Research on Educational Data Mining Based on Big Data

Xuping He, Wensheng Tang, Jia Liu, Bo Yang, and Shengchun Wang

College of Information Science and Engineering, Hunan Normal University, Changsha 410081, Hunan, People’s Republic of China
1216747124@qq.com, {tangws, yangbo}@hunnu.edu.cn, liujia@smail.hunnu.edu.cn, chuyf@qq.com

Abstract. Educational data mining (EDM) is a cross-disciplinary technology involving computer science, education, statistics, etc. It analyzes and mines education-related data to discover and solve various types of education problems. To make them better understand students and their learning environment and improve the teaching effect of teachers. Under the background of big data, EDM research will usher in a new development space. This paper first analyzes the latest research status of EDM at home and abroad, and then focuses on the progress of EDM in the context of big data in recent years. It summarizes the characteristics, shortcomings and development trends of EDM in the context of big data. Finally, it discusses the opportunities and challenges faced by EDM in the era of big data.

Keywords: Big data · Education data mining · Education big data

1 Introduction

Education data mining is a comprehensive use of mathematical statistics, machine learning and data mining techniques and methods to process and analyze education big data. Through data modeling, we can find the correlation between learners’ learning results and learning content, learning resources, teaching behavior and other factors, and predict learners’ future learning trend.

In recent years, with the development of a large number of educational applications such as online education platforms, social software, mobile phones, etc., which...
provide a large number of applications and data for the research of education data mining, the important report of “promoting education and learning through education data mining and learning analysis” issued by the U.S. government in 2012 [2] has triggered an upsurge of application research in the field of big data education. A large number of researches on education data mining and learning analysis, data-driven educational decision-making, large-scale personalized education, adaptive learning systems, and prediction-based teaching interventions have been carried out [2]. Education data mining in the context of big data has become a new research hotspot.

Since 2012, EDM research has entered a big data era of “Data driving schools, analysis reforming education” [1]. Big data provides effective improvement means for the lack of equity, accuracy, personalization, innovation and other aspects of the field of education, and has broad development space [2]. At present, there are three active research directions in this field: the first is the study of students dropping out based on the analysis of learning behavior data; the second is the application study of personalized learning services based on recommendation and adaptation algorithm; the third is analysis and prediction of learning behavior based on massive education data mining. This paper focuses on the development of EDM in the context of big data in recent years, summarizes the characteristics, shortcomings and development trend of EDM in the context of big data, and finally discusses the opportunities and challenges faced by EDM in the era of big data.

2 Research Overview

2.1 New Features

Education big data mining is a new sub direction of traditional education data mining. The traditional education data mining, because of the less sources of education data at that time, generally came from questionnaires and information management software. The mining method is relatively simple and plays a very limited role in promoting the development of education. With the advent of the information age and the development of online teaching platforms, education data sources have become very extensive: from the perspective of students, it includes life information, learning information and online second classroom information; from the perspective of teachers, it includes teaching tasks, courseware and other teaching information like paper works, scientific research data and other scientific research information; from the perspective of managers, it includes the school’s resources production information, teacher information, enrollment and employment information, etc. At the same time, with the rise of new technologies such as mobile Internet and Internet of things, more and more information is generated by teachers and students and collected automatically by devices [2]. Education has also entered a big data era.

Unlike traditional education data, education big data accurately covers all records related to education. Due to the large amount of data, diverse types, strong continuity, and low value density of educational big data [3], the traditional mining algorithms lag behind the big data analysis technology in algorithm efficiency, analysis accuracy, processing heterogeneous unbalanced data and so on, which can no longer meet the mining needs. In addition, the mining method has also changed from a single statistical analysis to the
use of visualization. Big data mining, such as clustering, regression, text mining and deep learning neural network, and has developed into education big data mining. Education big data mining can discover the essence of education problems more accurately and effectively, and then promote the development of intelligent education.

