Chapter 14
An Overview of Data Mining Techniques for Student Performance Prediction

Xiu Zhang and Xin Zhang

Abstract In order to better understand and optimize the learning process and learning environment, educational data mining technology is becoming more and more important in processing a large number of educational data. Through the analysis of large amounts of educational data, students' academic performance is predicted, identifying a "high risk" of dropping out and predicting their future achievement, e.g., on final exams. The predicted results can provide early warning for students' own learning, and provide suggestions for educators to allocate educational resources more reasonably and improve the teaching mode. The purpose of this chapter is to comprehensively introduce the more advanced supervised machine learning technology, different educational resource dataset, and the latest research results.

Keywords Educational data mining · Machine learning · Student performance prediction · Survey

14.1 Introduction

With the development of the network technology, a large amount of data information have been collected which cover a wide range of fields, such as business data, education data, agriculture data, military data and so on. In order to explore the meaning behind the data, the data mining technology has received more and more attention from all over the world. In the field of education, especially, it has become the research hotspot to predict the student performance using data mining technology [1, 2].

In the educational data mining (EDM) work, the result of student achievement evaluation is often one of the important indicators to evaluate students' development potential, development level and performance. Entity and efficient educators often collect all aspects of information from students through research and other methods,

X. Zhang · X. Zhang (✉)
Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin, China
e-mail: ecemark@tjnu.edu.cn

W. Wang et al. (eds.), *Artificial Intelligence in Education and Teaching Assessment*,
https://doi.org/10.1007/978-981-16-6502-8_14

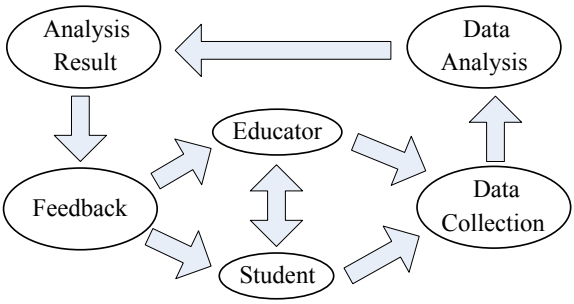

Fig. 14.1 The relationship between the students, educators and EDM

and organize them into relevant documents. With these documents as the data support [3], so as to dig deeply the value behind these data, for the teaching administrators to provide a basis, humanized, directional guidance. Teaching quality is an important factor to measure the grade of a school, while students' academic level is the main index to evaluate teachers' teaching effect and students' learning quality [4, 5].

The work of student's performance prediction using data mining technology has extensive instruction for educational work as shown in Fig. 14.1 [6, 7]. For students, it can help them to understand their learning efficiency and learning progress, so as to know more about their learning abilities. For teachers, it can help them to master the progress of teaching, and adjust their teaching schedule according to the predicted results. For educational administrator, it can provide decision support, improve the management system, and allocate educational resources scientifically.

The subjects most closely associated with EDM are computer science, education, and statistics. The workflow of educational data mining includes data collection, data preprocessing, data analysis and prediction as shown in Fig. 14.2. Data collection includes questionnaire survey, online course data acquisition, offline examination information collection and so on. The purpose of data preprocessing mainly focused on the following aspects: (i) Remove useless information from the data; (ii) Transform unstructured data into structured data; (iii) Split and merge the attributes. The most important step in EDM is data analysis. The technologies include statistic analysis such as descriptive statistics analysis and inferential statistics analysis, cluster analysis such as K-means cluster method, performance prediction approaches such as similarity-based methods, model-based methods and probabilistic method [1–5].

The chapter is organized as follows. Section 14.2 provides an overview of the dataset. Section 14.3 describes data mining technology. Section 14.4 summarizes the research results. Section 14.5 gives the conclusion.

