International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-9, Issue-6, November 2021
https://doi.org/10.55524/ijircst.2021.9.6.23
Article ID IRPS12033, Pages 100-104
www.ijircst.org

A Review of Data Mining Techniques and Its Applications

Madhav Singh Solanki[1], and Ms Anuska Sharma[2]

[1, 2] SOEIT, Sanskriti University, Mathura, Uttar Pradesh, India

Correspondence should be addressed to Madhav Singh Solanki; madhavsolanki.cse@sanskriti.edu.in

ABSTRACT- Knowledge Discovery in Databases (KDD) is another name for data mining. It's also known as the process of extracting interpretable, intriguing and valuable statistics from unstructured data. There are a variety of resources which generally produce huge amounts of raw data. This is the primary cause for the fast growth of data mining applications. This article discusses data mining methods and their applications, including scholastic data mining (SDM), life sciences, commerce, finance, and medicine among others. We put current methods together to see how data mining might be used to various areas. Our classification focuses on research that was published between 2007 and 2017. We provide a simple and brief perspective of various models used in data mining with this classification.

KEYWORDS- EDM, Educational Data Mining, KDD, Knowledge Discovery in Database, LMS, Learning Management System, SNA, Social Network Analysis.

I. INTRODUCTION

Raw data is transformed into usable information or expertise using data mining methods (DMT). Data is meaningless in and of itself, but processing it may be extremely helpful and fascinating. Many advanced technologies make clever use of data as valuable information. Knowledge Discovery in Databases (KDD), for example, is the procedure or method of extracting needed output in various forms from raw data [1] . KDD may alternatively be described as a method for identifying valuable patterns in data. Figure 1 is a typical and widely used data mining or KDD diagram.

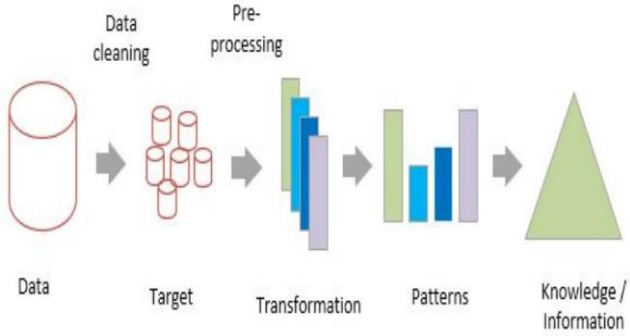

Figure 1: Process of discovering knowledge from raw data [2]

DMT is most often used in Educational Data Mining (EDM). It's utilized to create databases that help students and administrators make decisions more quickly. It is described as a new educational program that investigates various kinds of data generated by educational institutions. It examines data produced by the educational system so as to enhance educational and learning outcomes. It's part of a body of work on data mining, visualization, machine learning, and computing. Neural Networks, Nave Bayes, Decision Trees, K-Nearest Neighbor, and many more techniques are utilized in machine learning [3]. Data mining is also used in other areas. There are numerous suggested methods that combine data mining with semantic web, for example. Similarly, data mining in association with artificial intelligence (AI) and machine learning (ML) methods are utilized in a variety of applications.

Web or browser-based e-learning, which is popular nowadays, generates a large amount of data. This massive quantity of data or information is produced by online web servers, which may come from a variety of sources. User information relating to login history, location details, and other actions may all be found in this data. Traditional classrooms and remote education are the two primary sources of data generation in EDM. Scholars and educators are physically present in classes and elementary and higher either public or private education are all included. Attendance, information of educational courses, examinations, extra or co-curricular activities, and planning are all ways that educators monitor students' behavior. Educational data mining is beneficial to everyone connected with the institution [4-6]. Institutions, for example, need to know whether students are eligible to enroll in a specific course. The administration need data such as class size and entrance criteria. Students must learn the way to pick or choose educational courses depending on their predictions on which course would be the most beneficial. Instructors must know that from variety of available methods which teaching experiences are the most effective and beneficial to entire classroom. Learning systems produce data that is sent via the internet. On the data collected from the web, data mining methods (clustering, classification, pattern matching, text mining, and so on) are used. Data mining is utilized almost by all sector including accounting, administration, Human Resources (HR), and many others, in addition to

ducation. Table 1 lists the various EDM tools and their functions.