2.2 Literature Analysis

In order to understand the current research situation of education data mining, this paper searches “educational data mining” as the theme in the web of science database for the period of 2005–2019. Firstly, it makes a comparative analysis on the number of documents, and uses the function of classifying and displaying the retrieval results according to the year in the database to classify the retrieved documents, and obtains the large number of documents at home and abroad according to the number of research results each year when technology is applied to education and teaching.

Since 2005, there have been 1421 papers published by EDM, as shown in Fig. 1. Before 2008, the number of papers published by EDM research was relatively small.

Since 2008, the number of papers published has increased gradually, especially in recent years. The first International Education Data Mining Conference held in Montreal, Canada in 2008 attracted the attention of researchers. In 2012, the United States Department of Education issued a blue book “Promoting Teaching and Learning Through Education Data Mining and Learning Analysis” [4], marking that EDM has been widely concerned. Since 2015, with the advent of the era of big data, researchers have applied the new technology of EDM to online learning platforms such as MOOC and intelligent learning network, and education big data mining has developed rapidly and becomes a research hotspot.

![Fig. 1. Publications from 2005 to 2019](image)

2.3 Research Status

Research in the United States and other countries started earlier. For example, as early as 2009, some scholars pointed out in their paper [5] that big data will bring about
changes in biological research and teaching. In 2012, the U.S. government issued the important report “Promoting Teaching and Learning through Education Data Mining and Learning Analysis” [4], which triggered the upsurge of application research in the field of big data education. Many aspects emerged, such as education data mining and learning analysis, data-driven education decision-making, large-scale personalized education, adaptive learning system, and prediction-based teaching intervention research. It starts a preliminary attempt to transform from "discovery data" to "mining data", and focuses on the prediction and decision-making functions of education data mining technology.

Research in China started late, and there is a large gap in research breadth and depth compared with foreign countries. In the past 10 years, domestic research on EDM has made some progress [1, 8–13]. With the rise of the era of big data, education big data as a subset of big data has also begun to attract the attention of experts in education [8]. Xu Peng et al. [1] interpreted the 2012 report, analyzing education for change. In the era of big data, Chen Chi et al. [9] introduced big data technologies such as EDM and LA, and designed a big data model oriented to the field of online education, providing ideas for the study of big data in the field of online education. EDM has attracted unprecedented attention in 2015. Such as Zhou Qing et al. [10] mainly introduced the research results of EDM from different education environments. [11] introduced the characteristics and development process of education big data, and finally aimed at the current problems and challenges in the development of educational big data in China present six policy recommendations. Yanmei Chai et al. [12] collected relevant literature from the Web of Science database from 2008 to March 2017 for statistical and visual analysis, and personalized learning services introduced related research results. Yu Fang et al. [13] introduced the current state of EDM research in the past 10 years and designed based on the “waterfall model” in the field of software engineering. Based on the “user-centered” EDM application research framework.

At present, the research on Education Data Mining in foreign countries is in the stage of in-depth development. How to realize the discovery and prediction of teaching intelligence is the current research topic. The researchers in the field of Education Data Mining mainly include Romero of Cordoba University in Spain, Ryan Baker of Worcester Institute of Technology in the United States [5], Kalian ace of Sydney University in Australia [6], and the researches of these three researchers are the most representative. Spanish scholars Romero et al. [7], as the authority who noticed the role of education data earlier, first defined the concept of Education Data Mining (EDM). At present, the research work of scholars mainly focuses on: (1) personalized learning service application research based on recommendation and adaptation algorithm; (2) learning behavior analysis and prediction based on massive education data mining. With the rapid development of information technologies such as artificial intelligence, the realization of personalized teaching and behavior analysis and prediction that respect individual learning differences has become a new requirement for dynamic adjustment of teaching strategies in the era of big data.

