International Journal of Communication and Information Technology

E-ISSN: 2707-6628
P-ISSN: 2707-661X
www.computersciencejournals.com/ijcit
IJCIT 2022; 3(1): 37-40
Received: 05-01-2022
Accepted: 11-02-2022

Mohammad Atif
Undergraduate Student, Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India

Corresponding Author:
Mohammad Atif
Undergraduate Student, Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India

Data mining

Mohammad Atif

DOI: https://doi.org/10.33545/2707661X.2022.v3.i1a.44

Abstract
The advancement of technology has resulted in the birth of a number of new fields. Every day, several areas such as research, engineering, health, and business generate and accumulate large volumes of data. Data mining is an important field that organises & extracts necessary informations from massive amounts of data. This paper provides an overview as to how data mining has been used in many fields.

Keywords: Applications, data mining, data set, KDD, techniques, tools

1. Introduction

"Data mining is the processes of obtaining informations from a vast amount of data. Data mining is an important stage in Knowledge Discovery in Databases (KDD). The KDD processes involves extracting required data from big databases or data stores and converting it into various patterns, summary reports, and views, among other things. The stages of KDD are as follows:

1. **Data Extraction:** From massive databases or repository, we select sets of data or data samples.
2. **Data Cleaning:** Data sets from huge databases might not have been accurate. Employing data transformation methods, we needed to clean up the data by eliminating inconsistencies and missing values.
3. **Data Integration:** Information from various sources is merged and contained in a single location. In this step, we employ data migration and synchronisation tools.
4. **Data Selection:** Utilizing decisions trees, Naive Bayes, Clustering, Neural networks, and Regressions approaches, data that is valuable for the study is extracted from the data resource.
5. **Data Transformation:** Using summarization and aggregation processes, data is turned into an appropriate way for mining procedures.
6. **Data Mining:** Several strategies are used to turn data into useful patterns.
7. **Pattern Evaluation:** Interestingness metrics are used to identify useful patterns. In this stage, we apply visualisation and summarization approaches.
8. **Knowledge Presentation:** Data mining findings are visualised and presented in reports, tables, and other formats.

The most important aspect of data mining is selecting a dataset from such a large source [3, 7]. Record data, Graph-based records, and Sequential data are the three categories of dataset classifications. In the method of record collecting, data is being collected. Each record has data fields that are connected to one another. Take, for instance, market-basket data. Data having relationships between things, such as connected web pages, is referred to as graph-based statistics. The term sequential data refers to an ordered series of events. We have three types of sequential data based on the characteristics of events: time-series info, symbols sequence data, and biologically sequence data. To retrieve essential data, several theories and techniques of data mining are utilised depending on the applications and dataset.

2. Data mining techniques

Appropriate strategies are used depending on types of data mining assignment. Descriptive data mining jobs classify data in a targeted data-set given historical or recently incidents. Predicting tasks predict outcomes of future inquiries using historical data. A few of the most frequently utilized data mining techniques are classifications, clustering, regressions, outlier identification, association rules, pattern matching, and prediction.

Predicting data mining techniques include classifications, regressions, and outer detections. Description data mining approaches include clusters, sequential pattern identification, and association rules.

2.1 Classification

We develop a model in Classifications that recognises and allocates a class to new observative input data presented to the model. We split the input into 2 sets: the training set (which is used to develop the model) as well as the test set (which is used to model validation). The data out from Training set is separated into multiple classes. The concept we create assigns a class to the data from the test set. In this technique, we employ classifiers such as Decision Trees, Bayesian Classifiers, Neural Networks, K-Nearest Neighbor, Support Vector Machines, Linear Regressions, and Logistic Regression. For example, we assess whether a mobile phone is good or poor before purchasing it based on factors such as battery life, performance, and pricing. Direct marketing, sky surveys cataloguing, fraud detection, and other applications are among them.

2.2 Clustering

Clustering divides data into categories or clusters, having objects in the same cluster having similar features and those in separate clusters having less similar features. Depending on the application, several clustering approaches are utilised. Partitioning Method, Grid-Based Method, Density-based Method, Model-Based Method, Hierarchical Method, and Constraint-based Method are a few of them. Documents clustering, market segmentations, biology, medical imaging, and social network analysis are just a few of the uses.

2.3 Regression

The predictor variables (target) of the regression approach predicts the amount of a continuous-valued variables termed the responsive variable (which values are already known). Simplly Linear Regression, Lasso Regressions, Logistic Regressions, Supportive Vector Machines, Multivariate Regressions Algorithm, and Multiple Regressions Algorithm are among the regression methods used during data mining. Traffic accidents and reckless driving, for example, may be foreseen. Forecasting sales, financially planning, trend analyses, advertising, time series predictions, and determining fossils age are just a few of the uses available.