Table 1: Data Mining Tools [7]

Tool Name	Mining Task
Mining tool	Association and patterns
MultiStar	Association and classification
Data Analysis Center	Association and classification
EPRules	Association
KAON	Text mining and clustering
TADA-ED	Classification and association
O3R	Sequential patterns
Synergo/ColAt	Statistics and visualization
GISMO/CourseVis	Visualization
Listen tool	Visualization
TAFPA	Classification
iPDF-Analyzer	Text Mining

A. Categories of DMT

Because data received from many sources may be diverse and asynchronous, data mining methods are used to various elements of data mining. Data mining is a broad subject with applications in almost every area and department. As a result, a particular method or algorithm is used to tackle a given kind of issue effectively. Figure 2 depicts a taxonomy of data mining types. DMT is divided into nine categories, which are described below:

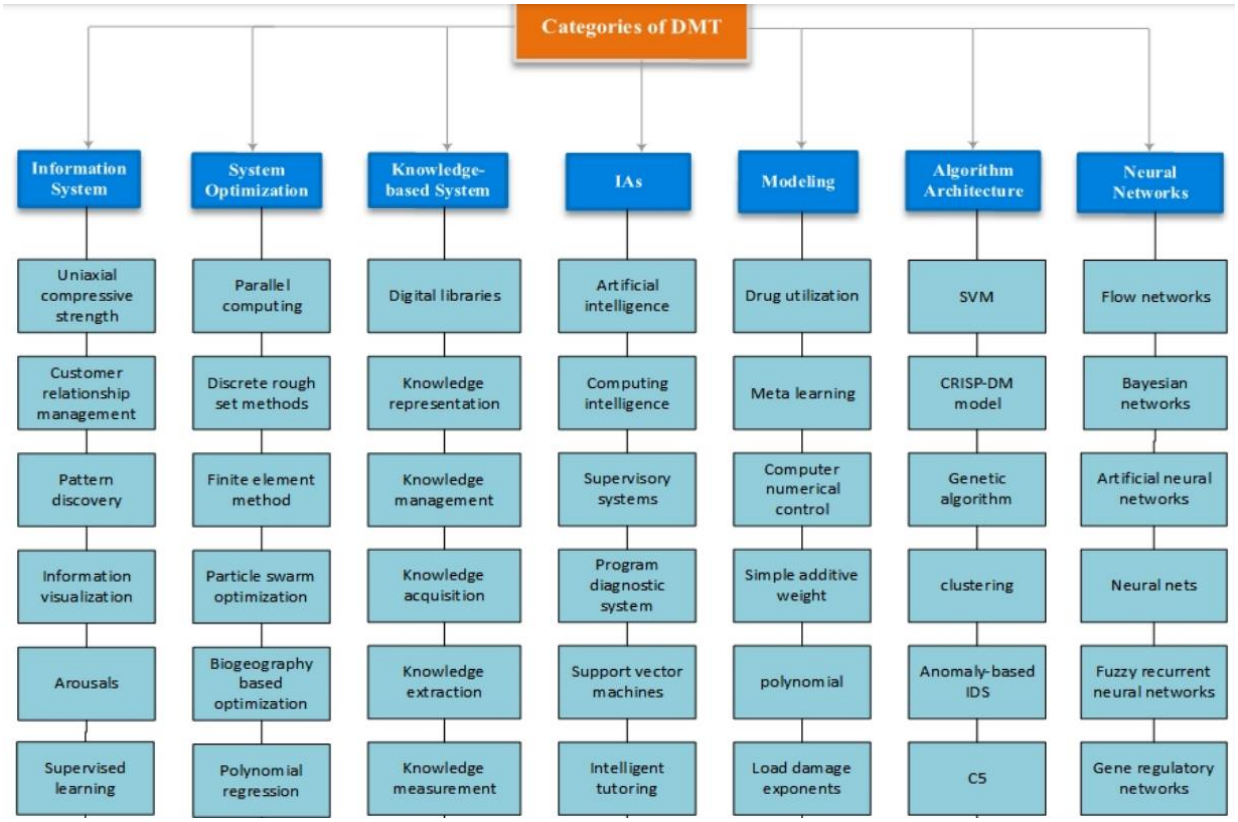

Figure 2: Taxonomy of data mining [8]

1) Information Systems

Information systems serve as link amongst the corporate sector and the area of computer science. Among all disciplines, information systems has become the most popular.

