Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/16/20230892

Educational data mining: Methods and applications

Dong Wang

Department of Software Engineering, Changan University, Shanxi Province, 710061, China

2020900971@chd.edu.cn

Abstract. Educational data mining is a rapidly growing field that applies various statistical and data mining techniques to analyze educational data. This paper provides a general review of the literature on educational data mining, focusing on the methods and applications. Methods used in education data mining include classification and clustering. A classification algorithm is a supervised learning technique that seeks to categorize a given set of data objects into specified categories, build a classification model using the input data that already exists, and then apply the model to categorize new data items. The Naive Bayes, Decision Tree, Neural Network, and K-Nearest Neighbors have commonly employed classification algorithms in educational data mining. Clustering is unsupervised learning, whose objective is to divide a collection of data objects into various groups, where samples within a cluster exhibit a high level of resemblance and those between clusters are dissimilar. In educational data mining, the K-means Clustering Algorithm, Grid-Based Clustering, and Hierarchical Clustering are common clustering techniques. Those data mining algorithms are used in education such as student behavior prediction, student bad behavior detection, and student grouping. Overall, this research demonstrates that education data mining has a significant potential to improve educational programmers and student results. To solve the legal and privacy issues associated with the collecting and use of educational data, however, more research and solutions are required.

Keywords. educational data mining, classification, clustering, machine learning.

1. Introduction

Data mining is renowned for its potent ability to unearth buried information from vast amounts of data [1]. Educational data mining is to analyze particular education-related data in order to collect and process the data generated in the process of education, dig into the laws existing in education and teaching, discover the core problems, and provide suggestions and countermeasures for stakeholders in the field of education [2, 3]. Advanced information technology has a significant impact on all facets of education and teaching due to the ongoing development of educational informatization. It is continually changing the course of educational development [2]. The swift evolution of information technology and internet technology has enabled the proliferation of data mining into the educational realm, encompassing higher education institutions such as colleges and universities [4]. Consequently, a critical challenge is to investigate the relationships among educational factors within these enormous educational big data sets, identify and diagnose existing issues, and forecast development trends. Data mining technology enables the collection and processing of large amounts of complex data. Its integration with traditional education

DOI: 10.54254/2755-2721/16/20230892

businesses can continuously advance university education system wisdom and teaching mode reform. It is crucial for advancing education evaluation system reconstruction, scientific research paradigm transformation, and the realization of truly individualized learning.

There have been a large number of research results and cases in the world for educational data mining techniques and methods. Garcia et al. constructed a model with an accuracy of nearly 60% by using data mining techniques and a plain Bayesian classifier to analyze the effect of socio-demographic characteristics and academic variable characteristics on academic performance [5]. Garima Sharma and K Santosh employed the ID3 decision tree algorithm to anticipate students' ultimate grades derived from their preceding academic achievements. They predicted students' academic performance according to "low, fair, and good" with 79%, 97%, and 67% correct rates, respectively. Online learning platforms, mobile apps, and social networks have all changed the information and education sectors in recent years, offering a variety of resources and material for EDM research [6]. For example, prior to 2013, MOODLE, an online education platform, had catered to over 60 million students and educators across the world. In June 2012, there were over a billion smartphone users worldwide, while over 2.2 billion people used Facebook as their social media site. By the end of 2014, the number of registered users on Coursera exceeded 10 million. Top international conferences in data mining and machine learning have also held workshops on educational data mining [3]. For example, Knowledge Discovery and Data Mining (KDD) 2011 and Neural Information Processing Systems (NIPS) 2013 have both held workshops on educational data mining.

The choice of mining technology is one of the most central steps in the process of educational data mining [2]. Through summary and analysis, this paper concludes that the commonly used educational mining techniques are classification techniques, clustering techniques, text mining techniques, association rules, test difficulty prediction, student cognitive diagnosis, personalized recommendation, and other common techniques [2, 7, 8]. Typical applications of data mining in the education field include the Civitas Learning project, i-Ready adaptive learning system, Course Signals system, and Zhongqing Intelligent Class System [2].

The remainder of this paper will be structured as follows: Firstly, a detailed introduction of the prevalent methodologies employed in educational data mining will be presented. Subsequently, successful instances of applying data mining techniques in education will be elucidated. Finally, potential challenges and future prospects in the field of educational data mining will be addressed.

2. Methods

2.1. General process of methods

To carry out data mining in the secondary education domain, it is imperative to initially aggregate a significant volume of information generated during the learning process, including data on grades and assignments [2]. After completing the data collection, the data will be pre-processed, and the common pre-processing includes denoising, de-duplication, and text data pre-processing. After completing the pre-processing, suitable data mining methods are selected according to different data mining purposes, and common mining methods include classification, clustering, and text mining. After obtaining the mining results, it is necessary to use suitable methods to present the mining results, commonly used for visualization, including line graphs, bar charts, pie charts, network structure diagrams, etc. Finally, the data mining outcomes are analyzed to offer relevant recommendations to diverse stakeholders, including teachers, students, and administrators, to foster the advancement of education and instruction.

