimination, mergers, segregation, mergers, and so on.

Classification based on the mining approaches that were applied: The data mining framework chooses and delivers certain tactics. This categorization divides data mining systems into subcategories based on the technique of data analysis that is used, such as machine learning, neural networks, genetic algorithms, statistics, observations, database or data storage, and so on as represented in **Table 1**.

Table 2 represents different classifiers their advances and their disadvantage.

There are a wide variety of search options at one's disposal. On the other hand, we shall discuss the data mining tools that are the most significant. In addition to that, a review of the tools used over the last several years [2-3].

Table 1: Tools of data mining and their uses in year-wise pattern

Data Mining Tools	2016	2017	2018	2019
R	38.5%	46.9%	49%	52%
Rapid Miner	44.2%	31.5%	32.6%	32.8%
SQL	25.3%	30.9%	35.5%	34.9%
Python	19.5%	30.3%	45.8%	52.6%
Excel	25.8%	22.9%	33.6%	28.1%
KNIME	15.0%	20.0%	18.0%	19.1%
Hadoop	12.7%	18.4%	22.1%	15.0%
Tableau	9.1%	12.4%	18.5%	19.4%
SAS	10.9%	11.3%	5.6%	9.12%
Spark	2.6%	11.3%	21.6%	22.7%

Standard Comparison of Classifiers Algorithms

AI and ML employ the data mining algorithm. Data miners have too many algorithms

Table 2: Advantage and disadvantages of the classifier

Classifier	Method	Advantages	Disadvantages
The Support Vector Machine	These supervised learning models use data-analysis methods.	1. Accurate. 2. Model complex nonlinear decision boundaries.	1. Complex algorithm and memory. 2. Kernel selection is tough.
K Nearest Neighbor	Most of an object's k closest neighbors vote to classify it.	1. Analyzable. 2. Simple implementation. 3. Uses local information for highly adaptable behavior.	1. Large storage needs. 2. Dimensionality-prone. Slow categorizing test tuples.
Artificial Neural Network	Its structure depends on external or internal information.	1. Less formal statistical training. 2. Data tolerance.	1. Unpredictability. 2. Oversizing.
Bayesian Method	Algorithm estimates class conditional probabilities.	1. Naïve Bayesian classifier simplifies calculations. 2. Accurately and quickly.	1. Class conditional independence assumptions no data.

Comparison of Data Mining methods

Classification:
Separation is a common data mining technique that uses pre-programmed samples to update a model that can distinguish data amounts. Always employs neural network configuration techniques. Division requires reading and division.
Categorization models:

Bayesian classifier
1) Neural networks
2) Vector support machine (SVM)
3) Tree planting Separation
4) Organizational separation

Clustering:
One way to describe integration is as the process of classifying groups of things that are analogous to one another. We are able to take a sample from a universal distribution, as well as identify a connection between mathematical symbols, when we use procedures that include confluence. These approaches allow us to locate tiny and compact spaces inside the space of an item. Integration is a pre-processing approach that may be used for the subset of adjectives and adjectives.
Types of meeting methods:
1) Divide Methods
2) Roads designed for congestion
3) Grid-based approaches
4) Model-based approaches

Predicting:
The return function supports configurable prediction settings. One or more of the relationships between the neutral variables and the systematic variables may be preserved via the use of reversal analysis. The response variable in the data mineral is what we choose to guess, whereas the random variable in the data mineral is an aspect that has already been experienced.
Types of retrieval methods:
1) Corresponding Modifications
2) Reduction of Multivariate Linear lines
3) Indirect postponement
4) Multivariate Nonlinear Regression

Association rule:
Integration and adjustment are often used to identify the results of establishing a common object amongst many huge data sets. This sort of acquisition assists businesses in making educated judgments, such as those pertaining to catalogue design, opposing marketing, and the study of consumer behavior. The capacity to develop guidelines with a confidence level of less than one is a prerequisite for the Association Rule algorithm family.

Types of organization law:
1) Multilevel organization law
2) Multi-sectoral law
3) Measurement organization law

Neural Networks:
The Neural Network is composed of a set of plug-ins that may be used with a variety of different devices, and all of its communications are cumbersome. It is possible to utilize neural networks to recover patterns and identify patterns that are too complicated to be observed by people or other computer systems. Neural networks offer a huge potential for extracting meaning from complex data, and they can be used to do so.

Types of neural networks:

Back Propagation
The use of data mining in the data management system is a fairly new development. Current data mining is done primarily on simple numeric and categorical data. In the future, data mining will include more complex data types. In addition, for any model that has been designed, further refinement is possible by examining other variables and their relationships. Research in data mining will result in new methods to determine the most interesting characteristics of the data. As models are developed and implemented, they can be used as a tool in the data management system [5-8]. **Table 3** represents the different techniques and their average accuracy.

Table 3: Techniques and their Average Accuracy

Techniques	Average Accuracy
Classification	83.10%
Clustering	82.07%
Prediction	82.76%
Association rule	74.72%
Neural Networks	82.85%

5. Conclusion

Mining data is crucial in a variety of commercial fields for identifying patterns, making forecasts, obtaining information, and so on. The applications and methods of data mining, including as segmentation, consolidation, and other similar processes, are helpful in identifying patterns that may be used to predict future trends in firms that are expanding. Because data mining has such a broad range of applications across the data processing sector, it is widely regarded as one of the most important aspects of database management and information, and it is particularly regarded as one of the parameters that hold the greatest potential for fostering the growth of a number of sub-sectors within the field of information technology. In this document, information is provided concerning comparative analyses of different data mining methods and algorithms, as well as different data mining platforms. These analyses may be classified into a variety of procedures.

References

[1] Mrs. Bharati M. Ramageri, Diy Process And Applications (7 April 2020).
[2] Jiawei Han And MichelineKamber Jian Pei, Analysis Of Thoughts And Diy Skills (Feb 2016).
[3] M.S Chen, J.Han, And P.S. Yu, A Comparative Study Of Details Of Details And Details (July 2017).
[4] Mr. Nilesh Kumar Dokania And Ms. Navneet Kaur, A Comprehensive Study Of Different Skills Of Data Information (May 2018).

[5] BerinaAlic, LejlaGurbeta, AlmirBadnjevic, Review Of Machine Learning Activities (June 2017).

[6] Smit Garg, Arvind K Sharma, Comparative Analysis Of Data Details (July 2013).

[7] Keshav Singh Rawat, Comparative Comprehension Of Data Conduct, Algorithm Materials And Machinery For Actual Data Analysis Reading (July 2017).

[8] Mr. Nilesh Kumar Dokania And Ms. Navneet Kaur, A Comprehensive Study Of Different Skills Of Data Information (May 2018).