An Empirical Study of Data Mining Issues in Higher Education

Sumit Garg, Arvind Sharma, Uma Kumari

Abstract— Now days, the huge amount of data stored in educational databases increasing rapidly. The educational databases contain hidden useful information with many important factors related to the student's learning. Data mining techniques have been applied to analyze these data and bring out the hidden knowledge. This paper focuses on the data mining and major issues associated with it. This study clarifies how data mining and KDD process are related together. The aim of this paper is how data mining techniques implemented on educational datasets to predict the performance of the students in higher education.

Index Terms— Data Mining, KDD, Data Mining Issues.

1 Introduction

In 21st century the human beings are used in the different technologies to adequate in the society. Each and every day the human beings are using the huge amount of data in the different fields. The data may be used in different formats like numbers, text, figures, hypertext formats and audio or video. As the data are available in different formats so that the proper action to be taken. Not only to analyze these data but also take a good decision and maintain the data. As and when the customer will be required the data should be retrieved from the database and make the better decision. This technique is generally known as Data mining or Knowledge Hub or KDD process. With the enormous amount of data stored in files, databases and other repositories, it is increasingly important to develop powerful tool for analysis and interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. The only answer to all of the above is Data Mining. Data mining is the extraction of hidden information from large databases [1,2]. It is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses [3,4].

The paper is organized in different sections. Section 2 explains the difference between KDD and Data mining. Section 3 summaries the literature review. Section 4 discusses different major issues in data mining. Section 5 contains data mining functions. Conclusion is shown in section 6 while references are mentioned in the last section.

- *Sumit Garg is currently pursuing M.Tech Degree in Computer Science & Engineering from Shekhawati Engineering College affiliated to Rajasthan Technical University, Kota, India. E-mail: sumit.grg1@gmail.com*
- *Arvind K Sharma has submitted his Ph.D in Computer Science in Jaipur National University, Jaipur, India and he is working as guest faculty in Dept. of Computer Science & Informatics, University of Kota, Rajasthan, India. E-mail: arvindsharma133@gmail.com*
- *Uma Kumari is working as Asst. Professor in Computer Sceince in Shekhawati Engineering College, Dundlod, Jhunjhunu, Rajasthan, India.*

2. KDD V/s Data Mining

KDD stands for Knowledge Discovery in Databases is a key field of Computer Science. It contains different tools and techniques which have been used for extracting useful and previously unknown information from large collection of database. Data mining is a step of KDD process shown in figure 1. The two terms KDD and Data Mining can be used interchangeably. KDD is the process of extracting knowledge from data while Data Mining is a step inside the KDD process. Data Mining is an application of the specific algorithm of the KDD process[5].

Fig.1: Data Mining is a key step towards KDD

2.1. Data Mining–Junction of Multiple Domains

Data mining is a technique that is used to extract knowledge from huge amount of data. Data warehousing is the storage of data from different sources in the organization while data mining is the exploration of information from the data stored in data warehouse. Data mining is a process of analyzing data

from different views and converts this data into useful information[6]. The extracted information can be used to increase the revenue and cut the cost [7]. Due to the importance of extracting information from the large data set, data mining has been become an important component in various fields of human life. Data mining use statistics, machine learning, artificial intelligence, pattern recognition, computational capabilities for the advancement in business, education, medical, and scientific etc.[8]. Different data mining applications have been successfully implemented in different domains such as health care, finance, retail, telecommunication, fraud detection, and risk analysis. Data mining is basically convergence of multiple key domains. The figure 2 below shows the same.

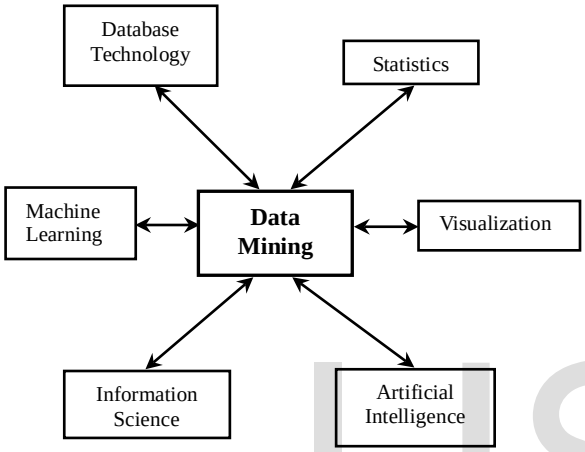

Fig.2: Data Mining-Junction of Multiple Domains

3. Literature Review

It is worthwhile to mention here that, although the multiple domains of data mining are shown above. Educational data mining is an area full of exciting opportunities for researchers. This field helps higher educational institutions with efficient ways to improve institutional effectiveness and student learning, yet different authors have given their research contributions in this field discussed below:

In one of the work[9] Banumathi and Pethalakshmi have given a novel approach for upgrading Indian education by using clustering data mining technique to analyze the performance of students through proposed UCAM clustering algorithm. This improves the scalability and reduces the clustering error.

