INTERNATIONAL JOURNAL FOR INNOVATIVE RESEARCH IN MULTIDISCIPLINARY FIELD
ISSN: 2455-0620 [Impact Factor: 6.719]
Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value : 86.87
Volume - 8, Issue - 1, JAN - 2022 Publication Date: 31/01/2022

DOIs:10.2015/IJIRMF/202201037 --:-- Research Article

Data Mining in the field of Higher Education

Shilpa K.

Asst. Professor, Dept. of Computer Science, Nrupatunga University, Bangalore
Email – shilpaktr@gmail.com

Abstract: *Data Mining is the method of extracting patterns and drawing inferences from large and complex datasets. Educational Data Mining(EDM) is about finding patterns and drawing inferences from data that comes from the education system. It is an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data obtained from educational system, and uses those methods to better understand students and the system in which they learn. This paper surveys the relevant studies in the EDM field and includes the methodologies used to explore the data from Education system and Applications of Data mining in Higher Education.*

Key Words: *Educational Data Mining, Data Mining Process, methods, clustering, association, prediction, classification.*

1. INTRODUCTION:

Good education forms the basis for a country to be healthy. Providing comprehensive education to its young minds should be the top priority for any government. Education should be addressed 360-degree angle viz. academics, personality development, soft skills, computers, etc. This is one of the measure to see any of the country's social and economic growth.

Government of India provide free education for all citizen from grade one to throughout the higher education or by minimal fee structure. So that a large number of students enter into universities every year for their higher education. For Ex. Due to corona pandemic all the students of 12th standard or PUC (in Karnataka) who had appeared for exams were promoted. Due to which there was a huge inflow of students at universities and colleges. The major challenge that arises is having good infrastructure, experienced teaching staff, etc. to such a large influx of students.

Due to pandemic all the educational institutions started conducting online classes, many students who were staying in remote villages had to face mobile network issues to connect to online classes, as there is no proper network, bandwidth and speed to connect to online classes. All students were promoted to next semester without any assessment, students were unable to secure any knowledge in their academics due to poor infrastructure for online classes. Due to this, students now after the pandemic when the colleges started conducting regular classes in colleges are facing difficulty in understanding the subjects, concepts, etc. they lack basics of any given topics. Using data mining techniques one can analyse the student's current understanding level and can bridge the gap.

It is a cumbersome process to analyse the data manually due to huge volume of data. The gap between data and intake has been reduced by using different tools. It can also be stated to as Knowledge discovery from data(KDD).

Knowledge Discovery from data: is the process of non-trivial mining of implied, unknown and potentially useful information from a large database. In Knowledge Discovery, data mining has been utilised to find patterns related to user requirements. The pattern definition is an expression that describes a subset of data [1][31].

The accurate detection of patterns through data mining depends on several factors, such as size, sample, data integrity and support from domain knowledge, all of which affects the degree of certainty needed to identify the pattern.

There is a huge volume of data in the field of education, the increase in technology has made this happen. With the help of this huge amount of data extensive information may be extracted [2]. DM techniques can be used to analyse this educational data. The main aim of EDM is to develop methods that use unique type of data. These developed methods are inturn used to enrich the learning process and improve the quality of teaching [3].

The EDM also allows user to gather information from students' data. This information will help in validating and evaluating the educational system, the learning process, etc. [5]

Most of the educational system's issues can be resolved using EDM. The four key area of EDM suggested by Baker [8] are improving student models, improving domain models, studying pedagogical support provided by learning software

and conducting scientific research on learning and learners. Five approaches/methods are available: prediction, clustering, relationship, mining, distillation of data for human judgement and discovery with models [8].

2. DATA MINING

Data mining is the process of looking for patterns in large data sets. As a result of this procedure, we may draw valuable inferences about the data. This also creates fresh data regarding the data that we currently have. Pattern tracking, categorization, association, outlier identification, grouping, regression, and prediction are some of the approaches used. Because there might be a quick change in the data provided, patterns are easier to detect. We have collected and categorized the data based on different sections to be analysed with the categories. Data is clustered into categories based on their commonalities.

In DM different algorithms can be used for solutions. Few algorithms can explore the data and few extract a specific result based on that data. Clustering algorithms recognise patterns and group data into different set of groups. This data which is available in each group is more consistent and help to create decision model.

The DM used in KD has discovered patterns with respect to user's need. It is an interactive process which examines decisions made by the users.

