A Study On Applications Of Data Mining

Koti Neha, M Yogi Reddy

Abstract: Advancement in technologies has led to the emergence of various new fields. Different fields like science, Engineering, Health, Business are generating, accumulating high amounts of data every day. Data mining is a vital area, which manages and extracts required information from these enormous amounts of data. This paper gives a gist of how data mining is utilized in different fields.

Index Terms: Applications, Data mining, Data set, KDD, Techniques, Tools

1 INTRODUCTION

DATA MINING is like extracting needed information from an ocean of data. Significant step in Knowledge Discovery in Databases (KDD) is Data mining [1],[2]. The process of KDD is retrieving required information from large databases or data marts and converts the data into different patterns, summary reports, views etc. KDD has several stages as below:

1. Data Extraction: We choose Data sets or data samples from large databases or repositories.
2. Data Cleaning: The data set from the large databases may not be proper. We need to simplify the data by removing inconsistent data, missing values using data transformation tools.
3. Data Integration: Data from different resources is combined and stored in a single source. We use Data migration tools, Data Synchronization tools in this step.
4. Data Selection: Data which is useful for the analysis is taken from the data source using decision trees, Naive Bayes, Clustering, Neural networks and Regression techniques.
5. Data Transformation: Data is transformed into a suitable format of mining procedure using summary, aggregation operations.
6. Data Mining: Data is converted into meaningful patterns using a few methods.
7. Pattern Evaluation: Useful patterns are identified using Interestingness measures. We use visualization and Summarization techniques in this step.
8. Knowledge Presentation: Data mining results are represented using visualization tools into reports, tables etc.

Choosing a dataset from a vast repository is the primary thing in the process of data mining [3],[7]. Classification of datasets is of three types, namely Record data, Graph-based data, Sequential data. Record data is in the mode of collection of records. Each record contains related information among data fields. For example, market-basket data. Graph-based data is the data with relationships among objects, for example, linked web pages. Sequential data is an ordered list of events.

- *Koti Neha is presently working as Assistant Professor in GITAM deemed to be University, Hyderabad, India. E-mail: nehak9104@gmail.com*
- *M Yogi Reddy is presently working as Assistant Professor in GITAM deemed to be University, Hyderabad, India. E-mail: iamyogireddy@gmail.com*

Based on the features of events, three kinds of sequential data we have, namely Time-series data, Symbolic sequence data, Biological sequence data. Based on the application and dataset different tools and methods of data mining are used to retrieve required data.

2 DATA MINING TECHNIQUES

Based on data mining task type, appropriate techniques are applied. Descriptive data mining tasks categorizes features of data in a target data set based on past or recent events. Predictive tasks provide the results of future queries based on past data. Classification, Clustering, Regression, Outlier detection, Association rules, Sequential patterns and prediction are few most commonly used data mining techniques. Classification, Regression and Outer detection are predictive data mining techniques. Clustering, Sequential pattern discovery, Association rules are descriptive data mining techniques.

2.1 Classification

In Classification, we build a model which identifies and assigns a class for the new observation input data given to the model. We divide data into two sets training set (used to build the model) and test set (used to validate the model). From the Training set data, various classes are divided. Test set data is assigned a class by the model we generate. We use Decision Trees, Bayesian Classifiers, Neural Networks, K-Nearest Neighbour, Support Vector Machines, Linear Regression, Logistic Regression, as classifiers in this technique. Example: Mobile phone before buying it whether it good or bad based on its features like battery life, performance, cost, we decide it as a good class or bad class. Some of its applications are Direct marketing, Sky survey cataloguing, Fraud detection etc.

2.2 Clustering

Clustering technique segregates data into groups or clusters where objects within the cluster must have similar features while objects in different clusters must be less similar to each other. Different clustering methods are used according to the application. Some of them are Partitioning Method, Grid-Based Method, Density-based Method, Model-Based Method, Hierarchical Method, Constraint-based Method. Some of the applications are Document clustering, Market segmentation, Biology, medical imaging, Social network analysis etc.

2.3 Regression

Regression technique predicts the value of a continuous-valued variable called predictor variable (target) based on the response variable (whose values are already known). Some of the Regression algorithms used in data mining are Simple Linear Regression model, Lasso Regression, Logistic

regression, Support Vector Machines, Multivariate Regression algorithm, Multiple Regression Algorithm. Example: Relationship between road accidents and rash driving can be predicted. Some of the applications are forecasting sales, financial forecasting, trend analysis, marketing, time series prediction, estimating fossil age etc.

2.4 Association Rule Mining

Association Rule mining technique finds patterns in data and relationships among large data sets and correlations between them. The occurrence of an item can be predicted according to the occurrences of other items. Association rules are if-then rules using which lift, support and confidence are calculated to discover frequent patterns and relations between objects. Some of the Association rule algorithms are the Apriori algorithm, FP-growth algorithm, Eclat algorithm, Market-basket analysis, cross-marketing, catalogue designing uses this technique.

