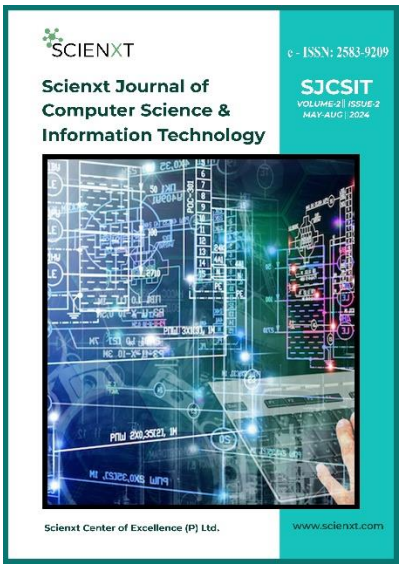

Scienxt Journal of Computer Science & Information Technology
Volume-2 || Issue-2 || May-Aug || Year-2024 || pp. 1-14

A Review on data mining techniques and applications

[*1]Priyanka Singh, [2]Avantika Singh Rajput, [3]Riteen Shaw

[*1]Assistant Professor, Department of Computer Science & Engineering, Bhopal Institute of Technology and Science, Bhojpur Road Bhopal, 462045 M.P. India

[2,3]Student, Department of Computer Science & Engineering, Bhopal Institute of Technology and Science, Bhojpur Road Bhopal, 462045 M.P. India

**Corresponding Author: Priyanka Singh*
Email: priyankasinghcse02@gmail.com

Abstract:

Data mining, also referred to as Knowledge Discovery in Databases (KDD), is a methodological process aimed at uncovering concealed patterns and insights from extensive databases and data repositories. It facilitates data exploration, analysis, and visualization on a large scale, devoid of predefined hypotheses. Central to its operation is the utilization of modeling techniques for predictive purposes. A plethora of algorithms and tools exist to support these endeavors. The scope of data mining extends across diverse domains, spanning from business to medicine to engineering. This paper offers an overview of data mining methodologies, models, tasks, applications, key challenges, and future research directions, focusing on select domains where data mining technologies find wide-ranging applications.

Keywords:

Data mining, Knowledge discovery in databases, Data mining applications.

1. Introduction:

With the rise of computers and their capacity for extensive digital storage, we began accumulating and storing vast amounts of data, relying on computers' capabilities to sift through this wealth of information. However, this accumulation quickly became overwhelming, prompting the development of structured databases and database management systems (DBMS). These systems efficiently handle large volumes of data, enabling the swift retrieval of specific information when needed. Additionally, they facilitate the ongoing accumulation of various types of information. This ability to retrieve information as and when required laid the groundwork for data mining technology.

Data mining is considered a natural progression of information technology, arising from the increasing availability of vast datasets and the growing demand to derive useful information and knowledge from them. It involves extracting insightful patterns or knowledge from massive amounts of data. Known by various names such as Knowledge Discovery in Databases (KDD), knowledge extraction, and business intelligence, data mining entails analyzing data in a database using tools that identify trends or anomalies without prior knowledge of the data's significance. It is primarily utilized by statisticians, database researchers, and business communities.

Data mining software not only presents data differently but also uncovers previously unknown relationships among the data. It operates on information stored in historical databases of past interactions. In theory, data mining is not limited to any specific type of media or data and should be applicable to any information repository.

2. Literature survey:

Fayyad et al. (1996) characterized Knowledge Discovery in Databases (KDD) as a complex process aimed at identifying valid, novel, potentially useful, and ultimately comprehensible patterns within data [1]. In this context, data represents a collection of facts accessible in electronic format. The term "patterns" refers to models and regularities discernible within the data, which must be valid, meaning they should hold true for new data to a certain extent.

Data mining, as a component of the KDD process, involves the application of data analysis and discovery algorithms to generate a specific set of patterns from the data, subject to computational efficiency constraints [2]. According to this definition, data mining focuses on extracting actionable knowledge from data. It underscores the importance of data mining

algorithms being capable of processing large datasets efficiently, ensuring that desired patterns can be identified within acceptable computational limits.

Rygielski et al. (2002) highlight the transformative impact of technologies such as data warehousing, data mining, and operations management software on relationship marketing [3]. These technologies have enabled firms to delve into customer relationship management, offering opportunities for gaining competitive advantages. Particularly through data mining, organizations can extract predictive insights from extensive databases, enabling the identification of valuable customers, anticipation of future behaviors, and empowering firms to make proactive, knowledge-driven decisions.