DOI: 10.54254/2755-2721/16/20230892

2.2. Machine learning algorithms

2.2.1. Classification. An essential method of data analysis, classification extracts models that represent significant data classes [8]. Data is divided into many classes according to the number of scans and the various features present in each data set using classification algorithms. The classification algorithm involves two steps, the first being the learning step, where a suitable classifier is generated using a predefined data or concept set, and the second being the classification step, where the accuracy of the classification rule is assessed using a known test sample set. If the accuracy is acceptable, the model is used to make predictions on samples of the class to be tested with unknown class labels. Classification algorithms commonly used in educational data mining include Naive Bayes, Decision Trees, Neural Networks, K-Nearest Neighbors, etc.

Based on Bayesian techniques with independent attributes, the Naive Bayes algorithm can be swiftly implemented using clearly defined IF-THEN rules in object-oriented programming. Thus, even novice users unacquainted with data mining can easily understand the output of these algorithms [9].

Decision trees are a prevalent prediction technique commonly employed in data mining research. This algorithm consists of a series of nodes that gather relevant information to aid the root node in making decisions.

A neural network consists of interconnected input/output units, where each connection has an assigned unique weight. During the learning phase, the network adjusts the weights to anticipate the correct class of input samples.

Classification algorithms have been widely used to detect and predict student behavior [10]. B K Francis suggests that classification techniques can enhance the quality of the higher education system by accurately predicting students' ultimate grades in a given course. The goal of the classification technique is to check participation levels and prevent students from dropping out of distance learning and online learning courses, monitor participation levels, stop students from dropping out of online learning courses, spot learners who lack motivation, evaluate learners' performance in learning activities, and forecast whether learners will finish assignments. Classification algorithms can also be used to classify learners' learning behaviors [2]. Sunar et al. classify MOOC participants into various categories to automatically categorize learners' learning habits in MOOC courses so that instructors can provide individualized learning advice.

2.2.2. Clustering. In contrast to classification algorithms, when utilizing clustering algorithms, there are no definitive criteria for categorizing the data. Therefore, it is essential to utilize machine learning methods to automatically assess the features and similarities inherent in the data to partition the data objects into distinct subsets or clusters [2, 8]. The clustering process requires that the intra-cluster data similarity be as large as possible while the inter-cluster data similarity be as small as possible [11]. Clustering aims to discover the characteristics of the data or to process the data through the obtained classes or clusters [10]. The common clustering algorithms in educational data mining are the K-means Clustering Algorithm, Grid-Based Clustering, Hierarchical Clustering, Density-Based Clustering, Model-Based Clustering, etc. [8].

The core of the K-means algorithm is to realize iterative clustering by calculating the mean value of the distance, giving the initial centroid and category K, and other steps [12]. The algorithm starts by randomly selecting k points in the dataset as centroids and then allocates each data point to the cluster corresponding to the nearest centroid based on the distance from each data point to these centroids. The K-means algorithm is widely used in the field of data mining because it can efficiently process large-scale data sets, simplify complex data structures into simple cluster structures, and facilitate analysis and application.

The data space is partitioned into a grid-like structure consisting of individual regions. Grid-based algorithms adopt a uniform mesh approach, creating a single grid that separates the entire problem domain into cells. Each cell contains data objects that are described using a combination of statistical attributes derived from the objects themselves.

DOI: 10.54254/2755-2721/16/20230892

In education, clustering enables schools to recognize students who are at risk for learning at an early stage and explores collaborative learning styles while increasing retention rates [10]. Educational decision-makers can see probable dropouts early by grouping students with comparable learning characteristics based on the content and learning qualities of the pages visited. In addition, clustering algorithms are often used to discover specific behavioral patterns of learners in the learning process [2]. For example, Linjing Wu et al. clustered learners' learning behaviors in online learning spaces and summarized four common behavioral patterns in online learning spaces, i.e., third-best student type learners, diligent learners, moderate learners, and negative learners; Rebecca et al. used clustering algorithms to model and cluster learners' participation behaviors in MOOC to uncover the typical behavioral engagement patterns in MOOC learning.

3. Application and discussion

3.1. Student behavior prediction

Student behavior prediction is advantageous for both students and educational institutions, enabling the timely identification of potential academic problems and focused interventions [9]. Ismail et al. investigated methods for forecasting academic performance in first-year computer science bachelor's degree students. They employed various techniques, such as Naive Bayes, Rule-Based, and Decision Tree, to examine student data and create a superior prediction model. The study found that the Rule Based approach achieved the highest prediction accuracy of 71.3%.