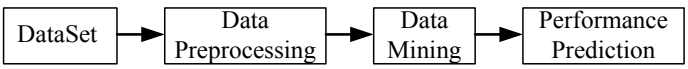

Fig. 14.2 Flow chart of EDM

Table 14.1 Dataset information freely available

Authors	Number of instances	Number of attributes	Associated tasks	Country	Date
Hussain [1]	300	22	Classification	India	2018
Hussain [2]	666	11	Classification	India	2018
Gunduz and Fokoue [3]	5820	33	Classification, Clustering	Turkey	2013
Vurkac [4]	10,800	20	Classification	America	2011
Vahdat et al. [5, 6]	230,318	13	Classification, Clustering, Regression	Italy	2015
Petkovic et al. [7]	74	102	Classification	America	2017
Kuzilek et al. [8]	32,593	12	Classification	Czech	2017

14.2 Overview of the Dataset

At present, because of the rise of online education, a large amount of education data has been produced. The collection of educational data includes the traditional questionnaire survey, the information stored in the educational administration system of each school, the data collection set by the teaching unit according to the actual situation, and the data collection in the online education system.

In this part, we list several publicly available data sets since 2010 that are available for download online as shown in Table 14.1.

14.3 Data Mining Method in Performance Predication

14.3.1 Data Mining Tools

In this part, we will give an overview of several commonly used data mining tools all over the world.

(1) Rapid Miner [9]
Rapid Miner is an environment for machine learning and data mining experiments which is applied in research and practical data mining tasks. This tool is developed in Java programming language and provides high-level analysis through a template-based framework.

It has rich data mining analysis and algorithm functions. The biggest advantage of the tool is that it doesn't require the user to write code. It already has many templates and other tools that make it easy to analyze the data.

(2) KNIME [10]
KNIME is a user-friendly, understandable and comprehensive open source for data integration, processing, analysis and exploration platform. It has a graphical user interface to help users easily connect nodes for data processing.

It is easy to integrate with third-party Big Data frameworks, such as Apache Hadoop and Spark, through the Big Data Extension. It is Compatible with multiple data formats, including plain text, database, document, image, network, and even Hadoop-based data formats. Meanwhile, it is compatible with multiple data analysis tools and languages including R and Python language support for scripts, so that the experts can use powerful visualization function to provide an easy-to-use graphical interface, which can show the analysis results to users through vivid graphics.

(3) Smartbi [11]
Smartbi Mining is a professional data Mining platform that provides predictive capabilities to businesses. This platform is integrated with rich algorithms and supports 5 categories of mature machine learning algorithms including classification, regression, clustering, prediction, correlation algorithms. In addition to providing the main algorithm and visual modeling functions, SmartBi Mining also provides essential data preprocessing functions. In general, this platform is easy to learn and use.

(4) TANAGRA [12]
TANAGRA is a data mining software for academic and research purposes. The software has exploratory data analysis, statistical analysis, machine learning. TANAGRA contains some supervised learning, but also includes other paradigms such as clustering, factor analysis, parametric and nonparametric statistics, relevant rules, feature selection, and building algorithms.

(5) Orange [13]
Orange is a suite of component-based data mining and machine learning software written in Python. It is an open source for data visualization and analysis. Data mining can be done through visual programming or Python scripts. It can be visualized using scenarios, bar charts, trees, networks, and heat maps.

(6) Weka [14]
Weka (Waikato Environment for Knowledge Analysis) is the best known open source machine learning and data mining software. It can invoke the analysis component including data preparation, classification, regression, clustering, association rules mining, and visualization through Java programming and the command line.

(7) Scikit-learn [15]
Scikit-Learn is a simple and efficient data mining and data analysis tool. It's a machine learning library in Python, built on top of Numpy, Scipy, and Matplotlib, and it's also open source. Its characteristics include classification, regression, clustering, dimensionality reduction, model selection and preprocessing.

14.3.2 Performance Prediction Approaches

In the educational data mining, classification and regression are commonly used to predict the student's performance. In the following, the main methods are briefly introduced and discussed.