In[10] Monika Goyal and Rajan Vohra have proposed a methodology to use data mining technique to improve the efficiency of higher education institute. OLTP (on line transaction processing) has developed as a complete ERP solution for academic institutes especially for engineering colleges.

In [11], Richard A Huebner has focused on analyzing data to develop models for improving learning experience and improving institutional effectiveness. This study has been focused exclusively on ways that data mining used to improve student success and process directly related to student learning.

In [12], Chady EI Moucary has presented a predictive model with a powerful decision making that considerably affects engineering programs in various types of higher education institutions. It assists in planning the courseware and designating needed faculty members.

In [13], Brijesh Kumar Baradwaj and Saurabh Pal have given a data mining model for higher education system in the university to increase the quality of education for students in educational institutions. In this work, the decision tree has used on student database to predict the student division on the basis of previous database.

In [14], Feng Shi and Qi Miao have been used data mining technique such as association rule in college curriculum to help the teachers to arrange courses, scientific guidance for teaching and learning, and improve the teaching management.

In [15], C.P. Samaranayake and H.A. Caldera presented the failure of a large number of talented students in the physical science stream of education advance level examination. For this some association rules are used.

In [16], Sunita B Aher et al. presented a survey of an application of data mining in education system and presented result analysis using WEKA tool. It uses ZeroR algorithm for classification, and DBSCAN algorithm for clustering.

In [17], Bhise R.B. et al. has applied educational data mining methods to discover knowledge from educational databases. K-means clustering has used to evaluate the factor likes midterm and final exam assignments that helps the teacher to reduce drop-out ratio to a significant level.

In [18], Munpreet Singh Bhullar and Amritpal Kaur has used J48 data mining algorithm to predict the result of the students and the data presented in the form of decision tree.

4. Major Issues in Data Mining

Data mining contains some major issues regarding user interaction, performance and diverse data type, which are discussed below:

4.1 Mining Methodology and User Interaction Issue: This issue shows the types of knowledge mined the ability to mine knowledge at multiple granularities, domain knowledge, ad hoc mining and knowledge visualization.

4.2 Mining Different Kinds of Knowledge in Database: Different users can be interested in different types of knowledge. Data mining should use a large number of data analysis and knowledge discovery task [19] such as characterization, discrimination, association, classification, clustering, tread and deviation analysis and similarity analysis. These tasks use the same database in different ways.

4.3 Interactive Mining of Knowledge at Multiple Level of Abstraction: Data mining process should be interactive to know exactly what can be discovered within a database users can be focus on the search for patterns by using interactive data mining.

4.4 Incorporation of Background Knowledge: Knowledge may be used to guide the discovery process and

allow discovered patterns to be expressed in concise terms and at different level of abstraction.

4.5 Data Mining Query Language and Ad-hoc Mining: Relational Query language i.e. SQL allows users to pose ad-hoc queries for data retrieval.

4.6 Presentation & Visualization of Data Mining Result: Discovered knowledge should be represent in high level language, visual representation or other user friendly forms. So the knowledge can be easily understood by human.

4.7 Handling Noisy Data: The data stored in database may contain noise, exceptional case or incomplete data object. As a result the accuracy of the system became poor. Noise can be removing by the data cleaning methods and data analysis method. Some outlier methods should be use for discovery of exceptional cases.

4.8 Pattern Evaluation: A data mining system can be discovering many patterns. Some of these patterns may be unused for user because they contain a common knowledge.

4.9 Performance Issue: These include efficiency, scalability and parallelization of data mining algorithm.

4.10 Efficiency and Scalability of Data Mining Algorithm: Data mining algorithm must be efficient and scalable for the extraction of information from a huge amount of data in database. Running time of a data mining algorithm must be predictable and acceptable in large database.

4.11 Issue Related to the Diversity of Database Types: In the current scenario different types of relational database and data warehouse are used. Data may be hypertext and multimedia, spatial data, temporal data or transaction data. One system cannot be managing all these data. So specific data mining system should be constructed for specific type of data.