First, try to gather information on the domain in which the application is to be developed. Have prerequisite knowledge and know the goals of the user. Second, select the target dataset and get the variable subset or data samples which is targeted for examination. Third, clean up the noise and inconsistent data. Fourth (the data reduction and projection phase), get the features that are useful for representing the data viz. reduction of dimensionality or transformation methods. Fifth, use the KD goals and choose the correct DM strategy. Sixth, search for correct patterns by matching the dataset with DM algorithms. Seventh, get the correct patterns from a set or a representational form. Eighth, understand these gathered patterns and go to any of the above steps for more iteration. Finally, document, prepare the report or take action of the knowledge discovered.

3. DATA MINING PROCESS

Every author comes up with his/her own steps for data mining. [6] data mining is "the process of selection, exploration, and modelling of large quantities of data to discover regulations or relations that are at first unknown with the aim of obtaining clear and useful results for owner of the database."

Data mining process rendered by different authors that helps to compare the process with each one of them.

Giudici [6] explained DM in four steps, they are as follows:

1 Strategic
2 Training
3 Creation
4 Migration

Bryman [7] explained DM in six stages, they are:

1 Selection
2 Pre-Processing
3 Transformation
4 Data Mining
5 Interpretation
6 Evaluation

Hsu [8] defined DM in four steps as follows:

1 Data Collection
2 Data Pre-processing
3 Data Mining
4 Information interpretation and Visualisation

Berry [9] explained DM in 11 steps, they are as follows:

1 Translate the business problem
2 Select appropriate data
3 Get to know the data

4 Create a model set
5 Fix problems with the data
6 Transform data to being information to the surface
7 Builds models
8 Assess models
9 Deploy models
10 Assess results
11 Begin again

Kantardzic [10] explained DM in five stages, they are as follows:

1 State the Problem
2 Collect the data
3 Pre-Process the data
4 Estimate the model
5 Interpret model and draw conclusions

CRISP-DM (Cross-Industry Standard Process Data Mining) is the standard process generally used by industries. The steps of CRISP-DM Process are - Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment are.

4. DATA MINING METHODS

There are many tasks of data mining practicing by various industries. Prediction, classification, association and clustering are the most important tasks of data mining.

4.1. Prediction

Sen et al. (2012) built models to predict secondary education placements-test using sensitive analysis on predicting models (Decision Tree Algorithm, Support Vector Machine, Neural Network and Logistic Regression). They have identified the following important predictors of placement-test: previous test experience, student has a scholarship or not, students' number of sibling and previous years' grade point average (GPA).

They indicate that decision tree analysis is the best predictor, followed by support vector machine, and neural network. Logistic regression is the least predictor.

[11] built model to predict slow learners in education field using prediction based data mining algorithm like Sequential minimal optimization, Decision tree algorithm, reduced error pruning decision tree and Waikato environment for knowledge analysis. They have identified the factors associated with students whose academics performance is below average and to improve the quality of education by identifying slow learners so that they are assisted individually by their teachers to improve the performance.

4.2. Classification

Classification analyse a set of data and generate a set of grouping rules which can be used to classify future data [12]. Different algorithms used for classification are Decision Tree induction, Hunt's algorithm, ID3, C4.5 algorithms, etc.

4.3. Association

According to [13] association rule is "the process of discovering interesting association or relationship among data items." It will give the summary of entire data. [14] said association rules for two numeric attributes and one Boolean attribute.

[15] indicates that association rule can be applied to discover relationship between the characteristics of the students and helped to find relationship perfectly [1]. [3] association rule is 'discovering of if-then rule' which means if the value of one variable is found, the value of another
variable will have specific value.

4.4. Clustering

Clustering is grouping of similar objects, its task is unsupervised classification. [16] they used algorithms like ANN, exception maximization, hierarchical clustering, apriori algorithm, C-Means clustering methods, Markov clustering, K-Means clustering, etc.

5. OTHER METHODS OF EDM:

5.1 Outlier Detections:

Outliers are extreme values that deviate from other observations on data, they may indicate a variability in a measurement, experimental errors or a novelty. In other words, an outlier is an observation that diverges from an overall pattern on a sample. Outlier detection approach is divided into three types, statistical approach, the distance-based approach and deviation based approach [17].

Outlier detection methods can be used to identify decreased student or teacher performance that is not normal, to identify students at the extreme ends of the performance spectrum [18][19].

Outlier is used to detect students with learning problems. Ueno [20] proposes a method of online outlier detection of learner's irregular learning processes by using the learners' response time data for e-learning data.