3 Recent Research Progress

Under the background of education big data, the teaching process will produce massive data such as log, student behavior and teacher behavior every day. Big data technology
provides new solutions, standards and tools in terms of storage, processing and knowledge discovery, which can help the education field solve many technical problems in dealing with massive data. From the existing literature, the current research focuses on the following aspects: (1) personalized learning service under education big data; (2) the study of learning behavior mining under big data; (3) the study of students who drop out of school under big data;

3.1 Research on Personalized Learning Services

The goal of personalized learning recommendation is to correctly understand the individual differences and provide learning guidance in accordance with learners’ learning habits, knowledge mastery, resource preference, learning objectives, log recording data, etc., so that learners can better understand the learning process, promote the utilization of learning resources and improve learning efficiency and individual personality development. Education big data mining provides effective technical support for personalized learning. It uses big data technology to process and analyze massive education data, and then finds some associations and rules existing in education, selects appropriate teaching methods and contents for students’ individual characteristics, recommends corresponding learning contents and learning paths, and realizes “individualized teaching, personalized development”.

Traditional personalized learning service recommendation algorithms mainly include content-based recommendation algorithms, collaborative filtering recommendation algorithms and hybrid recommendation algorithms. Among them, collaborative filtering recommendation algorithms are one of the most widely used recommendation algorithms at present. The basic idea of collaborative filtering is to filter by “neighbor set”, which considers that there are similarities between users who choose the same project. In the personalized online learning system, the main research recommends the test questions to the students based on the user’s recommendation. First, it calculates the difficulty of the test questions. Secondly, in order to find the target user neighbor set, similarity between users is calculated. Then, the recommendation results are generated, and finally the recommendation quality is evaluated [20]. The two bottlenecks of collaborative filtering in recommendation process are “data sparsity” and “cold start”. Sparse data refers to the data in which most values are missing or zero. Sparse data is not useless data, but incomplete information. Much useful information can be mined by appropriate means. In view of the shortcomings of the algorithm, most of the strategies are to optimize the algorithm, but cannot fundamentally overcome the shortcomings of the algorithm. The traditional recommendation algorithm can solve the personalized learning recommendation problem when the data volume is small, but with the advent of information and big data era, in order to solve the recommendation problem under the background of large data volume, researchers propose intelligent recommendation based on deep neural network, which is also the hotspot of personalized learning research under the background of big data.

The main process of Intelligent Recommendation Based on deep neural network is to continuously learn the characteristics of massive data and construct a model. Neural network can automatically correct the weight bias and other parameters through continuous irrigation data to fit better learning effect and ultimately improve the accuracy
of recommendation. In terms of efficiency, the intelligent recommendation based on the deep neural network uses the trained model to process data, which greatly reduces the computation compared with the traditional algorithm. The traditional unified algorithm needs to modify the framework code to adjust the model in the process of solution, which makes the cost of improvement huge. By adjusting the parameters of the neural network, deep learning can change the model to produce results. This model is more flexible, and it is more suitable to use the principle of deep learning to solve the problem under the background of big data.

For example, Zhang Yongfu [21] and others put forward a personalized learning recommendation system based on LSTM. Long short-term memory (LSTM) recurrent neural network is a kind of recurrent neural network. LSTM has better output in sequence prediction and capturing the evolution of user taste. Compared with traditional collaborative filtering recommendation algorithms, LSTM recommendation is better in recommendation performance than traditional collaborative filtering algorithms. Yang Heng et al. [22] proposed a personalized recommendation method based on deep belief networks in the MOOC environment. By deeply mining the demographic characteristics of learners and the attribute characteristics of curriculum resources and combining the characteristics of learners’ learning behavior, a DBN based model of learners’ interest was constructed. Based on the interest model of learners, a personalized recommendation model of learners is constructed by using the method of DBN classification, which integrates the demographic characteristics of learners and the attribute characteristics of curriculum resources, and processes them into the feature vector of learners. When training the DBN classification model, the feature vector of learners and the feature vector of learners’ behaviors jointly affect the updating of model parameters. This method effectively solves the common problems of cold start and data sparsity of traditional collaborative filtering-based recommendation methods.