(1) Decision Tree (DT) [16]
Decision tree is a basic classification and regression method, which makes decisions based on tree structure and can be considered as the set of if–then rules. Generally, a decision tree contains a root node, several internal nodes and several leaf nodes. The root node contains all the sample points, the internal node serves as the partition node (attribute test), and the leaf node corresponds to the decision result. The advantages of the algorithm are low computational complexity, easy to understand the output results, insensitivity to the absence of intermediate values, and the ability to process irrelevant feature data. The downside is that it can cause overmatching problems.

For the decision tree construction based on ID3 algorithm, the criterion of feature selection is information gain. ID3 algorithm originated from concept learning system (CLS). C4.5 algorithm is a kind of classification decision tree algorithm, whose core algorithm is ID3 algorithm. C4.5 algorithm uses information gain rate to select feature, which overcomes the shortcoming of choosing feature with more values when using information gain to select feature. However, the disadvantage is that in the process of constructing the tree, the data set needs to be scanned and sorted for many times, which leads to the low efficiency of the algorithm. C4.5 algorithm was developed in Java in Weka as J48.

(2) Naïve Bayes (NB) [17]
Naive Bayes model (NBM) originated from classical mathematical theory, which has a solid mathematical foundation and stable classification efficiency. At the same time, the NBC model requires few parameters to estimate and is not sensitive to missing data, and the algorithm is relatively simple. In theory, the NBC model has the smallest error rate compared with other classification methods. However, in fact, this is not always the case, because the NBC model assumes that the attributes are independent of each other, which is often not valid in practical application, which has a certain impact on the correct classification of the NBC model. When the number of generics is large or the correlation between attributes is large, the classification efficiency of NBC model is inferior to that of decision tree model. When the attribute correlation is small, the performance of NBC model is the best.

(3) Support vector machines (SVM) [18]
SVM is a kind of supervised learning method, which is widely used in statistical classification and regression analysis. The support vector machine maps the vector into a higher dimensional space and establishes a hyperplane with maximum spacing in this space. Two parallel hyperplanes are built on both

sides of the hyperplanes separating the data. Separating hyperplanes maximizes the distance between two parallel hyperplanes. The larger the distance or gap between the pseudo-definite parallel hyperplanes, the smaller the total error of the classifier.

The advantages of the SVM are low generalization error rate and low computational overhead. The disadvantage is sensitive to parameter adjustment and kernel function selection.

(4) K-Nearest Neighbor (KNN) [19]

KNN classification algorithm is a relatively mature method in theory and one of the simplest machine learning algorithms. The idea of this method is that if most of the K most similar samples in the feature space of a sample belong to a certain category, then the sample also belongs to this category.

The advantages of KNN are high accuracy, insensitivity to outliers and assumption of no data input. The disadvantages are high computational complexity, and high space complexity.

(5) Random Forest (RF) [29]

Random forest is composed of many decision trees, and there is no correlation between different decision trees. When we carry out the classification task, new input samples come in, and each decision tree in the forest will be judged and classified separately. Each decision tree will get its own classification result. Which one of the classification results of the decision tree has the most classification will be regarded as the final result by the random forest.

The advantages of RF are that it can use very high dimensional data, and don't have to reduce dimensions and do feature selection. The disadvantage of RF has been shown to overfit for some noisy classification or regression problems.

(6) Artificial Neural Network (ANN) [31]

ANN can simulate the activity of neurons by mathematical model, which is an information processing system based on the structure and function of the Neural Network of the brain. The multi-layer forward neuron network (also called multi-layer perceptron, MLP) proposed by Minsley and Papert is the most commonly used network structure at present.

Compared with traditional data processing methods, neural network technology has obvious advantages in processing fuzzy data, random data and nonlinear data, and is especially suitable for systems with large scale, complex structure and unclear information.

(7) Classification and Regression Tree (CART) [20]

CART algorithm is a binary recursive segmentation technology. The current sample is divided into two sub-samples, so that each non-leaf node generated has two branches. Therefore, the decision tree generated by CART algorithm is a binary tree with simple structure.

14.4 Results and Discussions

In order to study the influence of different attributes on students' performance and to mine the meaning behind the data, different researchers have studied different attributes and analyzed their importance in students' performance prediction as shown in Table 14.2.