2.5 Outlier Detection

Outlier detection detects and excludes outliers (sample data which completely behaves differently compared with other data sets) from the data set. Some of the outlier detection methods are Z-Score, DBSCAN, Isolation Forest, Linear Regression Models (LMS, PCA), Proximity Based Models (non-parametric), High Dimensional Outlier Detection Methods. Some of the applications are Fraud detection, Intrusion detection, Medical and health outlier detection, Fraud detection of Insurance claim etc.

2.6 Sequential Patterns

Sequential patterns technique is used to predict sequential dependencies and sub sequences. Methods used for finding sequential patterns are GSP (Generalized Sequential Pattern), Free span, Prefix span, SPADE (Sequential PAttern Discovery using Equivalent Class). Some of the applications are DNA sequences, weblog click streams, telephone calling patterns, stocks and markets etc.

3 APPLICATIONS OF DATA MINING

Many diverse areas use data mining methods for technical, commercial and research purposes. Summary of most widely Data mining techniques and tools in various applications are listed in Table 1 below:

3.1 Bioinformatics

Bioinformatics [15] is the collection of various methods to manage, store and study biological data using computers. The data in this field were increasing every day and used extensively for research purposes. Data mining applications in this field include gene sequence finding, protein sequence analysis, gene and protein communication network construction, disease detection, DNA sequencing and aligning etc. Sequence data set is used in bioinformatics. Based on the application type, this sequence dataset is given to appropriate data mining tool to retrieve required results. Some of the data mining tools used in bioinformatics are BLAST (Basic Local Alignment Search Tool), FASTA, CS-BLAST for finding sequence alignment, GenScan, GeneMark for gene finding, Pfam, BLOCKS, ProDom for protein analysis etc.

3.2 Financial Banking

Digitalized banking generates vast amounts of transactional data every day [9]. Data mining is utilized in this filed for applications like financial banking are finding customer loyalty, issues loans and credit cards based on customers previous data, identifying stock market risks from historical data etc. These applications use Classification algorithms like Bayes classification, Boosting, Decision tree, Random forest. Data mining tools used in business and finance are Rapid Miner, R programming, Weka (Waikato Environment for Knowledge Analysis), Orange, KNIME, NLTK (Natural Language Tool Kit) etc.

3.3 Education

Educational data mining field is a new sector, which focuses on developing methods which discover required information from various educational areas [12],[16]. Data mining applications in this field are predicting students results, students learning behaviours, finding weak students etc. Learning patterns of students are used to develop teaching methods. Record data set is used in Education application. Data mining tools used in Education are SPSS, KEEL, Weka, Spark MLLib etc.

3.4 Criminal Investigation

Criminal analysis includes detecting crimes and criminal's relationships with these crimes. From different crimes like cyber-crimes, violent crimes, fraud detection, drug offences, we get high volumes of criminal datasets [7],[8]. Data mining is utilized in this field for applications like counter-terrorism activities, crime matching, crime trends, etc. Data mining tools used in this field are Weka, H2o, Orange etc. are field.

3.5 Market Basket Analysis

Market Basket Analysis [5],[6] is used to predict customer behaviour in the retail industry. It is based on theory like if a specific set of items are purchased, then the customer may buy another set of items. Association Rule Mining technique is applied here. It helps in increasing sales and also to arrange the store layout according to the customer's purchase behaviour. Data mining tools used in this field are R, SAS (Statistical Analysis System), MEXL, XLMINER etc.

3.6 Future Health Care

Electronic Health Records [4] are used these days widely, and we get large volumes of patient data. Techniques of data mining like Classification, Association Rules, Clustering are used to detect relationships among diseases and treatments, identify new drugs, fraud detection and abuse, decrease costs in this field. Data mining tools used in health care are Rapid miner, R programming, Weka, Orange, NLTK (Natural Language Tool Kit).

3.7 Manufacturing Engineering

Manufacturing enterprise contains data related to its company's products [10],[11]. Techniques like Classification, Association Rule mining, Regression in data mining are used to predict product development time and cost, the relationship between product architecture, customer needs, dependencies among tasks etc. Data mining tools used in this field are Rapid miner, Data melt, Board, Weka.

3.8 Web Mining

Web mining uses methods of data mining to discover relevant web documents and patterns from websites [13],[14].

Classification, Clustering, Regression techniques are used in applications like Web content Mining (to extract useful information from web documents), Web Structure Mining (to discover structure information from the website), Web Usage Mining (log mining). Data mining tools used here are SAS (Statistical Analysis System), Scrapy, Page rank etc.