3.2 Research on Learning Behavior Mining

In the era of big data, the analysis of learning behavior can deeply understand the learning habits and learning characteristics of learners. According to the characteristics of students’ learning behavior, teachers can make teaching plans or divide students into learning groups with complementary learning styles to improve learning efficiency. The current research is mainly carried out in the following two aspects: (1) the research on the potential behavior patterns of students based on the optimized association algorithm; (2) the research on the learning behavior mining based on the social network analysis algorithm.

Based on Association Rules Mining Algorithm

Since the education informatization was put forward, the data in the field of education has been growing, and mining technology for education data has attracted more and more attention and has become a new hotspot of data mining research. Most of the related management systems in colleges and universities only stay in the simple operation stage of the original data storage, but the relationship and influence between the courses cannot be determined. Therefore, it is very necessary to use association rule analysis to go deep into the data surface, carry out correlation mining analysis, and find out the
correlation between students’ data. Association rule analysis is one of the most active research methods in data mining. Its purpose is to find out the association relationship among various items in a data set. Through association rule mining, we can get useful information with potential value hidden in massive student data, and discover the behavior patterns and association relationships of students. The most famous association rule algorithm is the Apriori algorithm proposed by Agarwal. Most of the mining algorithms are based on the Apriori algorithm. However, with the advent of the era of big data, the Apriori algorithm is facing challenges in both time efficiency and space scalability. Therefore, researchers proposed an improved algorithm for massive data processing to mine its behavior patterns more efficiently.

Lu Xinyuan et al. [23] proposed a frequent and efficient association rule mining algorithm based on domain knowledge for mining the association relationship between the academic performance of middle school students in educational administration data. At present, most association rule mining algorithms mainly focus on frequent itemset mining, but only mining frequent itemset cannot meet people’s requirements for efficient use of results, and in the practical application field, they do not consider the rich domain knowledge closely related to the data itself, resulting in a large number of redundant rules. These rules have been well known by the industry, so that the result of the fruit is not very interesting. Based on the sequential pattern mining algorithm, the Fui-DK algorithm generates frequent candidate sets. Based on the support and confidence of the classical association rule algorithm, the two parameters of utility degree and interest degree are added to get the efficient set of interesting terms. Then, the support, confidence degree and utility degree of the Association rules that meet the conditions are sequenced and output. Finally, the efficient set of interesting terms obtains Interesting association rule results. The experimental results show that the algorithm can reduce the computing time, and the elimination rate of the known association rules in the field can reach 43%, which can help colleges and universities to carry out time-saving and effective education data mining. Luna [24] et al. proposed an evolutionary algorithm to discover association rules of rare categories in the learning management system, which is used to mine the association of students’ learning behaviors on the Moodle platform, and compared the algorithm with other five association rule algorithms. Geigle et al. [25] added a layer of HMM based on hidden Markov model (HMM) to form TL-HMM unsupervised learning of many student behavior observation sequences to discover potential student behavior patterns.

Based on Social Network Analysis Algorithms
Rabbany et al. [26] use social network analysis algorithms to evaluate the participation of students in the forum in the course management system, such as tracking the subjects replied by students, the number of posts published, etc., so that teachers can quickly understand the hot topics discussed by students. Mohammed Saqr [27] et al. proposed that visual analysis of social network analysis algorithms and quantitative network analysis (concentration measurement) should be used together to analyze the position and role of students in collaborative knowledge sharing, to monitor online collaborative learning, find gaps and traps in application, guide the potential of informed intervention, and design relevant data-driven stem from the information obtained through monitoring
Pre measures and using experimental, observational, repeated measurement designs to evaluate its effectiveness to promote teaching and learning.

### 3.3 Research on Student Drop-Out

Dropout is not only considered as a serious educational problem, but also a serious social problem. In addition to the high risk of unemployment or underemployment, dropout students are more likely to suffer from mental health problems, such as depression, gang participation or other criminal activities. In the era of big data, there have been new achievements and new methods in the studies of dropout. By using the technology of education big data mining, students with high dropout risk can be identified efficiently and accurately, and factors related to the risk of dropout faced by students can be analyzed. The current research is mainly divided into two categories: one is online education dropout research; the other is traditional classroom dropout research. In recent years, the main literature of student dropout research is shown in Table 1.