Before the classification algorithms applied to analyze the data, the feature selection approach was used to select 12 highly influential attributes from 24 attributes [1]. The results showed that it can greatly improve the accuracy of predictions. The researchers [7] collected the data about student team project activities. It can predict the student teams' performance.

The researcher [21] studied the effect of student background and social activities on the student's performance. It came to a conclusion that the student background and social activities had significant to the student's performance prediction in the binary classification. Different from other researches on academic prediction after the end of the course, the researchers [22, 23] studied the prediction of students' academic performance while the course is in progress, so as to give early warning to students and provide suggestions to teachers. In addition, Kahraman et al [24] developed an Intuitive Knowledge Classifier to analyze the web-based adaptive learning environment. It can greatly improve the accuracy of the classification. The authors [25–27] use data collected in a traditional teaching setting to learn how to predict students' academic performance in early stage. Among these, the authors [27] considered the role of students' self-assessment in the performance prediction.

The above researches focused on analyzing the effect of the student information on the performance prediction. Khan et al. [28] studied the impact of teaching on the student's performance. It indicated that teaching had a positive impact on the

Table 14.2 Attributes affect the performance

Authors	Attributes affect the performance
Kiu [21]	Student background, student social activities and student coursework result
Hu et al. [22]	Time-dependent variables
Huang et al. [23]	Student's cumulative GPA, grades earned in four pre-requisite courses and scores on three dynamics mid-term exams
Hussain et al. [1]	The 12 high influential attributes were selected among 24 attributes
Petkovic et al. [7]	Team Activity Measures
Kahraman et al. [24]	Web-based adaptive learning environments in different domains
Carter et al. [25]	Affirms the importance of social interaction in the learning process
Yu et al. [27]	Self-evaluation comments can play an important role in improving the accuracy of early-stage predictions
Khan et al. [28]	Teaching
Liu et al. [33]	Historical learning records, learning target and prerequisite graph

student's performance. The researchers [30] developed a performance prediction models with less information for predicting at-risk students. The results indicate that the subject which relied on knowledge of other subjects in the program generally performed better than those which relied less on previous subjects. Lee et al. [32] investigated the course dropout in a mobile learning environment. The researchers in [33] proposed a Cognitive Structure Enhanced framework for Adaptive Learning which combined knowledge levels of learners or knowledge structure of learning. The framework can dynamically provide the suggestions and guidance for the next learning during the whole learning process.

At present, there are many data mining techniques that can be used to predict students' academic performance. We list and summarize the classification rates of the current commonly used algorithms as shown in Table 14.3. There are other algorithms can be used to predict the students' performance. For example, Bendangnuksung et al. [31] proposed the Deep Neural Network (DNN) model to analyze the students' performance. The results indicated that DNN outperformed other algorithms (DT, NB, ANN) in accuracy.

As shown in Table 14.3, the same algorithm has different classification accuracy in different dataset. However, most of the algorithms have high accuracy for binary classification type. As shown in Fig. 14.3, the classification accuracy for multi-class classification problem is relatively low.

It can be seen from Fig. 14.3 that RF and ANN attains better performance compared with NB, DT and SVM. The minimum values of classification accuracy of the five methods have small gap compared with the maximum classification accuracy values.

Table 14.3 Classification accuracy

Authors	Type of classification	EDM accuracy (%)						
		NB	DT	RF	ANN	CART	SVM	KNN
Kiu [21]	Binary	88.9	92.4	89.4	86.3	–	–	–
	5-level	71	79.1	74.9	68.7	–	–	–
Hu et al. [22]	Binary	–	93.4	–	–	95	–	–
Huang et al. [23]	Binary	–	–	–	88.5	–	86.5	–
Hussain et al. [1]	5-level	65.3	73	99	–	–	–	–
Hussain et al. [2]	4-level	57.8	64.7	–	90.8	–	–	–
Petkovic et al. [7]	Binary	–	–	70	–	–	–	–
Kahraman et al. [24]	4-level	73.8	–	–	–	–	–	85
Bucos et al. [25]	Binary	–	82	84	–	–	84	–
Yu et al. [27]	Binary	–	–	–	–	–	74	–
Ahmed et al. [29]	Binary	–	–	83.6	–	–	–	–
Chanlekha et al. [30]	9-level	57.5	65	62.5	62.5	–	57.5	–
Bendangnuksung et al. [31]	5-level	80	82.2	–	80	–	–	–

