Study of techniques and applications of data mining in Education

Dr.Nikhat Khan
Institute of Company Secretaries of India

ABSTRACT

This paper discusses about the data Mining process in the field of education. A database is a collection of data. It stores information which can be useful in the future. This data can be utilized for extracting useful information by applying various data mining techniques. These techniques when applied in the educational data obtained from various sources like offline sources, online sources or from Learning Management System is beneficial for the Universities, Institutes in decision making in various ways. Educational data mining can be used for classifying and predicting students' performance, dropouts as well as teachers' performance. It can help educators to track academic progress to improve the teaching process, it can help students in course selection and educational management to be more progressive, efficient and effective.

Keywords: Knowledge Discovery in data base, Education Data Mining (EDM), Learning Management System (LMS), Data Mining Techniques

1. INTRODUCTION

The term Data mining which is also referred as Knowledge Discovery in Databases (KDD), is an important area of finding novel and potentially beneficial information from large amounts of data. Data mining is applied in various fields, including medicines, bioinformatics, banking, weather prediction, engineering and counter-terrorism. The recent development has been to use data mining within the area of educational research. This area is termed as educational data mining.

Educational data mining (EDM) is an important and fast growing new area that combines multiple disciplines toward understanding how students learn and toward creating better support for such learning. There are various disciplines which include, human-computer interaction, machine learning, artificial intelligence coming under the fields of Computer Science, cognitive psychology, cognitive science, statistics and fields of education which include psychometrics, educational psychology, learning sciences. The data used for educational data mining possible is obtained from varied sources. This data is analyzed and then further used for addressing the questions related to improvement of student learning and the psychology of human learning. This paper discussed about the various fields in which education data mining can be used.

2. DATA MINING FRAMEWORK

The data mining is extensively used now a days in the field of educational research. Education data is collected and mined to get useful information to better understand students, teachers, curricula to be framed for students, business model for educational institutes etc.

Educational data mining (also referred to as “EDM”) is defined as the area of scientific inquiry centered around the development of methods for making discoveries within the unique kinds of data that come from educational settings, and using those methods to better understand students and the settings which they learn in.[1] Educational data mining methods often differ from methods from the broader data mining literature, in explicitly exploiting the multiple levels of meaningful hierarchy in educational data. In this various methods from the psychometrics literature are also integrated with methods from the machine based learning and data mining techniques to achieve the goal.

3. DATA OBTAINED FOR EDUCATIONAL DATA MINING

With the greater use of technology in educational systems, a large amount of data is available. Educational data mining (EDM) provides significant information .Based on this the data used for education data mining can be categorized as follows:

3.1 ***E-learning and learning management systems (LMS)*** *provide students with content available in system, instruction, communication, and reporting tools that allow students to learn by themselves. This type of education may occur as self-paced learning or may be instructor-led learning. Data mining techniques can be applied to the data stored by the systems in the databases.*

3.2 ***The data can be available from Offline education*** *where knowledge is transferred to learners based on face-to face contact. Data can be collected by traditional methods such as observation and questionnaires. It studies the cognitive skills of students and determines how they learn. Therefore, the statistical technique and psychometrics can be applied to the data.*

3.3 *Data can available from massive open online course (MOOCS), Intelligent tutoring systems (ITS) and adaptive educational hypermedia systems (AEHS).The data is obtained for different profiles of students. As a result, applying data mining techniques is important for building user profiles. The data generated by these systems can then assist in further research.*

Educational data mining is a significant field which uses educational data obtained from various sources to generate useful information. It uses multiple algorithms for getting results which helps in decision making. Learning is based on behavioral, psychological and constructive models. The learning outcome is based on observable changes in behavior of the student. The teacher’s involvement in the learning procedure results in the psychological model. Then there is constructive model in which the students learn on their own from the available resources.

Education is no longer an isolated activity. There is a drastic shift of conventional learning environment into community based learning conditions. There are various data mining algorithms which can be applied on the data obtained from various educational data sources which will help in improved learning outcomes and will be beneficial for the students in particular and the teachers and the institutions in general.

EDM is the process of transforming raw data obtained from educational systems into useful data that can be used to make data-driven decisions. Data mining is the most powerful technique with incredible potential to help various schools and colleges focus on the most significant information in the data sets collected about student's behavior and potential learners.