S.NO	Application	Data Mining Techniques	Dataset Type	Tools
1.	Bioinformatics	Clustering, Classification, Association Rule Mining	Biological sequence data, Symbolic Sequence data	BLAST, Electronic PCR, Entrez Gene, ORF, Model maker, Pfam, GenScan
2.	Financial Banking	Classification, Sequential Pattern mining	Record data, Time-series data	Rapid Miner, R programming, Weka, Orange, NLTK, KNIME, Teradata
3.	Education	Classification, Regression, Clustering, Association Rule Mining	Record data	SPSS, KEEL, Weka, Spark MLLib
4.	Criminal Investigation	Clustering, Classification	Record data, Graph based data	Weka, H_2o, Orange
5.	Market Basket Analysis	Association Rule Mining	Record data	R programming, SAS, MEXL, XLMINER
6.	Future Health Care	Classification, Clustering, Association Rule Mining	Record data, Sequential data	Rapid Miner, R programming, Weka, Orange, NLTK, KNIME
7.	Manufacturing Engineering	Classification, Regression, Association Rule Mining	Record Data	Rapid Miner, Data melt, Board, Weka
8.	Web Mining	Classification, Clustering, Regression	Graph-based data	SAS, Scrapy, PageRank

Table 1: *Data Mining Techniques and tools used in various Application*

4 CONCLUSION

Every field is digitalized these days, and because of this, a large volume of data is generated every day. Data mining plays a vital role in managing, analyzing and extracting the required information from these large databases. An overview of different applications of data mining and its techniques and tools used in each application are discussed.

5 REFERENCES

[1] Aarti Sharma, Rahul Sharma, Vivek Kr. Sharma, Vishal Shrivastava, "Application of Data mining-A Survey Paper" in International Journal of Computer Science and Information Technologies, Vol. 5 (2), 2023-2025 2014, ISSN: 0975-9646

[2] Bharati M. Ramageri, "Data Mining Techniques and Applications" in Indian Journal of Computer Science and Engineering, Vol. 1 No. 4 301-305, Dec 2014.

[3] Saima Anwar Lashari, Rosziati Ibrahim, Norhalina Senan, N. S. A. M. Taujuddin," 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

[4] D.Usha Rani, "A Survey on Data Mining Tools and Techniques in Medical Field" in International Journal of Advanced Networking & Applications (IJANA), Volume: 08, Issue: 05 Pages: 51-54 (2017) Special Issue.

[5] 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 85 (2016) 78 – 85.

[6] Dr. M. Dhanabhakyam, Dr. M. Punithavalli, "A Survey on Data Mining Algorithm for Market Basket Analysis" in Global Journal of Computer Science and Technology Volume 11 Issue 11 Version 1.0 July 2011, Publisher: Global Journals Inc. (USA) Online ISSN: 0975-4172 & Print ISSN: 0975-4350.

[7] MohammadReza Keyvanpoura, Mostafa Javidehb, Mohammad Reza Ebrahimia," Detecting and investigating crime by means of data mining: a general crime matching framework" in Procedia Computer Science 3 (2011) 872–880. Available: http://www.sciencedirect.com

[8] Uddin, Osemengbe O., P. S. O. Uddin, "Data Mining: An Active Solution for Crime Investigation" in IJCST Vol. 5, SPI - 1, Jan - Mar 2014.

[9] Abhijit A., Sawant, P. M. Chawan, "Study of Data Mining Techniques used for Financial Data Analysis" in International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 2, Issue 3, May 2013.

[10] Pushpesh Pant, SriramPandey, "Application of Data Mining Tools and Techniques in Material Selection" in International Journal of Scientific & Engineering Research, Volume 8, Issue 4, April-2017.

[11] V.K. Jha, R.K. Singh "Application of Data Mining in Manufacturing Industry" in International Journal of Information Sciences and Application. ISSN 0974-2255 Volume 3, Number 2 (2011), pp. 59-64.

[12] Ashish Dutti, Maizatul Akmar Ismaili, Tutut Herawan, " A Systematic Review on Educational Data Mining" in Digital Object Identifier 10.1109/ACCESS.2017.2654247, Volume 5, 2017.

[13] Dr. S. Vijiyarani, Ms. E. Suganya, "Research Issues in Web Mining" in International Journal of Computer-Aided Technologies (IJCAx) Vol.2, No.3, July 2015.

[14] Brijendra Singh, Hemant Kumar Singh, "Web Data Mining Research: A Survey" in IEEE International

Conference on Computational Intelligence and Computing Research, 2010.

[15] Stefano Lonardi, Jake Chen, “Data Mining in Bioinformatics: Selected Papers from BIOKDD” in IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 7, no. 2, April-June 2010.

[16] Katrina Sin, Loganathan Muthu, “Application of Bigdata in Education Datamining and Learning Analytics- A Literature Survey” ICTACT Journal on Soft Computing: Special Issue on Soft Computing Models for Big Data, July 2015, Volume: 05, Issue: 04.