2.4 Association Rule Mining

The associations rules mining approach identifies patterns in data, as well as linkages and associations between huge data sets. An item's recurrence may be anticipated based on the occurrences of another objects. Associations rules are that if rules that are used to compute lift, support, and conviction in order to find common patterns and relationships between items. The Apriori algorithm, FP-growth algorithms, Eclat algorithms, Market-basket analyses, cross-marketing, and catalogue creating are examples of Associations rule algorithms.

2.5 Outlier Detection

Outlier identification identifies and removes outliers (sampling data that acts in a radically different way from the rest of the data set) from the data collection. Z-Score, DBSCAN, Isolation Forests, Linear Regressions Models (LMS, PCA), Proximity Dependent Models (non-parametric), and High Dimensional Outlier Identification Approaches are among the outlier detection methods. Fraud detections, intrusion detections, medically and healthy outlier identification, insurance settlement fraud detection, and so forth are some of the uses.

2.6 Sequential Patterns

To forecast sequential interconnections and sub processes, the sequential patterns approach is utilised. GSP (Generalized Sequential Patterns), Loose span, Prefix period, and SPADE are some of the methods used to detect sequential patterns (Sequential PAttern Discovery utilizng Equivalent Classes). DNA sequences, weblog visit streams, telephone calling habits, equities and marketplaces are some of uses.

3. Applications of data mining

Data mining techniques are used in a wide range of fields for technical, commercially, and scientific goals. Table 1 shows a summary of the most extensively used data mining approaches and tools in diverse applications:

3.1 Bioinformatics

Bioinformatics [15] is a set of computer-assisted technologies for managing, storing, and analysing biological data. The amount of data in this subject was growing by the day and was being utilised extensively for study. Gene sequences analyses, gene and protein communicative network constructions, diseases detections, DNA-sequencing and alignment, and other data mining apps in such field involve gene sequences findings, protein sequence analyses, gene and protein communicative network construction, diseases detections, DNA sequencing and alignment, and so on. In bioinformatics, a sequence data collection is employed. This sequencing dataset is delivered to the appropriate data mining tools depending on the application kind to acquire the desired results. BLAST (Basic Local Alignment Search Tools), FASTA, CS-BLAST for identifying sequence alignments, GenScan, GeneMark for gene discovery, Pfam, BLOCKS, ProDom for protein analyses, and so on are a few of the data mining tools being used bioinformatics.

3.2 Financial Banking

Every day, digitalized bankings creates massive volumes of transactional datas [9]. In this field, data mining is used for applications such as financial banking, where it is used to determine client loyalties, issues loans & credit cards based on past data, and predict stocks market dangers from historical information. Classification methods such as Bayes classifications, Boosting, Decisions tree, and Random forest are used in these applications. Rapid Miners, R programming, Weka (Waikato Environmental for Knowledge Analysis), Orange, KNIME, NLTK (Natural Languages Tool Kit), and other data mining technologies are used in business and finance.

3.3 Education

The discipline of educational data mining is a recent one, focusing on creating techniques for discovering necessary data from diverse educational domains [12, 16]. Predicting student outcomes, learning behaviours, and identifying weak pupils are just a few of the data mining uses in this subject. Students' learning patterns are utilised to build instructional strategies. In the Education applications, the Record data set is utilised. SPSS, KEEL, Weka, Spark MLLib, and other data mining technologies are utilised in education.

3.4 Criminal Investigation

Criminal analysis entails identifying crimes as well as the criminals' connections to them. We get large amounts of criminal datasets from various crimes such as cybercrime, violent crimes, fraud detections, and drug offences [7, 8]. In this discipline, data mining is used for applications like as counter-terrorism, criminality matching, and crime trends, among others. Weka, H2o, Orange, and other data mining technologies are employed in this industry.

3.5 Market Basket Analysis

In the retail business, Market Basket Analyses is used to forecast client behaviour. It is founded on the notion that if a certain set of things is bought, the buyer will be able to purchase another sets of items. Rule of Association The process of mining is used here. It aids in the increase of sales as well as the layout of the business in accordance with the purchasing habits of the customers. R, SAS (Statistical Analyses System), MEXL, XLMINER, and other data mining tools are utilised in this discipline.

3.6 Future Health Care

These days, Electronical Health Data [4] are extensively utilised, and we get vast amounts of patient data. Classifications, Associations Rules, and Clustering are data mining techniques that are used to uncover links between illnesses and treatments, identify novel medications, detect fraud and abuse, and reduce expenses in this industry. Rapid miners, R programming, Weka, Orange, and NLTK are some of the data mining technologies used in healthcare coverage (Natural Languages Tool Kit).

3.7 Manufacturing Engineering

Data about a company's goods is stored in a manufacturing business. Data mining techniques like classifications, associations rule mining, and regressions are used to anticipate products development time and costs, as well as the link among products architecture, customer demands, and job dependencies. Rapid miners, Data melt, Board, and Weka are some of the data mining technologies utilised in this industry.

3.8 Web Mining

Online mining employs data mining techniques to identify useful web pages and trends on websites. Online content mining (extracting relevant information from web publications), Web Structural mining (discovering structure informations from the website), and Web Usages mining are examples of applications that employ classification, clustering, and regression methods (log mining). SAS (Statistical Analyses System), Scrapy, Page rank, and other data mining tools were employed in this study.

