Data Mining (DM) Techniques and Applications – Decades Review

Ms. Kruttike Bang[1] Ms. Madhuri Bidwe[2]
[1,2]DYPCOE, India

Abstract— Data Mining (DM) is also known as Knowledge Discovery in Database (KDD). It is also defined as the process which includes extracting the interesting, interpretable and useful information from the raw data. This is the main reason the applications of Data Mining (DM) are increasing rapidly. This paper reviews Data Mining (DM) techniques and its applications such as educational Data Mining (DM) (EDM), finance, commerce, life sciences and medical etc. We group existing approaches to determine how the Data Mining (DM) can be used in different fields. Our categorization specifically focuses on the research that has been published over the period 2009-2019. With this categorization, we present an easy and concise view of different models adapted in the Data Mining (DM).

Keywords: Educational Data Mining (DM) (EDM), Knowledge Discovery in Database (KDD), Learning Management System (LMS), Social Network Analysis (SNA)

I. INTRODUCTION

Data Mining (DM) techniques (DMT) are used to transform raw data to useful information or knowledge. Data itself is nothing, but to process it, is very useful and interesting [1]. There are many advance techniques that uses data as useful information smartly. For example, Knowledge Discovery KD in Database also known as KDD is the process of required output extraction in different formats from raw data. KDD is also defined as the process to view useful patterns in data [2]. A generic and most common diagram of Data Mining (DM) or KDD is shown in Fig.1.

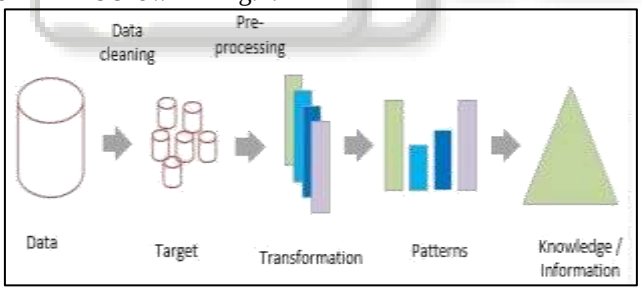

Fig. 1: Knowledge discovery from raw data

EDM is the major application of DMT. It is mentioned as an emerging program in education which explores many types of data produced by the educational institutions. It belongs to the literature that is related to Data Mining (DM), visualization, machine learning and computation. Moreover, methods used in machine learning are Naïve Bayes, Neural Networks, K-Nearest Neighbor, Decision Trees and many others [4]. Other fields are also merged with Data Mining (DM). For example, there are many proposed approaches that are the combination of Data Mining (DM) and semantic web [5]. Similarly, Data Mining (DM) techniques along with machine learning are used in many applications [6].

Large volume of data is produced from web based e-learning which is common today. This huge amount of data is generated by web servers that can be different obtained from multiple web servers. There are two main sources of data production in EDM that are traditional classrooms and distance education. In case of classrooms, teachers and students are present physically. Educators observe students' behavior by attendance, course information, exams, curricular activities and planning. Educational Data Mining (DM) helps each individual associated with institution. Student needs to understand how to choose or select courses based on prediction which course is the best. Instructors need to know which teaching experiences are the best and most contributive to the class. Data Mining (DM) techniques (classification, clustering, text mining, pattern matching etc.) are applied on the data obtained from web. Not only in education, Data Mining (DM) is used by all departments like administration, accounting, Human Resource (HR) and many more. Table below, shows us different tools used in EDM with their task.

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

Table 1. Data Minig Tools

II. RESEARCH WORK IN DATA MINING (DM)

As Data Mining (DM) has become most popular and its use has become most common. It makes automated systems by applying different Data Mining (DM) techniques to data flow. Many algorithms are applied in Data Mining (DM) techniques to solve real life problems. There are many advantages of Data Mining (DM) like it is helpful in banking, finance, accounting, retail, marketing, manufacturing, governments and many more [7]. In the same way, it also has many disadvantages as there are security issues, privacy issues, misuse of information, use of inaccurate information, risk of data loss etc. With the passage of time, Data Mining (DM) is growing and has been improved. There are many journals and articles written about it. Data Mining (DM) can be used in different perspective with respect to dataset given to solve a specific real time issue [8].