3.2. Student bad behavior detection

The purpose of this application is to use data mining techniques, such as classification, clustering, decision tree, and neural network, to detect bad student behaviors like academic failure, low motivation, and cheating [9]. These techniques aid in identifying problematic behavior early on and taking corrective measures based on predictive models generated from large amounts of data. Bayer et al. aimed to anticipate school failures and dropouts by adding extracted social behavior data from students into their analysis. Using data from discussion board conversations and emails, a social graph was formed to create new features for both the represented behavior and student data. A unique approach using cost-sensitive learning was introduced to minimize the misclassification of unsuccessful students. The study found that incorporating Social Network Analysis (SNA) significantly increased prediction accuracy to 92.89%.

3.3. Student grouping

The objective of the student grouping application is to categorize students into clusters or groups based on numerous features present in their profile information [9]. Different educational stakeholders employ this program with a range of tasks aimed at facilitating the learning process. Unlike merely sorting comparable students, clustering may involve matching students that complement one another. Additionally, when clustering students, dissimilarities between the various clusters may be maximized. Harley et al. implemented clustering techniques on data collected from 106 students to discern distinct profiles. The obtained results revealed that three clusters could be identified, and multivariate statistics (MANOVAs) were employed to validate the analysis. Statistically significant differences were found in all twelve variables used for cluster formation, demonstrating variations among the different profiles regarding the perceived prompts through the system. The overall prediction accuracy for the clusters was determined to be 78.8%.

3.4. Discussion

Compared to fields such as commerce, transportation, environment, and medical care, the field of education possesses greater uniqueness and complexity [13]. Educational data exhibits distinct features such as real-time availability, coherence, comprehensiveness, and naturalness, resulting in intricate and versatile analysis and processing requirements. As a result, the applications of educational data are more

DOI: 10.54254/2755-2721/16/20230892

diverse and profound. As information technology and artificial intelligence continue to advance, the volume of data in the education realm is increasing [2]. This data holds significant value and presents promising opportunities for educational data mining. A huge number of different sources and types of educational data is not only a valuable asset for education but also contains sensitive information about educators and learners [13]. Insufficient protection of this data can result in serious security incidents. Therefore, it is crucial for education managers to collaborate with various stakeholders to establish efficient data governance practices, enhance the quality of education data, safeguard data privacy and security, ensure rational application of education data, and promote legal sharing of education data.

4. Conclusion

This paper reviews the current state of progress in educational data mining, highlighting several successful cases of its application, and delving into various classification and clustering techniques routinely employed within the field. Based on the analysis of the application in the education field conducted in this paper, it proves that data mining has a huge impact on the education field, and prediction accuracy has reached a relatively trustworthy level. The use of educational data mining technology presents a significant challenge for stakeholders with regard to legal and efficient implementation, necessitating the development of effective strategies to address this issue. Despite this challenge, the outlook for the continued growth and development of educational data mining is optimistic.

References

[1] Mohamad S K and Tasir Z 2013 Educational data mining: A review Procedia-Social and Behavioral Sciences 97 pp 320-324
[2] He P L and Zhang Z S 2019 Educational data mining in the era of big data: methods, tools and applications (in Chinese) China Educational Technology Equipment (23) p 4
[3] Jiang B Qiu F Y and Li H J 2014 A review of educational data mining research - a technological perspective Computer and Education: Practice, Innovation, Future - Proceedings of the 16th Annual Conference of the National Society for Computer-Assisted Education
[4] Zhang Z G 2020 Research and exploration of educational data mining (in Chinese) Journal of Changchun Normal University 39(2) p 3
[5] Shen L H Tang H and Xu R 2022 A study of factors influencing academic performance based on educational data mining (in Chinese) Journal of Jiamusi University: Natural Science Edition 40(6) pp 139-144
[6] Zhou Q Mou C Yang D 2015 A review of research advances in educational data mining (in Chinese) Journal of Software 26(11) p 17
[7] Xia C F 2020 Research on the application of key technologies for educational data mining (in Chinese) Light textile industry and technology
[8] Yang Z Y Wu S Y Wang Z Q and Ding G Y 2022 Analysis of application technologies in the field of domestic educational data mining in the last decade (in Chinese) China Education Informatization
[9] Alhakami F et al 2020 Educational data mining applications and techniques International Journal of Advanced Computer Science and Applications 11(4)
[10] Yang G M Fang X J Ge B and Zhang Y 2020 Advances in educational data mining and learning analytics research (in Chinese) Journal of Mudanjiang Normal College: Natural Science Edition (3) p 5.
[11] Xu Y 2019 Research on personalized adaptive learning system based on educational data mining (in Chinese) China Education Informatization (11) p 6
[12] Faizan M Zuhairi M F Ismail S and Sultan S 2020 Applications of clustering techniques in data mining: a comparative study J Adv Comput Sci Appl 11(12)
[13] Yang X M Tang S S and Li J H 2016 Developing big data in education: connotations, values and challenges (in Chinese) Modern Distance Education Research 1 pp 50-61