2) System Optimization

In the past, the phrase "linear programming" was used to describe improvement or enhancement of the systems. The finest or “grade A” element from collection of various accessible elements is chosen via system optimization.

3) Knowledge-based Systems

Artificial intelligence is built on knowledge-based systems. Many artificially intelligent technologies exist, and they use justification to make intelligent choices. Artificial intelligence serves as their foundation. Frames, scripts and other rules are utilized usually to express familiarity or understanding in these systems.

4) Modeling

It is a method of developing software that creates a data model using various data modeling approaches. A software engineering paradigm makes it simple to implement software. It is used to comprehend a system's complicated structure and movement from many angles. The data is quantitatively analyzed using modeling methods.

5) Architectural Analysis of System

The theoretical or abstract model that describes structures, perspectives, and behaviors of systems are used in system architecture analysis. The exemplification of system's structure along with the formal description is known as architecture. It reveals all of a system's components that work together to implement the whole system, as well as their relationships and behaviors that influence the overall system. The internal interaction of system components is the subject of architecture of the system.

6) Algorithm Architecture

A set of instructions or step by step procedure to calculate a function or resolve some problems is defined as an algorithm. Algorithms are used to process data and calculate results. The time and price involvedness and effectiveness of systems to solve an instantaneous issue are all influenced by the algorithm. Figure 3 depicts the steps required in developing an algorithm.

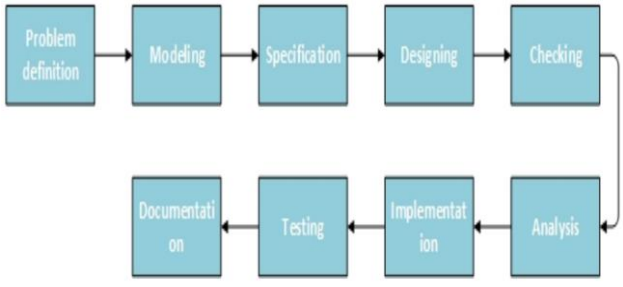

Figure 3: Algorithm development [9]

7) System for Intelligence Agent (SIA)

System for Intelligent Agent (SIA) is a kind of AI that has the ability to influence the environment. They may study and use knowledge to attempt to accomplish their objectives.

8) Approach Based on Dynamic Prediction

A statistical prototype used in modeling mollecules is the dynamic prediction-based method. It's also utilized to look for applications in the stock market.

9) Neural Networks

Biological neuron circuits are usually referred to as artificial neural networks (ANNs) or simply as neural networks. ANNs, which are made up of artificial nodes or neurons, are referred to as artificial neural networks in contemporary use. They are computational models that may be used in computer science and a variety of other areas of study. Neural networks are made up of neural units, often known as neurons. As a result, each neuron is connected to a large number of other neurons. These systems aren't pre-programmed in any way. They are, on the other hand, self-taught and trained. The aim of ANNs is to solve problems in the same way that a human brain does.

II. DISCUSSION

A. Data Mining Techniques

Because data mining is such a broad area with so many applications, it has become a fascinating study topic. Characterization, generalization, and association are the three types of data mining methods. Data mining methods may be utilized in a variety of ways, since they are difficult to use yet beneficial when done correctly. The following are some data mining techniques that are categorized and briefly discussed:

1) Clustering

Data may be kept as large data in either physical or digital form. Such data is stored in a variety of repositories. Big data is a data set that exceeds the processing capacity of software. Clustering is the process of forming groups of distinct items and their classes based on their various characteristics such as location, connectivity, and so on. Schools, for example, may be categorized based on their similarities and variances. Similarly, pupils may be grouped based on their actions. Clustering is used to look for data points that are naturally clustered together.

2) Prediction

Predictions are often based on prior information and experience. Predictor variable refers to the emphasis on a particular element of data in relation to another feature of data. Prediction is a technique for predicting an unknown outcome based on prior knowledge.

3) Relationship Mining

For relational databases, relational mining (RM), also known as relation data mining (RDM), is widely employed. A connection between various variables within

a data collection is found via relationship mining. The RDM algorithm searches for a specific pattern among various patterns available in a database. The relationship between variables must fulfill two criteria: it must be intriguing and it must be significant.

4) Outlier Detections

In general, an outlier is a fresh observation that differs from the revious one being compared. Outlier detection compares various values in a dataset to the lowest & biggest values and detects the deviation between them.