In case of distance education, different techniques are applied to grant access to the students who are far from space and time of lectures in traditional class rooms. Distance education involves internet education, web-based education, multimedia education and videotape education. This type of education creates the history of users' accesses in web logs [9].

It is compulsory to convert the data into particular format to use in a suitable Data Mining (DM) algorithm [10]. Some important processes used to format the data before implementation of Data Mining (DM) algorithms, are given in Table 2.

Process name	Objective
Data Cleaning	Irrelevant data is removed from the raw data. Only useful data is left that is needed for the specific mining algorithm
User identification	Referring associated the specific page to the user
Transaction identification	It makes smaller units of sessions as per transaction
Data transformation	It is creation of new attributes from the existing data
Data integration	Data is integrated and synchronized
Data reduction	Data is reduced according to dimensions

Table 2: Data Formatting Processes

III. CATEGORIES OF DMT

Data Mining (DM) techniques are applied with respect to different aspects of Data Mining (DM) as data obtained from different sources can be different and asynchronous. Data Mining (DM) is a vast field and found in every field and department. There are nine categories of DMT [13], that are discussed below:

A. Information Systems

Information systems provide a bridging tool between business world and computer scienc. Information systems have become the popular field among all other fields.

B. System Optimization

'Linear Programming' was the original term used for systems optimization, in the past.

C. Knowledge-based Systems

Knowledge-based systems are the core of artificial intelligence. Their base is Artificial intelligence.

D. Modeling

It is used for understanding of a complex structure and flow of a system through different perspective. Modeling techniques are used to analyze the data quantitatively.

E. System Architecture Analysis

A system has hardware architecture, operating system architecture, enterprise architecture and software architecture.

F. Algorithm Architecture

An algorithm is defined as a finite list of instructions to solve a problem. Algorithms are used for data processing and calculation. Steps involved in the development of an algorithm are shown in Fig. 2.

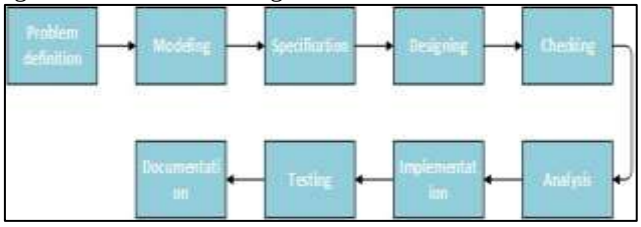

Fig. 2: Algorithm development

IV. DATA MINING (DM) TECHNIQUES

Data Mining (DM) is a vast field and it has a large number of applications, so it has become interesting subject to research. Data Mining (DM) techniques are classified into characterization, generalization and association [15]. There are different measures to use Data Mining (DM) techniques as the use of Data Mining (DM) is tricky but helpful if properly used [16]. Some Data Mining (DM) methods are classified and briefly discussed below:

A. Clustering

Special diverse repositories are used to store such data. In clustering, groups of different objects and their classes are made on the basis of their different aspects like location; connection etc. For example, schools can be grouped on the basis of their similarities or differences.

Similarly, students can be clustered on the basis of their behavior. The purpose of clustering is to search data points that are naturally grouped together.

B. Prediction

Prediction often depends on previous knowledge and experience. It is the focus on a single aspect of data with respect to some other aspect of data, called predictor variable.

C. Relationship Mining

Relation mining also known as relational Data Mining (DM) is commonly used for relational database. In database, relational Data Mining (DM) algorithm search for pattern among different patterns. Relationship between variables must satisfy two things: interestingness and significance [17].

D. Outlier Detections

Generally, if the new observation is different from the existing one compared, it is named as outlier. Outlier detection compares different values with smallest or largest values in a data set and finds the deviation among values.

E. Text Mining

This Data Mining (DM) technique described as the text data in Data Mining (DM) is specific with text data. Text data include documents, emails, messages, and html files. Text mining can be classified as document processing, document summarization, indexing, topic clustering and mapping [18]. It is commonly used in education and business. Text mining involves machine learning, statistics and natural language processing.

F. Social Network Analysis (SNA)

In the process of social media sites analyzing, relationship between different entities in network information is detected. It is commonly used to analyze the activities of a group or community.

G. Process Mining

It extracts the business processes knowledge that is related to process of event log.

H. Data Distillation for Judgment

In this method, data is represented intelligently. This technique uses visualization and summarization. This is useful to see and explore large amount of data at a time.