Data mining uses its tools to find previously unknown patterns and relationships in huge data sets. These tools can incorporate machine learning techniques, statistical models and mathematical algorithms. These techniques will help in acquiring useful information through the available data.

4. PROCESS OF DATA MINING IN EDUCATIONAL DATA:

Data mining which is a branch of computer science is used to operate on huge data sets to find hidden patterns and relationships, which is helpful for many organizations to make data-driven decisions. [2].Various techniques and algorithms such as Clustering, Association Rules, Classification, Regression, Neural Networks, Artificial Intelligence, , Genetic Algorithms, Decision tree, etc. are used for knowledge discovery from databases.[3]

An educational system typically has a huge educational database. This data may include teacher's data, accounts data, data related to various colleges, various modes of teaching, data related to several curricula, colleges, student's data, predicting students' results, students learning behaviors, weak students, alumni data, etc. The educational data mining focuses on the development techniques for exploring the special types of data that originate from an educational context. These data originate from various sources, including data from the traditional face -to- face classroom environment, online courseware, educational software, etc.

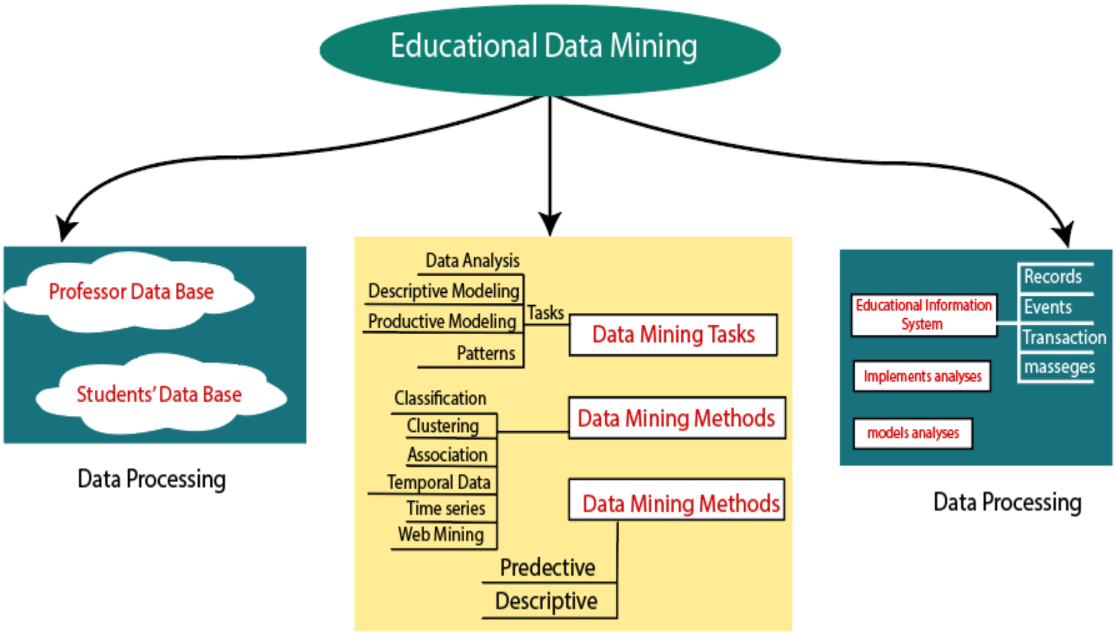

Fig 1: Process of Data Mining in educational data

In educational data mining, some techniques like Classification, association rule, clustering, Decision Tree, KNN, K-Clustering, Random Forest has been widely used by authors in their research work. [Fig 1]

Classification:

In Classification a model is built which identifies and assigns a class for the new observation input data given to the model. We divide data into two sets training set which is used to build the model and test set which is 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. In other words classification is the task of learning an objective function that maps each attribute set A to one of the predefined class level B. There are various classification techniques, namely Decision Tree-based Methods, Memory-based reasoning, Rule-based methods, Naïve Bayes and Bayesian Belief Networks, Neural Networks. In classification, test data is used to estimate the certainty of the classification rules. If the certainty is acceptable, the rules can be applied to *the new data tuples. The classifier-training algorithm utilizes these pre-classified examples to determine the arrangement of parameters needed for proper discrimination.*

Clustering:

Clustering refers to the process of classification and identification of objects into different groups, the segmentation of a data set into clusters such that the data in each cluster share some common characteristic of similar classes of objects.