4. Programs for Data Mining Data mining

Programs for Data Mining Data mining is a relatively new technique with a lot of room for improvement. Notwithstanding this, it is already being utilized on a regular basis by a range of enterprises. Retail stores, hospitals, banks, as well as insurance companies are just a few examples. Many of these companies are combining data mining with some other important technologies such as statistics and patterns recognition. Data mining can be used to find patterns and correlations that otherwise would be difficult to find. Many businesses use this technologies since it allows them to learn more about their customers and make good marketing decisions. The following is a list of corporate difficulties and solutions identified using data mining.

4.1 The FBTO Dutch Insurances Company Faces Difficulties to save money on direct mail

- Increases cross-selling to current clients utilizing inbound channels including the company's sales centre as well as the internet a 1-year testing of the solution's performance Results The marketings staff now has the capacity to forecast the success of its efforts.
- Made marketing campaign development, optimization, and implementation more efficient.
- Decreased mailing expenses by 35% while increasing exchange rates by 40%.

4.2 ECtel Ltd., Israel Challenges Fraudulent activities in telecommunications services

- Outcomes More than 150 telecoms firms saw a significant reduction in telephone fraud.
- All throughout the globe by providing real-time fraud detections, we were able to save money.

4.3 Dependable Financials United Kingdom Home Credit Divisions Challenges There is no mechanism in place to identify and prevent fraud

- Outcomes Fraud by agents and customers has decreased in frequency and extent.
- Saved money by detecting fraud early.
- Increased prosecution rates and saved time for investigators.

4.4 Challenges Facing Standards Life Mutual Financial Services Companies determine the essential characteristics of customers who are interested in their mortgages offer.

- Cross-sell Standards Life Banks products to others Standard Life firms' customers.
- Create a remortgages model that can be used on the group's website to assess the profit abilities of the mortgage businesses that Standard Life Bank accepts."

5. Conclusion

Each field is now digitalized, and as a result, a tremendous amount of data is created every day. Data mining is critical for organising, analysing, and retrieving the information needed from these massive datasets. There is a discussion of the many uses of data mining, as well as the methodologies and tools employed in every application.

6. References

1. Bharati Ramageri M. Data Mining Techniques and Applications‖ in Indian Journal of Computer Science and Engineering. 2014;1(4):301-305.
2. Saima Anwar Lashari, Rosziati Ibrahim, Norhalina Senan, Taujuddin NSAM. Application of Data Mining

Techniques for Medical Data Classification: A Review‖ in MATEC Web of Conferences 150, 06003, (2018), MUCET, 2017. Available: https://doi.org/10.1051/matecconf/201815006003

3. Usha Rani D. A Survey on Data Mining Tools and Techniques in Medical Field‖ in International Journal of Advanced Networking & Applications (IJANA). 2017;08(05):51-54. Special Issue.
4. Manpreet Kaura, Shivani Kanga. Market Basket Analysis: Identify the changing trends of market data using association rule mining‖ in International Conference on Computational Modeling and Security (CMS 2016), Procedia Computer Science. 2016;85:78-85.
5. Dr. Dhanabhakyam M, Dr. Punithavalli M. A Survey on Data Mining Algorithm for Market Basket Analysis‖ in Global Journal of Computer Science and Technology, 2011, 11(11). Version 1.0, Publisher: Global Journals Inc. (USA) Online ISSN: 0975-4172 & Print ISSN: 0975-4350.
6. Mohammad Reza Keyvanpoura, Mostafa Javidehb, Mohammad Reza Ebrahimia. Detecting and investigating crime by means of data mining: a general crime matching framework‖ in Procedia Computer Science. 2011;3:872-880. Available: http://www.sciencedirect.com
7. Uddin Osemengbe O, Uddin PSO. Data Mining: An Active Solution for Crime Investigation‖ in IJCST, 2014, 5(SPl-1).
8. Abhijit A, Sawant Chawan PM. Study of Data Mining Techniques used for Financial Data Analysis‖ in International Journal of Engineering Science and Innovative Technology (IJESIT). 2013;2:3.
9. Pushpesh Pant, Sriram Pandey. Application of Data Mining Tools and Techniques in Material Selection‖ in International Journal of Scientific & Engineering Research. 2017;8:4.
10. Jha VK, Singh RK. Application of Data Mining in Manufacturing Industry‖ in International Journal of Information Sciences and Application. ISSN 0974-2255. 2011;3(2):59-64.
11. Ashish Dutti, Maizatul Akmar Ismaili, Tutut Herawan. A Systematic Review on Educational Data Mining‖ in Digital Object Identifier 10.1109/ACCESS.2017.2654247, 2017, 5.
12. Dr. Vijiyarani S, Suganya E. Research Issues in Web Mining‖ in International Journal of Computer-Aided Technologies (IJCAx). 2015;2:3.