Educational Data Mining: A state-of-the-art survey on tools and techniques used in EDM

Ginika Mahajan, Bhavna Saini
Manipal University, Jaipur.

Abstract

Over the last decade, limitless data is generating in the field of education. To process this huge raw data, enormous potential is required beyond the manual and time consuming tasks. With the span of time, analytics and data mining is used to extract useful information from large data repositories. Educational Data Mining (EDM) exploits statistical, machine learning and data mining in the domain of education to analyze and predict the educational data using various approaches. EDM seeks to use online modes of learning to better understand learners and their learning, and develop computational approaches to analyze the facts and figures so as to benefit learners. Various machine learning algorithms and research tools are used in Educational Data Mining for analysis and prediction on different types of educational data. This paper presents a survey of applications and tools used in Educational Data Mining. Also, it presents detailed review of current trends in EDM where techniques and results of recent work done in this field are compared.

1 Introduction

Educational Data Mining (EDM) is an evolving interdisciplinary research domain that deals with development of computational and statistical models to analyze and explore educational data [3]. The core objective of EDM is to analyze education-based different types of data to resolve research issues in the domain of education [12]. When applied in the domain of education, these approaches are referred to as learning analytics (LA) and educational data mining (EDM). These two research communities, educational data mining and learning analytics, have embraced similar perspectives on analysis of educational data. The only difference is that EDM researchers are more concerned in applying automated models to get information from educational data while researchers in learning analytics are more towards human-led methods for exploring educational data [4]. Practically EDM allows to determine new knowledge based on students learning data in order to endorse and evaluate educational system, and hence to improve some aspects of quality of education [10].

2 Background

2.1 Data Mining and its Applications

Over the last decade, limitless data is generating in every field. To process this huge raw data, enormous potential is required beyond the manual and time consuming tasks. Data mining or “Knowledge discovery in databases” are the processes that extract the useful or most relevant insights out of the large datasets. There are a number of applications where data mining has been applied to discover the hidden facts and to predict the future trends. Healthcare is one of the emerging field where huge data was collected and analyzed to understand the illness, symptoms and complications [23][24]. Various previous studies had helped in prediction and in taking right decision at early stage [25][26][27]. Agriculture is another major application area of data mining. Using previous agriculture data, an early detection can be done for the food quality, plant diseases, color of affected area, nutrient deficiency and many more factors [28][29]. Along with this, data mining is also proving helpful in understanding the agricultural supply chain and operations such as production, storage, retails, distribution, etc. [30][31] Other than these banking sector is utilizing the data mining methods in evaluating the customer satisfaction and their needs [33]. It also helps in generating credit rating and identifying credit card frauds [34].
Data Mining methods are also major role in e-commerce, to understand the diversity of market, to meet the increasing demand of customers, etc. [36]. It also helped a lot in improving customer relationship [37].

2.2 Educational Data Mining

In this era, teaching and learning platforms have completely change. Nowadays everything is online whether its related to teaching content, teacher, evaluation, tests and quizzes. There is an abundant diversity in educational systems and environments like E-learning, Learning Management System (LMS), Adaptive Hypermedia (AH) educational systems, tests or quizzes, contents. Various other platforms are available that aid in learning such as various social networks, forums, educational game-based learning environment, Learning repositories, virtual environments, ubiquitous computing environments, etc. [3].

Traditional classroom method which is considered as offline mode, various Psychometrics and statistical techniques have been applied to understand the student/learners behavior and performance in classroom. In online mode like learning management system (LMS) and E-learning, the EDM techniques are applied on student’s data which is stored digitally in database. Extracting various features and factors from this data, EDM techniques can be applied to gain relevant information which can then be analyzed for further improvements.

EDM applies computational techniques for analysis and visualization of educational data. This analysis can be used to predict student’s performance or student’s strong and weak skills and knowledge. It can be used to detect undesirable student’s behaviors and providing recommendations for students. These models can assist instructors in grouping students, getting feedback, developing course with proper planning and scheduling. This paper presents a review of most relevant and current studies done in EDM.

3 Surveys on EDM

Various researchers are working in the field of EDM to analyse and do prediction in the field of education. This paper presents recent survey of various tools and techniques used in EDM.