Clustering technique segregates data into clusters or groups 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, Constraint-based Method, Grid Based Method, Density-based Method, Model-Based Method, Hierarchical Method, Clustering is a very efficient technique in grouping where it divides the data into groups or clusters with similar characteristics [4]. However, the size of data is reduced in clustering so some details are lost.

Regression

Regression technique predicts the value of a continuous valued variable called predictor variable (target) based on the response variable having known values. Regression techniques can be adapted for prediction. It can be used to demonstrate the relation between one or more independent and dependent variables. In data mining, independent variables are attributes that are already known, and response factors are what we need to predict. Many advanced techniques such as logistic regression, neural nets, and decision trees can be used to forecast future values.

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 student’s course curricula and student’s result can be predicted.

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.

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), and 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.

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

.

5.APPLICATIONS OF DATA MINING IN EDUCATION

5.1 Performance of Students:

This paper helps in the study of performance of students using different data mining techniques such as classification and clustering and provides a suitable technique that could be used by student advisors. This study will help the universities to improve the performance of the students. It introduces student marks prediction models using predictive model approaches based on student behavior [5].In this study the decision tree (J48) classification algorithm was used to analyze student performance. Further prediction of performance of students is also possible.

5.2 Building the smart Infrastructure:

Universities or institutes can build the smart infrastructure for education and allocate the budget or funding direction based on the decision making by doing the analysis of various data models and studying the success rate of various modes of learning. Smart infrastructure is a smart educational environment or a smart platform providing implementation, use and development of smart technologies in education. They can take decisions by doing the assessment of models having smart platform, smart technology, smart knowledge management system, use of learning tools and mobile devices for teaching.

5.3 Profiling of students:

The profiling of the student is made by using clustering algorithm. The students are divided into various clusters using data from various attributes of data like results scored in the semester, gender, break between semesters, average score in each semester, overall score. The students are profiled based on strugglers, Keen completers, Average, Late Completers, Based on the profiling the students the intelligent tutoring system is designed to get the best learning outcomes. Students of each cluster could benefit from the same intervention.

5.4 Students' Enrolment Prediction

Aksenova et al. (2006) build predictive model for fresher, existing and returned students at both graduate and undergraduate levels. This model is built based on population, unemployment rates in the region, institutional tuition fees, household income, enrolment past data of institutions. Data is mined with the help of Cubist tool.They conclude that data mining has an enormous application in higher education.Kovacic, J. Zlatko (2010) predicted students' success based on enrolment data (socio-demographic and environment variables) using data mining techniques such as CART (Classification and Regression Technique).They concluded that classifying students based on pre-enrolment information helps to identify students 'at-risk'of dropping the course and suggest using orientation, advising and mentoring programs to make them success.

5.5 Predicting Students' Profiling

Romdhane et al. (2010) indicate that data mining allows building customer models each describing the specific habits, need and behaviour of group of customers. It classifies new customers and predicts their special need. Consequently, data mining can help management to identify the demographic, geographic and psychographic characteristics of students based on information provided by the students at the time of admission.Profiles are often based on demographic and geographic variables (Berry and Linoff, 2004).According to Chen et al. (2005) data mining can be applied to describe behaviour of customers. Furthermore,surveys are one common method of building customer profiles. Neural networking technique can be used to identify different types of students. In addition, Discriminant analysis can also be used to identify patterns.Regression analysis, decision tree and Bayesian classification can be applied. Consequently, cluster analysis can be done to students' profiling and separate marketing strategies can be prepared to target segmented students.Cluster analysis is also called data segmentation (Sinha et al., 2010).

5.6 Library facility

Data mining techniques can be used in library to explore students' reports in relation to books selections, loan accounts and books shelves to gain information. In addition, clustering analysis can be applied to understand books selection and ordering system based on age, gender and grades of students.

5.7 Curriculum Development

Hsia et al. (2008) study course preferences, completion rates and enrolled professionals by using data mining algorithm such as decision tree, link analysis and decision forest. They found the correlation between course category and enrolled professions. They lay emphasis on importance of data mining in designing curriculum and marketing in the field of higher education.