#### Research on Online Teaching Platform

With the popularization of Internet education platforms, many online education platforms such as MOOC and intelligent learning networks continue to emerge, and many excellent teachers in colleges and universities in China have also opened excellent MOOC courses, which allow students in other colleges and universities to enter the classroom. However, according to the research, the high dropout rate of online students both at home and abroad has become more and more prominent, and the high dropout rate in online education platforms has also attracted the attention of many researchers.

Online dropout prediction research predicts whether students will insist on learning or drop out in the next week by analyzing the data of students’ learning activity logs on the platform. Its research significance lies in its use as an intervention by creators of online courses and education researchers, so that researchers can gain insights into the reasons why students drop out, and predict various signals of dropout for online education platforms to create customized intervention strategies for learners. In the early stage, the research on the reasons for MOOC students’ dropping out was based on statistical analysis. Generally, the reasons for dropping out were analyzed by filling in the after-school questionnaire survey and artificial statistical analysis data. The personalized analysis and intervention on the reasons for students’ dropping out could not be carried out in depth, which made the accuracy of prediction results poor. With the increasingly prominent advantages of machine learning methods in big data processing, the application of machine learning algorithms in MOOC dropout prediction tasks has been promoted in the context of education big data.

The general process used in the classic algorithm of dropout prediction model can be divided into three steps: (1) Data Preprocessing; feature selection of dropout data, screening out significant features and discarding non-significant features; (2) Dropout model training and tuning; selecting model according to the actual situation of dropout data and specific problems to be solved, such as sample number, feature dimension, and comprehensive consideration of data characteristics. In the optimization problem, we use cross validation to observe the loss curve, test result curve and other analysis reasons. The adjustment parameters are: optimizer, learning rate, batch size, etc.; (3S)
Research on Educational Data Mining Based on Big Data

Model effect evaluation; prediction and evaluation with test set, comparative study on accuracy rate, recall rate, F-score and other evaluation indicators. The flow chart is shown in Fig. 2.

In recent years, research of MOOC dropout prediction mainly focuses on two shortcomings of classical machine learning algorithms. The first one is the improvement of data preprocessing steps, mainly including the processing of feature rule making and the processing of extremely unbalanced data. The second is the improvement of the prediction model, including model fusion to make up for the shortcomings of the model and the new model.

In order to improve the data preprocessing, the feature rules are formulated based on classic machine learning methods. This part of the work requires a lot of manual operations and carrying out a variety of complex feature extraction operations manually. When there are many complex linguistic phenomena in the text, the process of feature rule making becomes very difficult. Sun Xia et al. [14] proposed a use of convolutional neural networks to automatically extract useful features from student behavior data in order to solve the problem of manual feature extraction in the past. CNN learned the changes of each category from many data to realize the stable classification of feature changes in the same category. Wang Xiyu et al. [16], based on the data of eight MOOC courses with the largest number of students selected from the dream course platform of National University of science and technology of national defense, extracted three dimensions from the course factors, learners’ own factors and other personnel factors. A total of more than 40 learning data were used to study the prediction of dropouts, and the most helpful behavior data of each course was analyzed. In the extremely unbalanced data processing, because the dropout students belong to a few categories, the data set is unbalanced. Usually, most students continue to study, and only a few students drop out. In this case, the accuracy may be misleading, because most of the default classifiers will get high accuracy, while a few are easy to be ignored. Therefore, in the process of data processing, the processing of data imbalance is particularly important. It is necessary to design a specific algorithm that can focus on a few categories, which is conducive to the improvement of prediction accuracy. In the case of education data, the research activity is focused on resampling algorithms and cost sensitive algorithms. Data resampling modifies the training data set by adding instances belonging to a few classes to generate a more balanced distribution of data classes. Smote is a resampling method
that has been shown to improve the classification of imbalances, especially when used in combination with C45 and SVM [19]. On the other hand, the cost sensitive learning algorithms consider the classification errors in other categories. Cost sensitive learning allocates a higher misclassification cost to a few classes of samples, and a smaller misclassification cost to most classes of samples. In this way, cost sensitive learning improves the importance of samples of a few categories artificially in the process of training the learner, thereby reducing the preference of the classifier for most classes. Chaplet [15] et al. put forward a cost sensitive learning algorithm for data imbalance to punish the given class for error classification and adjust the weight of a few classes of data, so that the algorithm has better accuracy and better false negative rate. This method is superior to the previous algorithm in the kappa value of Cohen.