5.2 Text Mining:

The process of transforming unstructured text into a structured format to identify meaningful patterns and new insights. It is a part of DM to extract valuable text information from a text repository [21]. DM and text mining techniques are used to analyse two different data sets for educational courses. The study found the discrepancies and similarities in the students' pattern [22]. Ueno [20] uses data mining and text mining technologies for collaborative learning and a discussion board with evaluation between peers in an ILMS.

5.3 Social Network Analysis (SNA):

Seeks to understand the connection and relationships that form between individuals or communities, most commonly expressed as node and edge diagram. It is commonly used to analyse collaborative social networks such as those seen in social media, or in student interaction within in MOOCs or online courses [23]. The main goal is to better understand and how students learn and identify the settings in which they learn to improve educational outcomes. There are some case studies about SNA for e.g. García-Saiz [23]and others used a multiple regression analysis to evaluate the factors that influence participation in learning communities. Rabbany and others built Meerkat-ED a tool designed to access the students' participation in online courses [24]-[35].

6. APPLICATIONS OF DATA MINING IN HIGHER EDUCATION

6.1. Predicting Admission of Students

The model built by Aksenova et al. (2006) for freshers, existing and the students who dropped off from the college at graduate and undergraduate levels. The model took different parameters into consideration, viz. economic strata, population, tuition fees, past students' data of the institution, etc. When the data was mined using Cubist tool, the result was beyond the imagination and it was concluded that data mining has huge scope in higher education.

Predicting the students' success based on the admission data using CART (Classification and Regression Technique) technique by Kovacic, J. Zlatko (2010) could identify and predict the students who are vulnerable of dropping off from the course and could groom them with orientation programs, mentoring, etc. in coming out with flying colours.

6.2. Predicting Profiling of Students

Data mining can help the institution with some special needs depending on the information furnished by the student at the time of admission, viz demographical, geographical, psychological, etc. Romdhane et al. (2010) pointed out that data mining can be used for predicting such special needs.

Chen et al. (2005) mentioned that can be used to define customer's behaviour. A common and simple way of collecting customer's profile is by conducting surveys. Techniques like neural networking helps to identify different varieties of students. Also, different patterns can be identified using Discriminant analysis. Different techniques like Regression analysis, decision tree and Bayesian classification can also be applied. Subsequently, cluster analysis can be prepared on students' profiling with which different marketing strategies can be planned to target students. Cluster analysis can also be called as data segmentation (Sinha et al., 2010).

6.3. Developing Curriculum

The study by Hsia et al. (2008) shows that using data mining algorithms like decision tree, link analysis and decision forest one can study the course preferences, completion rates and profession of enrolled person. During the study it was found that there is a relation between enrolled profession and the category of the course. This will help in building better curriculum and better marketing strategies in higher education.

6.4.Feedback from Students

Chen et al. (2012) proposed a model/technique called 'PARA' (P=Primary Diagnosis, A=Advanced Diagnosis, R=Review and A=Action) for handling students complaints better. This technique is helped better in identifying and categorizing the complaints and improving the services.

6.5. Course Completion

Institutions can gather information regarding loyalty of the students, degree of satisfaction and complaints in order to understand the pattern of students' completion of course.
Dr. Mohd Maqsood Ali, International Journal of Computer Science and Mobile Computing Vol.2 Issue. 4, April- 2013, pg. 374-383

6.6. Library

Data mining methods can be used in library of the institution to explore students report which details the selection of books, loan accounts and book shelves to gather additional information. Apart from this, clustering analysis can be applied to understand, E.g. books selection, ordering system with regard to gender, age, and grades of students.

6.7. Selection of course by Student

Kardan et al. (2013) determined that the factors that influenced the students in selecting the course using neural network, such as student's workload, grade in the course, type of the course, duration of the course, conflicts if any, time of the exam, etc. These parameters act as the input for the neural network model. Adding to this Guo (2010) analysed and predicted student course satisfaction using neural networks. He came to a conclusion that the number of students enrolled for a course and coming out with more than distinction in final grading were the two important factors for student's satisfaction in the course.

6.8. Performance of Teachers' in teaching

Stepwise regression and decision tree approaches were utilised by Mardikyan and Badur (2012) to find the elements that impact teaching success in colleges/universities. Attitude of the teacher/lecturer, status of the employee, attendance of the student and the feedback from the students affect the performance of the teacher/lecturer.

6.9. Performance of the Students

Kumar and Uma (2009) investigated student performance in the course using data mining approaches, namely the Nave Bayes and Decision Tree classification algorithms. depending on the factors like student's ID, marks scored etc. Apart from this they recommended that the data mining process to be done on the teachers' performance that helps to perform better in colleges/universities.