Liu et al. (2010) present the technology of process knowledge discovery in the process database. After scrutinizing the process planning [4], they discuss the knowledge discovery flow and its key technologies, highlighting numerous advantages. Moreover, they emphasize its potential to expedite the standardization of process planning. Finally, they introduce the PPK discovery system and detail its structure and functionality.

Diamantini et al. (2011) introduce Designer, a web-based semantic-driven tool [5] designed to assist users in the collaborative design of a KDD process. This tool, tailored for supporting non-expert users in the collaborative design of KDD processes, leverages an SOA-based methodology to execute KDD tools as web services. This approach resolves interface heterogeneity issues and enables seamless communication protocols.

In summary, data mining is a methodology for uncovering previously undiscovered, valid patterns and relationships within vast datasets, whether they are qualitative, textual, or multimedia. This is achieved through the application of various data analysis tools, often utilizing datasets collected for different purposes.

Anshu (2019) highlights how data mining enhances the performance [16] of regular databases, resulting in faster operations and increased profitability due to informed decision-making. The paper elaborates on the steps involved in the data mining process and its utility across diverse industries for extracting valuable insights from large datasets.

Koti Neha et al. (2020) emphasize the critical role of data mining in managing and extracting essential information from vast datasets across various domains. Their paper [17] provides an overview of how data mining is applied in different fields.

Yang Yang (2023) proposes an examination score analysis and application [18] system based on data mining algorithms to delve into the intricate relationship between examination scores

and various factors. Through the development of a performance analysis algorithm model, the paper aims to predict and analyze students' performance more accurately.

3. Architecture and process of data mining:

3.1. Architecture of data mining:

Data mining involves discovering valuable knowledge within extensive datasets stored in data warehouses, databases, or other information repositories. A typical system architecture comprises the following major components, as depicted in Fig. 1.

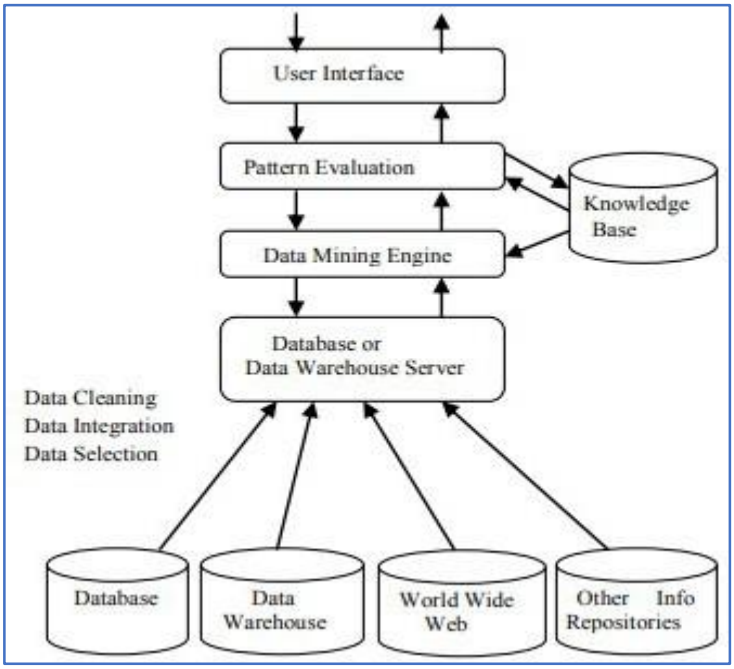

Figure. 1: Architecture of typical data mining system

3.1.1. Data repository, whether it's a data warehouse, database, World Wide Web, or any other information storage system:

This encompasses either a single or a collection of data warehouses, databases, spreadsheets, or similar repositories. Data cleaning and integration techniques may be employed to ensure data quality and coherence.

3.1.2. Database or data warehouse server:

This component is tasked with retrieving the pertinent data in response to the user's data mining query.

3.1.3. Knowledge repository:

This encompasses domain-specific knowledge utilized to guide the search or assess the significance of resulting patterns. Such knowledge may include concept hierarchies and user perspectives.

3.1.4. Data mining engine:

This core component of the data mining system comprises functional modules responsible for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier detection, and evolution analysis. It is of paramount importance to the data mining process.

3.1.5. Pattern assessment module:

This component typically incorporates measures of pattern interestingness and interacts with the data mining module to direct the search towards meaningful patterns. The pattern evaluation method can be seamlessly integrated with the data mining component depending on the chosen implementation technique.