In the improvement of the prediction model, the traditional machine learning method assumes that the probability of students dropping out of school in different time steps is independent, which is inconsistent with the actual situation in the scene, because the state of students at a certain time will be affected by the state of the previous time, making the accuracy of the prediction result poor; sun Xia et al. [14] can’t correlate the study of the state of learning in different time periods.

**Research on Traditional Classrooms**

In traditional education, teachers and school administrators have spent a lot of time and energy to reduce dropout, but until now, it still exists in schools. How to efficiently and accurately identify at-risk students in many students, put forward personalized intervention programs in time, and reduce dropout rate is a major problem faced by schools. With the in-depth application of big data mining technology in the field of education, researchers use massive learning behavior data for data mining analysis to identify students with a particularly high risk of dropping out. It can effectively and accurately identify students with high dropout risk, and analyze which factors are related to students’ DROPOUT risk. Teachers can choose targeted methods to intervene, which can greatly improve teachers’ work efficiency and personalized care for students.

In recent years, the research on the traditional dropout field mainly includes the algorithm improvement for the mining of interpretable classification rules and the model evaluation algorithm for the dropout field. Xing et al. [17] took k-nearest neighbor, support vector machine and decision tree as the benchmark algorithm, and then proposed a drop out prediction model based on deep learning compared with the benchmark algorithm, and got higher drop out prediction accuracy. Márues et al. [18] proposed a new icrm2 algorithm based on the mining of interpretable classification rules, in order to obtain more accurate and shorter classification rules than other existing algorithms, to obtain better performance, with the classification accuracy as high as 91%. Lakkaraju [19] et al. put forward an evaluation algorithm for dropout prediction, using evaluation algorithm to evaluate the results of SVM, decision tree, random forest, logistic regression and other classification models. The evaluation algorithm has a strict qualitative and quantitative comparison for the classification model, changing the situation that only general evaluation indicators can be used, and the evaluation of dropout field algorithm is very important great significance.
Table 1. Summary of student dropout studies under big data

| No.      | The main method                                                                 | Results                                                                 |
|----------|---------------------------------------------------------------------------------|------------------------------------------------------------------------|
| Online teaching platform | Literature [14] Combining CNN and LSTM sensitive learning algorithms | F1 value is 0.948 higher than other algorithms                          |
|         | Literature [15] Emotion Analysis Scoring Algorithm Improved neural networks       | Kappa in Cohen beats the previous algorithm with a value of 0.432       |
|         | Literature [16] Multi-linear regression model BP neural network model            | Forecast accuracy of more than 90%                                    |
| Traditional classrooms | Literature [17] Comparison between drop-out prediction model and benchmark algorithm based on deep learning | Higher accuracy than the baseline algorithm                           |
|         | Literature [18] An ICRM2 algorithm is proposed                                   | Up to 91% of classification accuracy                                  |
|         | Literature [19] An evaluation algorithm specifically for drop-out prediction    | Changed the situation where only common assessment indicators can be used |

Model Experiment and Analysis

Experimental Settings

(1) Problem setting: predict the dropout behavior of Xuetang Online, one of the largest MOOC platforms in China. Through the interpretation of the data set, based on the user’s previous behavior, he predicts whether he will skip class in the next 10 days. The defining criterion is that a registration number has no log record within 10 days after a certain point in time.