6.10. Students' Dropouts

Massa and Puliafito (1999) conducted study on the issue of dropout student in the university, they used the Makov Chains mining technique. They concluded that using this technique the behaviour of identical group of students can be studied, it can also define clusters of students related to different degree of dropout risk.

6.11. Relationship management of the students

Data mining techniques can be used to understand and analyse the critical data of student's relationship management. It can be used to gather, maintain and preserve the students to achieve effectiveness of the organisation.

CRM, according to Maqsood (2013), is "a combination of a company strategy, organisational process, and technology for gaining most profitable and unprofitable consumers by offering value-added services and maintaining stronger customer relationships." by offering customer loyalty programs." [36]-[40].

7. CONCLUSION:

Technology is being used vastly in education and it is generating huge volume of data every minute. Researchers around the world are taking advantage of this data for mining in the field of education. Data mining is growing at a fast pace where newer algorithms and techniques are developed to meet the need. Mining the educational data is opening up new challenges and is creating interest in extracting the data which is interpretable and useful. This extracted data when processed will help in better understanding the productivity of students and their by setting benchmark for each student

to learn. It also helps in understanding the risk factors that the students are facing, prioritising the learning needs, increasing the performance of the students, effectively analyse the standard of the institution, fine tuning of curriculum and many more. Modern methods, tools and techniques help the higher educational institution to bridge the gap between teaching – learning process.

REFERENCES:

1. Aggarwal, C. Charu and Yu, S. Philip. "Data Mining Techniques for Associations, Clustering and Classification." in Zhong, Ning and Zhou, Lizhu (Eds.) methodologies for knowledge discovery and data mining, third pacific Asia Conference, PAKDD, Beijing, China, April 26-28, 1999 proceedings, Springer, New York.
2. Ahn, Jin. Sook, and Sohn, S. Young. "Customer pattern search for after sales services in manufacturing." Expert System with Applications, 6 (2009): 5371-5375.
3. Baker, R.S.J.D., Data Mining for Education, in B. McGraw, P. Peterson, and E. Baker (Eds.), international Encyclopaedia of Education, Third Edition, Vol. 7, 2010, 112-118.
4. Aksenova, S. Sretlana., Zhang, Du and Lu, Meilin. "Enrolment prediction through Data Mining" IEEE Conference Publications (2006): 510-515.
5. Baritchi, Andi. Data Mining and Knowledge Discovery, in (Eds.,) Business Intelligence in the Digital Economy, Raisinghani, S. Mahesh, 2004, Idea Group Inc, 2004,
6. Giudici, Paolo. Applied Data Mining: Statistical Methods for Business and Industry, John Wiley and Sons Ltd., 2013.
7. Bryman, Alan and Cramer, Duncan. Quantitative Data Analysis with SPSS 12 and 13: A Guide for Social Scientists, 2005, Routledge.
8. Hsu, Hui-Huang. Introduction to Data Mining in Bioinformatics, in Hsu, Hui-Huang Reichgett, (Eds.), Advanced Data Mining Technologies in Bioinformatics, Idea Group Inc, 2006, 1-12.
9. Berry, J.A. Michael and Linoff S. Gordon., "Data Mining Techniques for Marketing, Sales, and Customer Relationship Management." Second Edition, Wiley Publishing, 2004.
10. Kantardzic, Mehmed. Data Mining: Concepts, Models, Methods and Algorithm, Second Edition, John Wiley and Sons, New Jersey, 2011.
11. P Kaur, M Singh, GS Josan – Procedia Computer Science, 2015 – Elsevier.s Classification and prediction based data mining algorithms to predict slow learners in education sector.
12. G. Kesavaraj and Sukumaran, IEEE 2013 4th international conference on computing communications and networking technologies – a study on classification techniques on data mining.
13. Gopalan, N.P and Sivaselvan, B. Data Mining: Techniques and Trends, PHI, New Delhi, 2009
14. K. Yoda, T. Fukuda, Y. Morimoto, S. Morishita, T. Tokuyama, Computing optimized rectilinear regions for association rules Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (Aug. 1997).
15. Romero, Cristobal., Ventura, Sebastian and Garcia, Enrique. 'Data Mining in Course Management Systems: Moodle case study and tutorial." Computers and Education 51 (2008): 368-384.
16. Ashish Dutt, Saeed Aghabozrgi, Hamid Reza and Maizatul Akmal Binti Ismail – Clustering Algorithms Applied in Educational Data Mining, International Journal of Information and Electronics Engineering, Vol. 5, No. 2, March 2015.
17. Bakar, Zuriana; Mohemad, Rosmayati; Ahmad, Akbar; Deris, Mustafa (2006). [IEEE 2006 IEEE Conference on Cybernetics and Intelligent Systems - Bangkok, Thailand (2006.6.7-2006.6.7)] 2006 IEEE Conference on Cybernetics and Intelligent Systems - A Comparative Study for Outlier Detection Techniques in Data Mining.
18. Basuki, Wibawa, B., Siregar, J. S., Asrorie, D. A., & Syakdiyah, H. (2021). Learning analytic and educational data mining for learning science and technology. THE 2ND SCIENCE AND MATHEMATICS INTERNATIONAL CONFERENCE (SMIC 2020): Transforming Research and Education of Science and Mathematics in the Digital Age.
19. X. Shu, *Knowledge Discovery in the Social Sciences: A Data Mining Approach* (University California Press Oakland, California, 2020).
20. Ueno, M. Data mining and text mining technologies for collaborative learning in an ILMS ''samurai''. In IEEE international conference on advanced learning technologies, Joensuu, Finland, 2004a, (pp. 1052–1053).
Jo, T. [2019]. Text mining studies in Big Data.