I. Applications of data mining (DM) methods

There are many applications of Data Mining (DM) methods. Some of them are discussed below:

1) Statistics

In the Data Mining (DM), user of applications is the main subject. These visualized data can be about assignments, exams, courses and marks. Instructors can get information about their students and distance classes.

2) Web Data Mining (DM)s

Web Data Mining (DM) is also an application of DM. Here, information is filtered from data obtained from web. The main purpose of web Data Mining (DM) is to facilitate users with information they seek [20].

According to Paul B [21], classification technique is used to:

- Select students with same characteristics
- Find student misuse
- Find student who are hint-driven in multiple choice questions

In the Data Mining (DM), common web mining techniques are clustering, classification, text mining, association rule, outlier detection and sequential pattern. These are briefly discussed below:

a) Classification and clustering

Classification and clustering are almost defined the same. Clustering make groups of pages with same contents or users. Classification characterizes the group of user profile and course sessions.

b) Sequential patterns and association rules

This relation among attributes creates if-then statements. Sequential patterns tell that which content gives access to the other content.

V. Conclusion

Data being the core entity in every field needs to be managed in efficient way. Data Mining (DM) helps a lot in this regard. The main issue faced today, is data privacy and data security. In case of global data sharing, privacy becomes more important, especially for web. Therefore, our future work includes the data privacy and security by applying a specific security algorithm that would not harm the data efficiency.

References

[1] "Spatial Data Mining (DM) and geographic knowledge discovery—An introduction," Comput. Environ. Urban Syst., vol. 33, no. 6, pp. 403–408, 2009.
[2] M. Goebel and L. Gruenwald, "A Survey of Data Mining (DM) and Knowledge Discovery Software Tools," SIGKDD Explor. Newsl., vol. 1, no. 1, pp. 20–33, 1999.
[3] A. Dutt, M. A. Ismail, and T. Herawan, "A Systematic Review on Educational Data Mining (DM)," vol. 3536, no. c, 2017. an overview from a database perspective," IEEE Trans. Knowl. Data Eng., vol. 8, no. 6, pp. 866–883, 1996.
[4] N. Jain and V. Srivastava, "Data Mining (DM) Techniques: a Survey Paper," IJRET Int. J. Res. Eng. Technol., vol. 2, no. 11, pp. 116– 119, 2013.
[5] S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data Mining (DM) techniques and applications – A decade review from 2000 to 2011," Expert Syst. Appl., vol. 39, no. 12, pp. 11303–11311, 2012.
[6] C. F. Chien and L. F. Chen, "Data Mining (DM) to improve personnel selection and enhance human capital: A case study in high- technology industry," Expert Syst. Appl., vol. 34, no. 1, pp. 280– 290, 2008.
[7] J. Han and J. Han, "Data Mining (DM) techniques," in Proceedings of the 1996 ACM SIGMOD international conference on Management of data - SIGMOD '96, 1996, vol. 25, no. 2, p. 545.
[8] L. Geng and H. J. Hamilton, "Interestingness measures for Data Mining (DM)," ACM Comput. Surv., vol. 38, no. 3, pp. 1–32, 2006.
[9] R. Baker, "Data Mining (DM) for education," Int. Encycl. Educ., 2010.
[10] Y.-H. Tseng, C.-J. Lin, and Y.-I. Lin, "Text mining techniques for patent analysis," Inf. Process. Manag., vol. 43, no. 5, pp. 1216– 1247, 2007.
[11] W. Fan, L. Wallace, S. Rich, and Z. Zhang, "Tapping the power of text mining," Commun. ACM, vol. 49, no. 9, pp. 76–82, Sep. 2006.
[12] R. Iváncsy and I. Vajk, "Frequent pattern mining in web log data," Acta Polytech. Hungarica, vol. 3, no. 1, pp. 77–90, 2006.
[13] P. Baepler and C. Murdoch, "Academic Analytics and Data Mining (DM) in Higher Education," Int. J. Scholarsh. Teach. Learn., vol. 4, no. 2, Jul. 2010.
[14] M. J. Shaw, C. Subramaniam, G. W. Tan, and M. E. Welge, "Knowledge management and Data Mining (DM) for marketing," Decis. Support Syst., vol. 31, no. 1, pp. 127–137, 2001.
[15] S. H. Ha and S. C. Park, "Application of Data Mining (DM) tools to hotel data mart on the Intranet for database marketing," Expert Syst. Appl., vol. 15, no. 1, pp. 1–31, 1998.
[16] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining (DM) to Knowledge Discovery in Databases," AI Mag., vol. 17, no. 3, p. 37, 1996.
[17] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, "The application of Data Mining (DM) techniques in financial fraud detection: A classification framework and an academic review of literature," Decis. Support Syst., vol. 50, no. 3, pp. 559–569, 2011.
[18] S. Wang, "A comprehensive survey of Data Mining (DM)-based accounting-fraud detection research," 2010 Int. Conf. Intell. Comput. Technol. Autom. ICICTA 2010, vol. 1, pp. 50–53, 2010.
[19] E. Kirkos, C. Spathis, and Y. Manolopoulos, "Data Mining (DM) techniques for the detection of fraudulent financial statements," Expert Syst. Appl., vol. 32, no. 4, pp. 995–1003, 2007.
[20] M. Ramageri, "Data Mining (DM) Techniques and Applications," Indian J. Comput. Sci. Eng., vol. 1, no. 4, pp. 301–305, 2010.