Hui Chun et. al. [1] worked to explore learning behaviors of students in blended learning courses. Dataset was collected from a university in northern Taiwan where two classes of Python programming related courses of first-year students were considered for experimentation. Experimentation values of f1-score of random forest model was evaluated as 0.83. This score was better as compared with decision tree and logistic regression. Also authors implemented machine learning and symmetry-based learning algorithms to explore student’s learning behavior. Huei-TseHou et.al. [7] analyzed the videotaped learning process behaviors from 86 college students in simulation game-based learning activities. This study used an integrated technique of sequential analysis with cluster analysis to simulate learner behavioral patterns in games. Authors also categorized three clusters of learners with diverse pattern of learning processes. The results indicate that by using this integrated model of unsupervised learning one can explore the learners’ reflective behavior pattern in simulation games. Castro et.al. [2] used Machine Learning algorithms on Learning Management System. The proposed work has applied Learning Management System in teaching and learning process of Bulacan State University (BulSU) Graduate School (GS) Program. Authors applied Support Vector Machine which is a supervised machine learning algorithm for classification and to identify best video lecture topic-wise.

Acharya et.al. [8] predicted students’ performance using ML techniques - Naive bayes , C4.5, MLP (multi-layer perceptron), sequential minimal optimization (SMO), and KNN (1-Nearest Neighborhood). They considered four features and applied correlation-based feature selection (CBFS).
The results show that SMO attains effective average testing accuracy of 66% which is higher than other ML techniques used.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Author(s)</th>
<th>Year</th>
<th>Techniques</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1]</td>
<td>Hui-Chun Hung, Fan Liu, Che-Tien Liang and Yu-Sheng Su</td>
<td>2020</td>
<td>Random forest, decision tree and logistic regression</td>
<td>Experimentation values of f1-score of random forest model was evaluated as 0.83. This score was better as compared with decision tree and logistic regression.</td>
</tr>
<tr>
<td>[7]</td>
<td>Huei-TseHou</td>
<td>2015</td>
<td>Integrated cluster and sequential analysis</td>
<td>Applying this integrated model explored three clusters of learners with different patterns of learning processes and that learners with advanced flow levels reach more complete reflective process.</td>
</tr>
<tr>
<td>[8]</td>
<td>Acharya and Sinha</td>
<td>2014</td>
<td>Naive bayes , C4.5, MLP (multi-layer perceptron), SMO (sequential minimal optimization) and KNN (1-Nearest Neighborhood)</td>
<td>SMO attains effective average testing accuracy of 66% which is higher than other ML techniques used.</td>
</tr>
<tr>
<td>[13]</td>
<td>Shuangyan Liu , Mathieu Aquin</td>
<td>2017</td>
<td>K-Prototypes clustering Algorithm</td>
<td>Using this algorithm, one can identify groups of students (successful and weak students group) based on demographic characteristics and interactions learning in online environment, and can examine the learning achievement of each group.</td>
</tr>
<tr>
<td>[5]</td>
<td>EkanshMaheshwari, Chandrima Roy, Manjusha Pandey, and SiddharthSwarupRautray</td>
<td>2018</td>
<td>K-means, Naive Bayes and Random Forest</td>
<td>On applying naïve Bayes, a maximum accuracy of 82.35% and minimum accuracy of 57.35% is achieved. Using Random Forest, the maximum accuracy of 66.17% and minimum accuracy of 47.05% is achieved.</td>
</tr>
<tr>
<td>[9]</td>
<td>Elaf Abu Amrieh , ThairHamtini and Ibrahim Aljarah</td>
<td>2016</td>
<td>Artificial Neural Network, Naïve Bayesian and Decision tree ensemble methods - Bagging, Boosting and Random Forest (RF)</td>
<td>The ANN model outperformed other techniques, while Boosting was the best ensemble method.</td>
</tr>
<tr>
<td>[11]</td>
<td>Tsiakmaki,Kostopoulos, Koutsonikos, Pierrakeas, Kotsiantis, and Ragos</td>
<td>2018</td>
<td>LR, RF, 5NN, M5 Rules, M5, SMOreg, GP, Bagging</td>
<td>The obtained results show that RF, Bagging and SMOreg took precedence over other methods with MAE value ranging from 1.217 to 1.943.</td>
</tr>
</tbody>
</table>