21. He, Wu (2013). Examining students' online interaction in a live video streaming environment using data mining and text mining. Computers in Human Behavior, 29(1), 90–102.
22. García-Saiz, D., Palazuelos, C., & Zorrilla, M. (2013). Data Mining and Social Network Analysis in the Educational Field: An Application for Non-Expert Users. Studies in Computational Intelligence.
23. Rabbany, R., & Takaffoli, M. (2011). Za¨ıane O. Analyzing participation of students in online courses using social network analysis techniques. In International conference on educational data mining.
24. Kumar, A.N.V. and Uma, G.V. "Improving academic performance of students by applying data mining techniques." European Journal of Scientific Research, No. 4 (2009): 526-534.
25. Kovacic, J. Zlatko, "Early prediction of student success: Mining students' enrolment data, in proceedings of informing science and IT education Conferences (InSITE) 2010, 647-665.
26. Larose, T. Daniel. Discovering knowledge in data: An Introduction to Data Mining Techniques, John Dr. Mohd Maqsood Ali, International Journal of Computer Science and Mobile Computing Vol.2 Issue. 4, April- 2013, pg. 374-383.
27. Luan, Jing. Data mining and its applications in higher education, new directions for institutional research, No. 113, 2002, Springer.
28. Nisbet, Robert and Elder, John and Miner, Gary., Handbooks of Statistical analysis and Data Mining Applications, Academic Press Publications, 2009.
29. Ngai, E.W.T, Xiu, Li and Chau, D.C.K "Application of Data mining techniques in customer relationship management: A literature review and classification." Expert System with Applications, 36 (2009): 2592-2602.
30. Massa, S and Puliafito, P.P. "An application of data mining to the problem of the university students'
31. dropout using Markov chains." in Zytknow, J.M and Rauch, J (Eds.), Principles of Data mining and Knowledge Discovery, third European conference, PKDD 99, Prague, Czech Republic, September 15-18, 1999, proceedings, Springer, vol. 1784, 51-60.
32. Maqsood, A. Mohammed, "Customer Relationship Management in B-Schools: An Overview." International Journal of Computer Sciences and Management Research, Vol. 2 issue 4 (April 2013).
33. Mardikyan, Sona and Badur, Bertan. "Analysing teaching performance of instructors using data mining techniques." informatics in Education, Vol.10, No.2, (2011): 245-257.
34. Mark, Last. Data Mining, in Cyber Warfare and Cyber Terrorism, Janczewski, Lech and Colarik, Andrew., Information Science Reforms, IGI Global, (2007): 358-365.
35. Olson L. David and Delen, Dursun. Advanced Data Mining Technique, 2008, Springer.
36. Pena, Alejandro. Dominguez, R., and Medd, J., "Education Data Mining: A sample of review and study case". Academic World and Education and Research Center, 2 (2009):118-139.
37. Pujari, K. Arun. Data Mining Techniques, Universities Press India Pvt. Ltd., Hyderabad, 2001.
38. Rajamani, Karthick., Sung, Sam., and Lox, Alan. "Extending the applicability of association rules." The proceedings of Third Pacific Asia Conference, PAKDD 99, Beijing, China, April 26-28, 1999.
39. Ramageri, M. Bharti "Data Mining Techniques and Applications." Indian Journal of Computer sciences and Engineering, Vol.1 No.4, (Dec-Jan 2010-2011): 301-305.
40. Rokach, Lior. "A survey of clustering algorithm," in O. Maimon, L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook, 2nd Edition, Springer, 2010.