[21] C. Apte, B. Liu, E. P. D. Pednault, and P. Smyth, "Business applications of Data Mining (DM)," Commun. ACM, vol. 45, no. 8, pp. 49– 53, Aug. 2002.
[22] C. Yen and H. Wang, "Applying Data Mining (DM) to telecom churn," vol. 31, pp. 515–524, 2006.
[23] R. Bellazzi and B. Zupan, "Predictive Data Mining (DM) in clinical medicine: Current issues and guidelines," Int. J. Med. Inform., vol. 77, no. 2, pp. 81–97, 2008.
[24] "Selected techniques for Data Mining (DM) in medicine," Artif. Intell. Med., vol. 16, no. 1, pp. 3–23, 1999.
[25] B. D. Pitt and D. S. Kitschen, "Application of Data Mining (DM) techniques to load profiling," in Proceedings of the 21st International Conference on Power Industry Computer Applications. Connecting Utilities. PICA 99. To the Millennium and Beyond (Cat. No.99CH36351), 1999, pp. 131–136.
[26] A. M. Wilson, L. Thabane, and A. Holbrook, "Application of Data Mining (DM) techniques in pharmacovigilance," Br. J. Clin. Pharmacol., vol. 57, no. 2, pp. 127–134, Sep. 2003.
[27] S. Palaniappan and R. Awang, "Intelligent Heart Disease Prediction System Using Data Mining (DM) Techniques," IEEE/ACS Int. Conf. Comput. Syst. Appl., pp. 108–115, 2008.
[28] K. Wang, "Applying Data Mining (DM) to manufacturing: the nature and implications," J. Intell. Manuf., vol. 18, no. 4, pp. 487–495, Jul. 2007.
[29] S. Tahmasebian, M. Ghazisaeedi, M. Langarizadeh, and M. Mokhtaran, "Applying Data Mining (DM) techniques to determine important parameters in chronic kidney disease and the relations of these parameters to each other," vol. 6, no. 2, pp. 83–87, 2017.
[30] S. Bandaru, A. H. C. Ng, and K. Deb, "Data Mining (DM) methods for knowledge discovery in multi-objective optimization: Part A - Survey," Expert Syst. Appl., vol. 70, pp. 139–159, 2017.
[31] B. Boubacar, B. Kamsu-foguem, and F. Tangara, "Data Mining (DM) techniques on satellite images for discovery of risk areas," Expert Syst. Appl., vol. 72, pp. 443–456, 2017.
[32] C. Romero and S. Ventura, "Educational Data Mining (DM): A Review of the State of the Art," IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev., vol. 40, no. 6, pp. 601–618, Nov. 2010.
[33] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get another label? improving data quality and Data Mining (DM) using multiple, noisy labelers," in Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and Data Mining (DM) - KDD 08, 2008, p. 614.
[34] H.-P. Kriegel, K. M. Borgwardt, P. Kröger, A. Pryakhin, M. Schubert, and A. Zimek, "Future trends in Data Mining (DM)," Data Min. Knowl. Discov., vol. 15, no. 1, pp. 87–97, Jul. 2007.