(2) Data set: The data set used in the experiment is a publicly available data set provided by KDD Cup2015. This dataset contains 120,542 registered activity logs from 79,186 students across 39 courses, and each course can take up to 5 weeks. There are two main sources of learner behavior record information: browsers and servers, including seven events: objects visited, discussions, navigation courses, page closures, attempts to solve problems, watching videos, and browsing the wiki. Each student’s course behavior includes watching videos, trying to solve problems, participating in the course.

(3) Evaluation indicators: Accuracy, recall and F1 values commonly used in dropouts are used as evaluation criteria. TP indicates the number of dropouts correctly predicted, FP indicates that the average student was incorrectly predicted as the number of dropouts, and FN indicates the number of students that were predicted to be actually dropouts. Precision, recall and F1 values are calculated by formulas (1), (2)
and (3).

\[
precision = \frac{TP}{TP + FP} \tag{1}
\]

\[
recall = \frac{TP}{TP + FN} \tag{2}
\]

\[
F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{3}
\]

(4) Experimental model: A comparison analysis between traditional classification methods (LR (logistic_regression), DT (decision_tree), SVM (Support Vector Machine)) and deep learning methods (GBDT (GradientBoostingDecisionTree), MLP (Multilayer Perceptron)).

Experimental Results and Analysis

Table 2. Comparison of the five dropout prediction models

|       | LR  | DT  | SVM | GBDT | MLP |
|-------|-----|-----|-----|------|-----|
| Accuracy | 0.821 | 0.798 | 0.661 | 0.813 | **0.873** |
| Recall    | 0.833 | 0.815 | 0.712 | 0.843 | **0.892** |
| F1        | 0.827 | 0.806 | 0.686 | 0.828 | **0.882** |

Among these prediction models, the performance of SVM is slightly worse, because the amount of data used in the experiments in this paper is large, SVM is not suitable for large sample data mining, and it is better for small sample SVM classification. When the amount of data is large, MLP performs best as a method of neural networks, which shows that neural networks have superior performance in educational big data mining and are more in line with the current needs of data mining in the context of big data (Table 2).

The overall prediction accuracy in the experiment is low. After analysis, the existing research and methods have the following deficiencies: (1) the use of big data cannot be used to improve the prediction accuracy, and the use of fine-grained features of student behavior records is insufficient (2) neural network methods The prediction result can explain too poor (3) Cannot solve the common cold start problem in practical application scenarios.

4 Summary and Outlook

After decades of development, EDM has attracted more and more researchers’ attention, especially in recent years, online education platform, social software, mobile phones and so on provide many applications and data for the research of education data mining. In
recent years, EDM research has made great progress, but there are still some deficiencies, mainly reflected in four aspects: first, there is little research on data preprocessing technology in the data preparation stage, only a few scholars put forward some improvement methods. Second, in the process of building the model, deep learning technology cannot make full use of big data to improve the prediction accuracy, and cannot make full use of the fine-grained characteristics of student behavior records and the poor interpretability of the model. Third, in the new teaching environment, there is less research on the mining of teachers’ educational process, which is only limited to the mining and prediction of students’ learning behavior. However, teachers also play an important role in the process of education. Fourth, there is a lack of open data set in the aspect of information sharing, in which the disclosure of students’ information may have problems such as privacy and security, and the lack of data sharing It hinders the research of education data mining, which leads to each research aiming at the specific teaching scenes collected by itself, and fails to conduct research in more and wider teaching scenes. The coming of education big data era brings not only opportunities for the development of EDM, but also challenges in technology and management. In the future, it will be the research trend of EDM in the era of big data of education in the future to integrate the relevant knowledge of education field and the experience of teachers’ teaching process into the established model. We expect it to be more mature and prosperous.