3.1.6. User interface:

This module serves as the intermediary between the user and the data mining system, enabling users to interact with the system by specifying data mining queries or tasks, providing information to refine the search, and conducting exploratory data mining based on interim results.

3.2. Data mining process:

While some individuals consider data mining synonymous with the commonly used term "Knowledge Discovery from Data," others view data mining as a fundamental step within the broader process of knowledge discovery, as illustrated in Fig. 2.

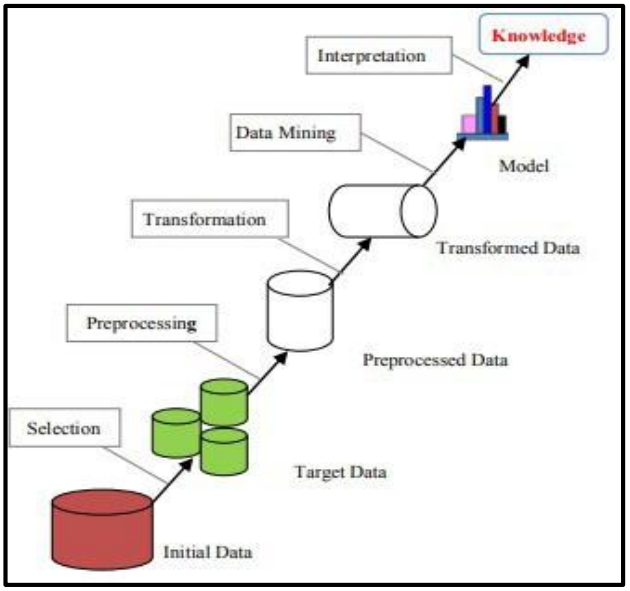

Figure. 2: Data mining process model

3.2.1. Data selection:

This involves selecting the necessary data for the data mining process, which may originate from various sources.

3.2.2. Data preprocessing:

This phase addresses erroneous or missing data. Various tasks may be undertaken during this stage, including correcting or eliminating flawed data and imputing missing values. Additionally, preprocessing encompasses noise and outlier removal, gathering pertinent information to model or account for noise, handling time sequence information, and addressing known changes.

3.2.3. Data transformation:

This step involves converting the data into a standardized format for further Processing. Some data may undergo encoding or transformation into a more suitable format. Technique ssuch as data reduction, dimensionality reduction (e.g., feature selection such as attribute subset selection or heuristic methods), and data transformation methods (e.g., sampling, aggregation, generalization) may be employed to reduce the number of potential data values being analyzed.

3.2.4. Data mining:

This crucial phase applies sophisticated techniques to extract patterns from the data.

3.2.5. Interpretation/evaluation:

This step entails elucidating how the data mining results are interpreted by users, which is of

paramount importance as the utility of the results depends on it. Various visualization and graphical user interface (GUI) strategies are employed in this stage. Different types of knowledge necessitate different forms of representation, such as clustering, classification, or association rules.

4. Data mining tasks:

Data mining tasks serve to categorize the types of patterns to be uncovered during the data mining process. Generally, these tasks fall into two categories: Predictive and Descriptive. A Predictive model aims to make predictions about data values using patterns discovered from different datasets, with the objective of uncovering strong relationships between variables within a dataset. Conversely, a Descriptive model identifies patterns or relationships within data, summarizing data in ways that enhance understanding of underlying processes. The key distinction between the two models lies in their objectives: while a Descriptive model seeks to uncover properties of the examined data, a Predictive model focuses on predicting new properties based on existing data.

Predictive model data mining tasks include classification, prediction, regression, and time series analysis. Descriptive tasks encompass methods such as Clustering, Summarization, Association Rule Discovery, and Sequence analysis.

4.1. Classification:

Classification involves identifying common properties among objects in a database and categorizing them into different classes based on a classification model. The objective is to develop accurate descriptions or models for each class using training data, which are then used to classify future test data. Common classification strategies include support vector machines, decision trees, and logistic regression.

4.2. Prediction:

Prediction involves forecasting occupied data values or trends, or predicting class labels for data. Once a classification model is established using a training set, predictions can be made based on feature values of the object and class characteristics. Prediction also includes forecasting missing numerical values or trends in time- related data.

4.3. Regression:

Regression techniques are adapted for prediction, with the predicted variable being continuous.

Regression entails learning functions that map data items to true-valued prediction variables. Common regression strategies include statistical regression, neural networks, and support vector machine regression.