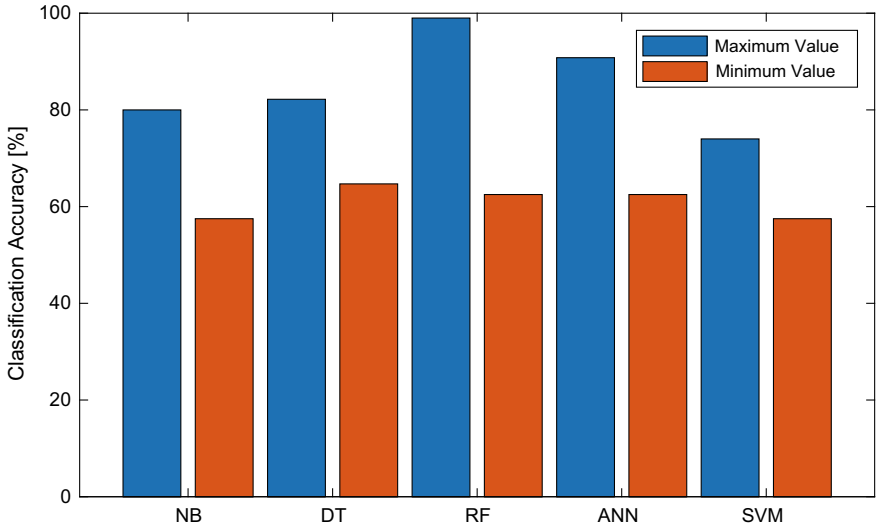

Fig. 14.3 The classification accuracy of different algorithms for multi-class classification problems

14.5 Summary

The rapid development of data mining technology has promoted the extensive application of educational data analysis. By mining the effective information behind the educational data and predicting the academic performance of students, it can not only help students understand their own learning state, but also help educators to specify corresponding strategies to improve the efficiency of education.

In the future, with the continuous progress of science and technology, online education will become more and more popular. A large number of online education data will provide more materials for data mining, and how to make better use of online and offline education data will provide better guidance for education.

References

1. S. Hussain, N.A. Dahan, F.M. Baalwi, N. Ribata, Educational data mining and analysis of students' academic performance using WEKA. Indonesian J. Electr. Eng. Comput. Sci. **9**(31), 447–459 (2018)
2. S. Hussain, R. Atallah, A. Kamsin, J. Hazarika, Classification, clustering and association rule mining in educational datasets using data mining tools: a case study, in *Cybernetics and Algorithms in Intelligent Systems. CSOC2018 2018. Advances in Intelligent Systems and Computing, AISC 765*, ed. by R. Silhavy (Springer, 2019), pp. 196–211
3. G. Gunduz, E. Fokoue, *UCI Machine Learning Repository* (University of California, School of Information and Computer Science, Irvine, CA, 2013)