Table 1. Survey of various techniques used by different authors and results obtained

Shuangyan et.al.[13] focused on performance of students in distance learning, forecasted how demographic variables and online learning activities affect student’s performance using unsupervised learning algorithm. They used clustering algorithm, k- prototype, to know and gather information about two group of students- Successful students group and weak students group. Categorizing these groups may help faculty to focus on students with poor learning outcomes and who need special attention. Using this algorithm, one can identify groups of students (successful and weak students group) based on demographic characteristics and interactions learning in online environment, and can examine the learning achievement of each group. Ekansh et.al [5] gave a prediction model to know direct and indirect factors that are affecting the dropout rate of Primary to High School Students in India. They used K-means algorithm to find classes and factors, for male is the percentage schools having toilet and the number of male teachers, while for female it is the percentage of schools having girl’s toilet and number of female teachers. For prediction naïve Bayes classifier and Random forest
was applied on the dataset. On applying naïve Bayes, a maximum accuracy of 82.35% and minimum accuracy of 57.35% is achieved. Using Random Forest, the maximum accuracy of 66.17% and minimum accuracy of 47.05% is achieved. Elaf et al. [9] proposed student’s performance prediction model based on student’s behavioral features. The obtained results forecasts a strong relationship between learner’s behavior and their academic achievement. Artificial Neural Network, Naïve Bayesian and Decision tree classifiers are used to evaluate the student’s performance in this model. Also to improve the performance of these classifiers, three ensemble methods are used- Bagging, Boosting and Random Forest (RF). The proposed model achieves an accuracy of 25.8% using ensemble methods. The ANN model outperformed other techniques, while Boosting was the best ensemble method. Tsiakmaki et al. [11] compared eight supervised learning algorithms on Weka environment – Linear Regression, Random Forests (RF), 5-NN, M5 Rules, M5 algorithm, Sequential Minimal Optimization algorithm for regression problems using SVM (SMOreg), Gaussian processes (GP), Bootstrap Aggregating (Bagging) – for predicting students’ marks. The evaluation metric used in this study to determine efficiency of regression methods is Mean Absolute Error (MAE). The obtained results show that RF, Bagging and SMOreg took precedence over other methods with MAE value ranging from 1.217 to 1.943. Table 1 provides a detailed survey of various techniques used by different authors and results obtained.

4 Tools used in EDM

Various machine learning algorithms and research tools are used in Educational Data Mining for analysis and prediction on different types of educational data. Although EDM is a rapidly emerging field and new tools and techniques are developing continuously, Table 2 provides a survey of various tools used in EDM for analysis in educational domain.

<table>
<thead>
<tr>
<th>Tool</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNIME [15]</td>
<td>Data cleaning and analysis package</td>
</tr>
<tr>
<td>Orange [16]</td>
<td>Orange is a data visualization and analysis package</td>
</tr>
<tr>
<td>SPSS [17]</td>
<td>A statistical package used for statistical tests, regression frameworks, correlations, and factor analyses.</td>
</tr>
<tr>
<td>KEEL [18]</td>
<td>KEEL used for analysis has classification and regression algorithms</td>
</tr>
<tr>
<td>The EDM Workbench [19]</td>
<td>tool for automated cleaning, organizing, and creating data with feature distillation and data labeling.</td>
</tr>
<tr>
<td>D3js [21]</td>
<td>Data visualization tool used for complex data visualizations that require data handling</td>
</tr>
<tr>
<td>PSLC DataShop [22]</td>
<td>Integrates data collection, construction, analysis, and visualization.</td>
</tr>
</tbody>
</table>

Table 2. Survey of various tools used in EDM

Conclusion

Educational data mining is an emerging research area which is needed in educational domain. It exploits statistics, machine learning and data mining techniques to get various students and instructor parameters for analysis and predictions. From a practical standpoint, this survey can be proved valuable for researchers working in the domain of Educational Data Mining. This paper presents a survey of tools used in EDM. Also it presents review of current trends in EDM where techniques of related work done in this field are compared. It is hoped that this paper will play a role in raising the profile of the educational data mining field and research community.
References