5.8 Students' complaints

Chen et al. (2012) proposes 'PARA' (P=Primary Diagnosis, A=Advanced Diagnosis, R=Review and
A=Action) model of service failure based on customer complaints. Data mining technique is used to identify categories of different complaints and evolve strategy to improve services.

5.9 Students' course selection

In his publication Kardan et al. (2013) determine the characteristics that influences student course selection using neural networks such as, students' workload, final examination time, course grades, course type, course duration, and number of time conflicts, and students' demand. These variables are used as input of neural network modelling. Further, Guo (2010) did the analysis and subsequently predicted student course satisfaction using neural networks. He found that number of students are enrolled to a course in which they attain high distinction.

5.10 Course Completion

Universities and Institutes can cluster students into groups based on students' loyalty, students' satisfaction and degree of complaints to understand students' patterns towards course completion [6]
Dr. Mohd Maqsood Ali, International Journal of Computer Science and Mobile Computing Vol.2 Issue. 4, April- 2013, pg. 374-383 © 2013, IJCSMC All Rights Reserved 380

5.11 Students' Targeting

Woo et al. (2005) defines customer targeting as "a process of defining strategy to target specific customers. "They indicate that customer map is the visualization method for customer targeting. Customer map helps in building customer-oriented strategy. It is a "novel technique to find right set of customers who are homogenous with similar characteristics, values and needs. It is organized with three dimensions of customer targeting: Customer Value (usage and behavior), customer characteristics (demographic and Psychographic), and customers' needs (complaints and satisfaction). Target customers are detected and Customer map derives the targeting strategies.

6. CONCLUSION

Educational data mining is an upcoming discipline with high potential for every participant in the educational process. Data mining techniques are developed and used to automatically discover hidden knowledge and recognize patterns from data. Educational data mining can be used for classifying and predicting students' performance, dropouts as well as teachers’ performance. It can help educators to track academic progress to improve the teaching process, it can help students in course selection and educational management to be more progressive, efficient and effective.

Educational data mining can be used to attract, maintain and retain the students to achieve the profitability of University. Analyzing students’ data is crucial for discovering, detecting and understanding which instructional practices are effective. In this paper, we presented the study of various data mining techniques and applications of the same in many educational areas.

The main goal of the paper is to reveal the benefits of educational data mining applications and to encourage people to use it before taking crucial decisions related to education.

References

[1] Data Mining on Educational Domain https://doi.org/10.48550/arXiv.1207.1535

[2] Dr.Nikhat Khan, Dr. F. H. Khan and Dr. G.S. Thakur, "Weighted Fuzzy Soft Matrix Theory and its Decision Making," International Journal of Advances in Computer Science and Technology, vol. 2, no. 10, pp. 214-218, October 2013.

[3] Dr.Nikhat Khan and Dr. F. Z. Khan, "Data Mining for Fuzzy Decision system in Banking," CiiT International Journal of Data Mining Knowledge Engineering, no. January 2013.

[4] R. Saxena, “Educational data mining: Performance evaluation of decision tree and clustering techniques using weka platform,” International Journal of Computer Science and Business Informatics, 2015.

[5] H. Alhakami, T. Alsubait, and A. Aljarallah, “Data mining for student advising,” International Journal of Advanced Computer Science and Applications, vol. 11, no. 3, 2020.

[6] Dr. Mohd Maqsood Ali, International Journal of Computer Science and Mobile Computing Vol.2 Issue. 4, April-2013, pg. 374-383 © 2013, IJCSMC All Rights Reserved 380

AUTHOR

Dr. Nikhat Khan is a Visionary IT Professional with more than 20 years of experience in the IT sector spanning multiple domains and worked in various Government, private and public sector organizations. Adept at working in fast-paced environments and picking up new skills on the go. Providing effective leadership in implementing innovative and technical solutions, planning strategies to drive comprehensive growth. Enthusiastic team player, with a focus on meeting organizational goals and surpassing expectations. She has visited US, UK and German countries for managing and executing IT projects for renowned international clients. She has worked in multiple national and international organizations. Currently she is working as a Director in the Institute of Company Secretaries of India. She has received MCA degree from D.A.V.V, Indore in the year 1998 and PhD in Computers from Maulana Azad National Institute of Technology, Bhopal (MANIT).She is eager to know about the advances in business intelligence, data mining and fuzzy logic. She has number of publications in National and International Journals.