References

1. Xu, P., Wang, Y., Liu, Y.: Analysis of learning change from the perspective of big data: interpretation and enlightenment of the report “promoting teaching and learning through education data mining and learning analysis” in the United States. J. Distance Educ. 11 (2013)
2. Sun, Q., Lu, C.: Research on the application of big data in colleges and universities. China Educ. Netw. 63–65 (2014)
3. Chen, D., Zhan, Y., Yang, B.: Application analysis of deep learning technology in education big data mining. Audio Vis. Educ. Res. 40(02), 70–78 (2019)
4. Bienkowski, M., Feng, M., Means, B.: Enhancing teaching and learning through educational data mining and learning analytics: an issue brief. Technical report. Office of Educational Technology, U.S. Department of Education, Washington, pp. 1–57 (2012)
5. Baker, R.S.J.D., Yacef, K.: The state of educational data mining in 2009: a review and future visions. J. Educ. Data Min. 1(1), 3–17(2009)
6. Merceron, A., Yacef, K.: Educational data mining: a case study. In: Conference on Artificial Intelligence in Education: Supporting Learning Through Intelligent & Socially Informed Technology (2005)
7. Romero, C., Ventura, S.: Educational data mining: a survey from 1995 to 2005. Expert Syst. Appl. 33(1), 135–146 (2007)
8. Liu, F.: A review of the application of big data in education. Mod. Educ. Technol. 24(8), 13–19 (2014)
9. Chen, C., Wang, Y., Li, C.: Research and application of big data for online education. Comput. Res. Dev. 67–74 (2014)
10. Zhou, Q., Mou, C., Yang, D.: Summary of research progress in education data mining. J. Softw. 26(11), 3026–3042 (2015)
11. Yang, X., Wang, D.D.: Application mode and policy suggestions of education big data. Res. Audio Vis. Educ. 54–61 (2015)
12. Chai, Y., Lei, C.: Overview of online learning behavior research based on data mining technology. Comput. Appl. Res. 1287–1293 (2018)
13. Yu, F., Liu, Y.: “User centered” education data mining application research. Audio Vis. Educ. Res. 39(11), 69–77 (2018)
14. Sun, X., Wu, N., Zhang, L.: Prediction method of MOOCS dropout rate based on deep learning. Comput. Eng. Sci. 893–899 (2019)
15. Chaplot, D., Rhim, E., Kim, J.: Predicting student attrition in MOOCs using sentiment analysis and neural networks
16. Wang, X., Zou, G., Li, X.: Prediction of learners’ dropping out of class based on MOOC data. Mod. Educ. Technol. 95–101 (2017)
17. Xing, W., Du, D.: Dropout prediction in MOOCs: using deep learning for personalized intervention. J. Educ. Comput. Res. (2018)
18. Márquez-Vera, C., Cano, A., Romero, C.: Early dropout prediction using data mining: a case study with high school students. Expert Syst. 33(1), 107–124 (2016)
19. Lakkaraju, H., Aguier, E., Shan, C.: A machine learning framework to identify students at risk of adverse academic outcomes. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1909–1918. ACM (2015)
20. Wu, J.: Literature review on collaborative filtering recommendation algorithm. Shang, p. 224 (2016)
21. Zhang, Y.: Research and design of personalized learning recommendation system based on LSTM (2018)
22. Yang, H.: Research on personalized recommendation method based on deep belief network in MOOC environment. Central China Normal University (2018)
23. Lu, X., Wang, X.: Mining association rules of academic data based on domain association redundancy. Comput. Sci. 427–430 (2019)
24. Luna, J.M., Romero, C., Romero, J.R., Ventura, S.: An evolutionary algorithm for the discovery of rare class association rules in learning management systems. Appl. Intell. 42(3), 501–513 (2014)
25. Geigle, C., Zhai, C.X.: Modeling MOOC student behavior with two-layer hidden Markov models. In: Fourth ACM Conference on Learning (2017)
26. Rabbany, R., Elatia, S., Takaffoli, M., Zaïane, O.R.: Collaborative learning of students in online discussion forums: a social network analysis perspective. In: Peña-Ayala, A. (ed.) Educational Data Mining. SCI, vol. 524, pp. 441–466. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-02738-8_16
27. Saqr, M., Fors, U., Tedre, M., Nouri, J.: How Social Network Analysis Can Be Used to Monitor Online Collaborative Learning and Guide an Informed Intervention (2018)