5) Text Mining

Mining of text data in data mining is a method that is unique only to the text data. Documents, emails, messages, and html files are examples of text data. Text summarization, document dispensation, topic grouping, indexing, and mapping are all examples of text mining. It is widely utilized in the fields of learning and professional occupation. Institutions generally have a large number of text data and employ text mining to extract the statistics they need. Natural language processing (NLP), ML and statistics are all used in text mining. Information technology (IT), publishing, telecommunications, banking, pharmaceutical businesses and public administration are just a few of the text mining uses.

6) Social Network Analysis (SNA)

The technique of social network analysis employs networks and graph theory to study social systems. The connection between various entities in network information is identified in this procedure. It's frequently used to examine a group's or community's activities.

7) Process Mining

On the basis of event logs, process mining examines business operations. It extracts information relevant to the event log procedure. The information system takes note of this event record for easy display.

8) Data Distillation for Judgment

Data is sensibly expressed in this manner. Visualization and summarization are used in this approach. This is helpful for seeing and analyzing huge amounts of data at once.

B. Applications of Data Mining Methods

Data mining techniques have a wide range of uses. Some of them are described below, while others are included in Table 2:

Table 2: Illustrates different Data Mining techniques with its limitations and applications [7]

Data mining technique / category	Description	Implementation / Tools	Limitation	Applications
Statistics	Graphical and tabular representation of data	WebStat, AccessWatch, Analog	Fail to analyze individual item, Not implemented on heterogeneous data, A small error brings to misleading	Measuring the number of visits on web, Large data is described as charts, graphs, 3D representation
Web data mining	Data mining related to web	Winautomation, import.io, CrawalMonster etc.	Invasion of privacy, Irrelevant contents	Determining the web structure, web contents, web usage
Classification	Grouping of data objects	Generation of groups with same attributes and characteristics	Useless for heterogeneous data	Reduction of information complexity, streamlining in data collection, helpful in planning
Clustering	Grouping of data with samilarities	Cluster 3.0, Java TreeView, PYCLUSTER etc.	Don't support shared storage, operational errors	Fault tolerance, maintenance
Sequential pattern	Ordering of objects with a particular sequence	XAffinity(TM), SPMF, Miningco	Big storage for database,	Shelf in a shop, disaster prediction, proceeding medication detection
Association rule	Antecedent and consequent or if then statements	FPM, Bart Goethals, FrIDA, KNIME, Magnum Opus	Research effort goes to improve the algorithm used, in e-learning, algorithm used has too many parameters	Used in LMS, stock trading
Prediction	Estimation based on previous data	EDM, business	Past result fail due to change in future trend	Weather prediction, student behavior, business
Correlation mining	Creation of patterns from signals, audios, videos, images, sequences	Google Trends, Google Flu Trends, Google Correlate	Doesn't provide the reason of relation among objects	Very helpful for researchers to collect more data than experiments, Used in neuroscience, material science and finance
Casual data mining	Prediction in data relationship	Weka, RapidMiner, KNIME, Rattle	Quality, security, privacy of data	Used in healthcare, business, finance, banking, education
Outlier detection	Detection of deviation of one observation from many other observations	CMSR Data Miner	Need mathematical justification, need probabilistic data model that is complex	Image processing, detection of industrial damage, fraud detection, intrusion detection, inside trading detection, public health and medical
Text mining	Driving information from text	Carrot2, GATE, Gensim, OpenNLP, Orange, Stanbol, KNIME, PLOS, PubGene	A lot of free text in data collection, data is unstructured, syntactic and semantic erros in data, resource development is difficult	Record management, intelligence, social media, searching, publishing, life sciences, security (encryption, decryption), customer relationship management, education, digital humanities
Social network analysis	Use of network to investigate social structures	Commetrix, Cytoscape, Cuttlefish, EgoNet, Gephi, Graph-tool, GraphChi, Graphviz etc.	Risk of fraud, Time wastage, Invasion of privacy	Worldwide connectivity, information sharing, targeted advertising
Decision trees	A tree like model of decisions and their consequences	SilverDicisions, Gambit, Simple Decision Tree, GATree, KNIME, RapidMiner, Smiles, YaDT,	Complexity, loss of innovation, a small change in data set brings a great change in decision trees, difficult to move because of its size and shape etc.	Modeling techniques, feature selection, data preparation, interpretation of data
Nearest neighbor technique	A method used for classification and regression	Face recognition, recommendation engines, spam filtering, Weka, Kaldi, MEKA, mlpy, MODLEM, sgmweka,	Finding the value of k, determining the parameters to be used, high computation cost	Adaptive websites, bioinformatics, cheminformatics, game playing, computer vision, marketing, medical, economics, search engines, stock market analysis, information retrieval, speech recognition
Process mining	Process management on the basis of event logs	Prom, XESame, OpenXS, ProMimport, MXMLib etc.	Timing problem, conFiguration issues, noise, incompleteness, complexity	Process discovery, conformance checking, compliance checking,