4. M. Vurkac, Clave-direction analysis: a new arena for educational and creative applications of music technology. J. Music, Technol. Educ. **4**(1), 27–46 (2011)
5. M. Vahdat, L. Oneto, A. Ghio, G. Donzellini, D. Anguita, M. Funk, M. Rauterberg, A learning analytics methodology to profile students behavior and explore interactions with a digital electronics simulator, in *EC-TEL 2014. LNCS, 8719*, ed. by S. de Freitas, C. Rensing, T. Ley, P.J. Munoz-Merino (Springer, 2014), pp. 596–597
6. M. Vahdat, A. Ghio, L. Oneto, D. Anguita, M. Funk, M. Rauterberg, Advances in learning analytics and educational data mining, in: *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges* (2015)
7. D. Petkovic, M. Sosnick-Pérez, K. Okada, R. Todtenhoefer, S. Huang, N. Miglani, A. Vigil, *Using the Random Forest Classifier to Assess and Predict Student Learning of Software Engineering Teamwork, Frontiers in Education (FIE)* (Erie, PA, 2016)
8. J. Kuzilek et al., Open university learning analytics dataset. Sci. Data **4**, 170171 (2017). https://doi.org/10.1038/sdata.2017.171
9. Available in https://rapidminer.com/
10. Available in https://www.knime.com/
11. Available in https://www.smartbi.com.cn/
12. Available in https://eric.msh-lse.fr/wricco/tanagra/
13. Available in https://orangedatamining.com/
14. Available in https://www.cs.waikato.ac.nz/ml/weka/
15. Available in https://scikit-learn.org/stable/
16. Available in https://www.investopedia.com/terms/d/decision-tree.asp
17. O. Okun, Feature Select. Ensemble Methods Bioinform. 13–31 (2011)
18. P. Andrzej, J. Luo, The more you learn, the less you store: memory-controlled incremental SVM for visual place recognition. Image Vis. Comput. **28**(7), 1080–1097 (2010)
19. P. Leif, K-nearest neighbor. Scholarpedia **4**, 2 (2009)
20. M. Krzywinski, N. Altman, Classification and regression trees. Nat. Methods **14**, 757–758 (2017)
21. C. C. Kiu, data mining analysis on student's academic performance through exploration of student's background and social activities, in *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)* (2018)
22. Y. Hu, C. Lo, S. Shih, Developing early warning systems to predict students' online learning performance. Comput. Hum. Behav. **36**, 469–478 (2014)
23. S. Huang, N. Fang, Predicting student academic performance in an engineering dynamics course: a comparison of four types of predictive mathematical models. Comput. Educ. **61**, 133–145 (2013)
24. H.T. Kahraman, S. Sagiroglu, I. Colak, The development of intuitive knowledge classifier and the modeling of domain dependent data. Knowl.-Based Syst. **37**, 283–295 (2013)
25. M. Bucos, B. Druagulescu, Predicting student success using data generated in traditional educational environments. TEM J. **7**(3), 617–625 (2018)
26. A.S. Carter, C.D. Hundhausen, O. Adesope, Blending measures of programming and social behavior into predictive models of student achievement in early computing courses. ACM Trans. Comput. Educ. **17**, 3 (2017)
27. L.C. Yu, C.W. Lee, H. I. Pan, C. Y. Chou, P.Y. Chao, Z.H. Chen, S.F. Tseng, C.L. Chan, K.R. Lai, Improving early prediction of academic failure using sentiment analysis on self-evaluated comments. J. Comput. Assist. Learn. (2018)
28. A. Khan, S.K. Ghosh, Data mining based analysis to explore the effect of teaching on student performance. Educ. Inf. Technol. **23**, 1677–1697 (2018)
29. N.S. Ahmed, M.H. Sadiq, Clarify of the random forest algorithm in an educational field, in *2018 international conference on advanced science and engineering (ICOASE)* (IEEE, 2018), pp. 179–184
30. H. Chanlekha, J. Niramitranon, Student performance prediction model for early-identification of at-risk students in traditional classroom settings, in *Proceedings of the 10th International Conference on Management of Digital Ecosystems—MEDES '18* (ACM, 2018), pp. 239–245

31. P.P. Bendangnuksung, Students' performance prediction using deep neural network. Int. J. Appl. Eng. Res. **13**(2), 1171–1176 (2018)
32. Y. Lee, D. Shin, H. Loh, J. Lee, P. Chae, J. Cho, S. Park, J. Lee, J. Baek, B. Kim, Y. Choi, Deep attentive study session dropout prediction in mobile learning environment, in *12th International Conference on Computer Supported Education* (2020)
33. Q. Liu, S. Tong, C. Liu, H. Zhao, E. Chen, H. Ma, S. Wang, Exploiting cognitive structure for adaptive learning, in *The 25th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD'19)* (2019)