4.4. Time series analysis:

Time series analysis examines attribute values as they vary over time. Various statistical techniques, such as auto-regression methods, are used to analyze time-series data, including ARIMA and long- memory time-series modeling.

4.5. Clustering:

Clustering involves grouping physical or abstract objects into classes of similar objects, also known as unsupervised classification. It is a major class of data mining and a standard technique for statistical data analysis used in various fields, such as pattern recognition and bioinformatics.

4.6. Summarization:

Summarization, also known as Description or Generalization, organizes data into subsets with corresponding descriptions. It retrieves actual parts of mined data and provides summaries based on these subsets. Summarization is an outcome of data mining rather than a mining technique itself.

4.7. Association rule mining:

Association rule mining discovers relationships among attributes within datasets, generating if-then statements about attribute values. It aims to extract interesting correlations, frequent patterns, or associations within transaction databases, commonly used for market basket analysis.

4.8. Sequence discovery:

Sequence discovery identifies sequential patterns within data, typically associations between variable data fields based on time. This method includes association rules and Markov concepts, often used to uncover patterns such as the sequence of purchases, like buying CDs after purchasing a music player.

5. Data mining major challenges:

Despite the considerable advancement in data mining and knowledge discovery technology, its practical application faces several challenges [7], outlined below.

5.1. Security and social concerns:

Security poses a significant issue when sharing data intended for strategic decision-making. Confidential data may be at risk of unauthorized access, potentially violating privacy policies. Data mining may unveil implicit information about individuals or groups, raising privacy concerns. Furthermore, the competitive advantage gained from discovered knowledge may lead to selective data disclosure and uncontrolled data usage.

5.2. User interface challenges:

The value of insights provided by data mining tools depends on their usability and comprehensibility to users. Challenges in user interfaces and visualization include limited screen space, effective data rendering, and interaction capabilities. Interactive exploration of data and mining results is essential for users to refine tasks, visualize information from various perspectives, and comprehend findings at different levels of abstraction.

5.3. Mining methodology challenges:

These challenges relate to the approaches and limitations of data mining techniques. The scale of data and the size of the search space significantly impact mining methodologies. The exponential growth of the search space with an increase in dimensions, known as the curse of dimensionality, severely affects the performance of some data mining techniques, necessitating urgent resolution.

5.4. Performance concerns:

Many AI and statistical methods used for data analysis and interpretation are not optimized for processing very large datasets encountered in data mining. Scalability and efficiency issues arise when dealing with massive datasets, including challenges in incremental updating and parallel programming.

5.5. Data source complexity:

Various issues are associated with data sources, ranging from practical concerns such as the diversity of data types to philosophical challenges like data glut. Heterogeneous data sources, characterized by differences in structure and semantics, pose significant challenges to both the database and data mining communities.

6. Applications of data mining:

Data mining finds applications across various domains, benefiting different sectors:

6.1. Healthcare:

Data mining plays a vital role in the healthcare industry, assisting health insurers in detecting fraud, aiding healthcare organizations in customer relationship management, assisting physicians in identifying effective treatments, and ensuring patients receive optimized healthcare services. The vast amounts of data generated by healthcare transactions are too complex for traditional methods to process efficiently, making data mining essential for transforming this data into actionable insights.

6.2. Educational data mining:

Educational data mining focuses on exploring and analyzing large datasets from educational contexts. Leveraging techniques from the data mining community, it addresses issues related to learning, cognition, and assessment. Recent advancements in computing power and data mining algorithms have made educational data mining increasingly successful in addressing these academic challenges.

6.3. E-commerce:

Data mining enhances electronic commerce (EC) by leveraging user-specific product information to personalize the shopping experience and construct effective EC strategies for businesses. Resolving issues related to complex activities in electronic commerce, such as negotiations between consumers and sellers, is crucial for facilitating transactions and satisfying both parties.

6.4. Sports data mining:

The sports industry collects vast amounts of statistics from players, teams, games, and seasons. Data mining techniques help derive meaningful insights from this wealth of data, aiding in performance analysis, player evaluation, and strategic decision-making.

6.5. Market basket analysis:

Market basket analysis utilizes data mining techniques to identify associations between different items purchased by customers. By discovering such associations, businesses can optimize their marketing strategies and increase profits by understanding customer buying patterns.

6.6. Customer relationship management (CRM):

Data mining is integral to CRM, offering insights into customer behavior, preferences, and trends. By analyzing customer data, businesses can enhance customer retention strategies, improve marketing campaigns, and tailor products and services to meet customer needs more effectively ongoing research in this area suggests a growing interest and potential for significant advancements in CRM applications leveraging data mining techniques.