1) Statistics

The user of apps is the primary focus of data mining. Access Watch, Web Stat, and Analog are some of the tools that utilize form use statistics [10,11]. Measuring the number of visitors is an example of use statistics. If data is stored in a relational database, SQL offers a variety of functions, including sample size and mode. All of the methods transform huge amounts of data into a particular visual representation. Charts, graphs, and 3D representations are often used to explain big data. These data visualizations may include assignments, examinations, courses, and grades. Mentors may acquire data relating to their pupils and online courses.

2) Web Data Mining

Web data mining is another DM application. Information is filtered from data received from the internet in this case. Web structure, web content, and web use are all examples of web data. The primary goal of online data mining is to provide consumers with the information they need.

C. Limitations and Open Issues

As previously said, data mining is an essential element that can be used in a variety of fields. It is divided into several categories and classifications. We investigated data mining, including its methods, classifications, and applications. It is a wide field that each person utilizes. Data mining is used by learners or students to customize e-learning. Data mining suggests that better learning experiences are possible. Data mining is used by educators or instructors to get feedback on their teachings. They look at how students behave and learn. They can also anticipate student performance in order to enhance course personalization. Data mining is used by researchers to determine the best data mining method and create data mining tools for particular purposes. Organizations and businesses utilize data mining to improve the efficiency of their decision-making processes. Data mining is used by administration to determine the best method to manage resources and to make the most effective use of those resources.

III. CONCLUSION

Data, being the most important component of any area, must be handled effectively. In this case, data mining is very beneficial. The most pressing problem now is data privacy and security. Privacy becomes increasingly essential in the context of worldwide data exchange, particularly for the online. Integration of qualitative, quantitative, and scientific techniques, as well as research of DMT methodologies, will help people comprehend the topic better. Finally, DMT methods' capacity to adapt and offer new knowledge is their primary benefit, and it will continue to be at the heart of DMT applications in the future. As a result, our future work will incorporate data privacy and security, which will be accomplished by using a particular security algorithm that will not degrade data efficiency.

REFERENCES

[1] Dhiman AK. Knowledge Discovery in Databases and Libraries. DESIDOC J Libr Inf Technol. 2011;

[2] Guarascio M, Manco G, Ritacco E. Knowledge discovery in databases. In: Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics. 2018.

[3] Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques. Int J Sci Res. 2016;

[4] Peña-Ayala A. Educational data mining: A survey and a data mining-based analysis of recent works. Expert Systems with Applications. 2014.

[5] Bakhshinategh B, Zaiane OR, ElAtia S, Ipperciel D. Educational data mining applications and tasks: A survey of the last 10 years. Educ Inf Technol. 2018;

[6] Dinesh Kumar A, Pandi Selvam R, Sathesh Kumar K. Review on prediction algorithms in educational data mining. Int J Pure Appl Math. 2018;

[7] Rangra K, Bansal KL. Comparative Study of Data Mining Tools. Int J Adv Res Comput Sci Softw Eng. 2014;

[8] Colonna L. A Taxonomy and Classification of Data Mining. Sci Technol Law Rev. 2013;

[9] Drumond M, Daglis A, Mirzadeh N, Ustiugov D, Picorel J, Falsafi B, et al. Algorithm/Architecture Co-Design for Near-Memory Processing. ACM SIGOPS Oper Syst Rev. 2018;

[10] Yordanova A, Yordanov M. WEB BASED SYSTEM FOR CHOOSING A STATISTICAL METHOD FOR DATA PROCESSING. Appl Res Tech Technol Educ. 2018;

[11] Kitchen AM, Drachenberg R, Symanzik J. Assessing the reliability of web-based statistical software. Comput Stat. 2003;