4.12 Mining Information from Heterogeneous Database and Global Information System: Local and wide area computer network connect many source of data, forming huge, distributed and heterogeneous database.

5. Functions of Data Mining

Data mining identifies facts or suggests conclusions based on shifting through the data to discover either patterns or anomalies. Data mining contains five main functions [20][21]:

5.1 Classification: It infers the defining characteristics of a certain group such as customers who have been lost to competitors.

5.2 Clustering: It identifies groups of items that share a particular characteristic. Clustering differs from classification in that no predefining characteristic is given in classification.

5.3 Association Rule: It identifies relationships between events that occur at one time such as the contents of a shopping basket.

5.4 Sequencing: It is similar to association, except that the relationship exists over a period of time such as repeat visits to a supermarket or use of a financial planning product.

5.5 Forecasting: It estimates future values based on patterns within large sets of data such as demand forecasting.

6. Conclusion

Data mining techniques have been frequently used to extract useful information from the large volume of data. The educational databases consists hidden knowledge for predicting the performance of the students. In this paper, data mining with KDD process and their relational behavior have been discussed. We have also tried to identify the research area in data mining where further work can be continued. In future, the classification techniques like decision trees, rule mining, Bayesian network etc. will be implemented with educational dataset for the prediction of student's performance.

Acknowledgment

The author wishes to express deep gratitude to Arvind Sharma for the encouragement and extensive support in preparing and publishing of this paper.

References

[1] Introduction to Data Mining and Knowledge Discovery, Third Edition, Two Crows Corporation, MD 20854 (U.S.A.), 1999.

[2] Larose, D. T., "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN 0-471-66657-2, John Wiley & Sons, Inc, 2005.

[3] Dunham, M. H., Sridhar S., "Data Mining: Introductory and Advanced Topics", Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006.

[4] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. "CRISP-DM 1.0: Step-by-step data mining guide, NCR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringenen Bank Group B.V (The Netherlands), 2000.

[5] Arun K Pujari, "Data mining Technique published by Universities Pras (I) Pvt. Limited, Hyderabad-500029.

[6] Arvind K. Sharma and P.C. Gupta, "Predicting the Number of Blood Donors through their Age and Blood Group by using Data Mining Tool", International Journal of Communication and Computer Technologies, Vol.1-No.6, Issue: 02 Sept. 2012.

[7] Venkatadri M., et al., "A review on data mining from past to the future", international journal of computer application, Vol.15- no.7. Feb 2011.

[8] Difference between KDD and data mining, http:// www.difference between.com/ difference between KDD and data mining.

[9] Banumathi & Pethalakshmi, "A novel approaches for upgrading Indian education by using data mining technique"

[10] Monika Goyal & Rajan Vohra, "Application of data mining in higher education", IJCSI International Journal of computer science Issues, Vol.9, No.1, March 2012

[11] Richard A Huebner, "A survey of educational data mining research"

[12] Chady El Moucary, "Data mining for Engineering Students (Predicting student's performance and enrollment in Master program)"

[13] Brijesh Kumar Baradwaj & Saurabh Pal, "Mining Educational Data to analyze student's performance", (IJACSA) Vol.2, No.6, 2011

[14] Feng Shi, Qi Miao, "The application of data association mining technology in university curriculum Management" 2012 IEEE Symposium on Robotics and application.

[15] C.P. Samaranayake, et al., "A Data Mining solution on high failure rate in Physical Science Stream at the University science stream at the university entrance examination", Tenth International Conference on ICT and Knowledge Engineering, 2012.

[16] Sunita B Aher et al., "Data Mining in educational System using WEKA", International Conference of Emerging Technology Trends (ICETT) 2011, Proceedings published by International Journal of Computer Application (IJCA)

[17] Bhise R.B., et al., "Importance of data mining in higher education system", IOSR Journal of Humanities and Social Science (IOSR-JHSS), Volume-6, Issue-6(Jan.-Feb. 2013),pp 18-21
[18] Munpreet Singh Bhullar & Amritpal Kaur, "Use of Data Mining in Education Sector", Proceedings of the World Congress on Engineering and Computer Science 2012 Vol I, WCECS 2012, October 24-26,2012,San Francisco,USA
[19] Osmar R. Zaina, 1999 Principal of Knowledge discovery in database Page-13 University of alberata, Deptt. of Computer Science.
[20] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.
[21] Campos, M. M., et al., "Data-Centric Automated Data Mining", www.oracle.com/technology/products/bi/odm/pdf/automated_data_mining_paper_1205.pdf

IJSER