7. Conclusion:

Data mining holds significant promise for organizations seeking to uncover hidden patterns in their data to predict customer behavior, product trends, and process outcomes. However, effective utilization of data mining tools requires guidance from users with a deep understanding of the business, the data, and the analytical techniques involved. Realistic expectations are key to achieving favorable outcomes across various applications, including revenue generation and cost reduction.

Addressing practical challenges related to data sources, such as heterogeneous databases and diverse data types, remains a critical concern. Given the varied nature of data stored across different repositories, it is challenging to develop a data mining system that can efficiently deliver high-quality results across all data types and sources. Tailoring algorithms and methodologies to suit different data types and sources is essential for effective data mining.

Furthermore, there is a growing recognition of the necessity for data mining, as evidenced by the increasing attention to its motivation and applications. This paper has provided an overview of the typical architecture and process of data mining, including the classification of data mining systems. Key challenges requiring attention have been identified, and several applications showcasing the use of data mining technology have been discussed.

In conclusion, effectively navigating the vast universe of digital data will be crucial for organizations' strategic success. The ability to manage and mine data effectively will play a pivotal role in leveraging insights to drive informed decision-making and achieve competitive advantage in today's data-driven world.

8. References:

(1) U.M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From Data Mining to Knowledge

Discovery in Databases," *AI Magazine Volume 17 Number 3 (1996) AAAI.*

(2) U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "The KDD Process for Extracting Useful Knowledge from Volumes of Data," *Communication of the ACM November (1996) /Vol. 39, No. 11.*

(3) C. Rygielski, J.C. Wang, D. C. Yen, "Data mining techniques for customer relationship management," *Technology in Society 24 (2002) 483–502.*

(4) S. Liu, X. Tian, Z. Zhang, "Process Planning Knowledge Discovery in the Process Database," *IEEE International Conference on Computer Application and System Modeling (ICCASM 2010).*

(5) C. Diamantini, D. Potena and E. Storti, "A Semantic-Aided Designer for Knowledge Discovery," *IEEE 2011.*

(6) M.S. Chen, J. Han, and P.S. Yu, "Data Mining: An Overview from Database Perspective," *IEEE transaction on knowledge and data engineering Vol. 8 no. 6 December 1996.*

(7) B.N. Lakshmi, G.H. Raghunandhan, "A Conceptual Overview of Data Mining," *IEEE Proceedings of the National Conference on Innovations in Emerging Technology Tamilnadu, India.17 & 18 February, 2011.pp. 27-32.*

(8) S.R Barahate, V.M. Shelake, "A Survey and Future Vision of Data mining in Educational Field," *IEEE Second International Conference on Advanced Computing & Communication Technologies 2012.*

(9) Z. Zheng, R. Kohavi, L. Mason, "Real World Performance of Association Rule Algorithms," *Proceedings of the Seventh ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY: ACM, 2001.*

(10) R. Agarwal, T. Imielinski, & A. Swami, "Mining association rules between sets of items in large databases," *In Proceedings of the ACM SIGMOD international conference on management of data, Washington DC, USA, 1993, (pp. 1– 22).*

(11) H. C. Koh and G. Tan, "Data Mining Applications in Health Care," *Journal of Healthcare Information Management Vol. 19, No. 2.*

(12) P.H. Chou, P.H. Li, K.K. Chen, M.J. Wu, " Integrating web mining and neural network for personalized e-commerce automatic service," *Expert Systems with Applications 37 (2010) 2898–2910*

(13) O.K. Solieman, "Data Mining in Sports: A Research Overview," *A Technical Report,*

MIS Master Project, August 2006.

(14) T. Raeder, N. V. Chawla, "Market Basket Analysis with Networks," *2011 Springer.*

(15) E.W.T. Ngai, L. Xiu, D.C.K. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification," *Expert Systems with Applications 36 (2009) 2592–2602.*

(16) Anshu, "Review Paper on Data Mining Techniques and Applications". *International Journal of Innovative Research in Computer Science & Technology (IJIRCST), Volume-7, Issue-2, March 2019*

(17) Koti Neha, M Yogi Reddy "A Study On Applications Of Data Mining," *International Journal Of Scientific & Technology Research Volume 9, Issue 02, February 2020.*

(18) Yang Yang "Application of Data Mining Technology in Exam Score Analysis," *International Conference on Machine Learning and Big Data Analytics for IoT Security and Privacy Procedia Computer Science 228 (2023) 98–111.*