An Educational Data Mining Model for Supervision of Network Learning Process

https://doi.org/10.3991/ijet.v13i11.9599

Jianhui Chen(✉), Jing Zhao
Zhengzhou University of Aeronautics, Zhengzhou Henan, China
jianhuichen10293@163.com

Abstract—To improve the school's teaching plan, optimize the online learning system, and help students achieve better learning outcomes, an educative data mining model for the supervision of the e-learning process was established. Statistical analysis and visualization in data mining techniques, association rule algorithms, and clustering algorithms were applied. The teaching data of a college English teaching management platform was systematically analyzed. A related conclusion was drawn on the relationship between students' English learning effects and online learning habits. The results showed that this method could effectively help teachers judge students' online learning results, understand their online learning status, and improve their online learning process. Therefore, the model can improve the effectiveness of students' online learning.

Keywords—data mining, statistical analysis visualization, association rule algorithm, clustering algorithm

1 Introduction

Informatization is applied in the field of education. Education informatization has played an important role in promoting education equity, realizing the sharing of educational resources, promoting the transformation of educational concepts, and cultivating innovative talents. In particular, the emergence of online learning has largely changed the traditional teaching methods, improved the teaching efficiency, and provided a way to achieve the fairness of education and the popularization of educational resources. However, in the application process, the process of network learning is difficult to supervise, and the learning effect is difficult to evaluate. The e-learning system stores a large amount of data based on the learner's learning trails and interactive process behavior. These data hide information such as the learner's online learning behavior. If educators can mine and analyze these data, it is possible to grasp the learner's online learning situation. The e-learning process is monitored and evaluated. It helps learners adjust the learning process in a timely manner. The analysis method of massive data generated during the network learning process was studied. Large data volumes and complex data analysis methods are huge challenges for most educators. The application of data mining technology in education provides powerful technical support for network learning data analysis. The era of big data has arrived and data is used to drive

school education. Data mining techniques are used to build educational models. The relationship between educational variables is explored, which provides decision support for the development of education and teaching. It has become an inevitable trend in the development of information education. Using the theories and techniques of pedagogy, computer science, psychology, and statistics, the problems in educational research and teaching practice are solved. According to the definition of the International Education Data Mining Working Group website, educational data mining refers to the use of evolving methods and techniques to explore data types in specific educational environments and to extract valuable information to help teachers better understand students. The learning environment was improved. This provides services for educators, learners, administrators and other educators. The main objectives of educational data mining are as follows. The learner model is built to predict the development of learning. The existing teaching contents and models are analyzed, and suggestions for improvement and optimization are put forward. The effectiveness of various educational software systems was assessed. Educational domain models are built to promote effective learning. The data of educational data mining can not only come from online learning systems or educational office software, but also from traditional learning classrooms or traditional test results. Data attributes can be either personal information (demographic information) or learning process information. The educational data mining process includes three stages: data acquisition and preprocessing, data analysis and result interpretation. The models of educational data mining can be divided into descriptive models and predictive models. Descriptive models are used for the description of the model, which provides a reference for decision making. Predictive models are primarily used for data-based predictions (such as predicting student achievement or course adoption). Based on the analysis of educational data mining technology and application status, the educational data mining model of network learning process supervision is established. Taking the university English teaching management platform of the university's FLTRP as an example, the case application research was carried out to verify the feasibility of the model.

2 State of the art

The rapid growth of educational data, the diversity of data types, the availability of data, and the development of data mining technologies have contributed to the development of educational data research. Learner models, academic achievement prediction, behavior pattern discovery, learning feedback and evaluation are the main hotspots of current educational data research.

At present, foreign scholars have relatively perfect research on the application of educational data mining in online learning. For example, Akhtar et al. [1] used clustering algorithms to analyze learning data in the Moodle platform. Students with similar learning characteristics are identified. According to the classification results, the student's learning effect is judged. This proves the feasibility of data mining technology in the network learning process. Aali et al. [2] used a decision tree algorithm to predict the influencing factors of students' academic success. Multiple model views were used

to build a complete educational data mining system. Hahsler and Karpienko [3] studied the learning behavior data of a distance learning platform. The association rule algorithm is used to explore the inherent law between learning style, learning behavior and academic achievement. It provides help and advice for teaching decisions and teaching optimization. Luna et al. [4] studied the application of data mining in curriculum management systems. Data mining techniques were used to analyze learner preferences. Studies have shown that field-independent learners often use forward or back buttons but spend less time on navigation. On-site dependent learners often use the main menu and have more repeat visits. Natek and Zwilling [5] use knowledge discovery techniques to analyze historical student course performance data to predict student dropout probability. Data from more than 100 students were tested using a logistic regression model. Data is analyzed by data mining techniques to predict student behavior. According to the results, the student's counseling behavior improved. Pardo et al. [6] believe that predicting student performance may be helpful in assessing students' learning mentality. By predicting the student's final outcome, bad learning behavior is improved at an early stage. The J48 decision tree algorithm was used to study three students and establish a predictive model. According to the completion time of students, the difficulty of students' curriculum and situational attributes is also different. This proves the importance of the mode adopted by students in curriculum performance and learning time. Verma et al. [7] used data mining technology to realize personalized teaching of distance learning system. The resources of the distance learning system can be configured around the requirements of the students. The student's individualized learning needs are supported and the student's academic achievement is predicted. Yukselturk et al. [8] proposed a learning behavior analysis model and a classification theory of learning outcomes. The academic performance prediction framework is designed. It includes three modules: analysis of learning content, analysis of learning behavior, and analysis of learning prediction. The correlation between student online behavior and student final academic performance was studied. Multiple regression analysis was used to analyze learners' online learning behavior data, and early warning factors affecting academic performance were judged.

In summary, many researchers have used educational data mining to conduct research in the process of online learning, and have obtained useful information and conclusions. It provides decision support for e-learning improvements. However, most of the research focuses on relational research, and there are few studies on the related network learning process supervision and the overall model construction. Therefore, an effective e-learning process governance model is designed. Educational data mining results provide decision support for the supervision and management of the online learning process. In the future, it is still an important issue worthy of further study.

3 Methodology

3.1 Process analysis

According to the special attributes of network learning and the educational data mining process, the educational data mining model of network learning process supervision

is constructed. The data source mainly comes from the network learning platform database, as well as the student course test scores and personal information in the educational management platform database. Due to the variety of data sources, data must be pre-processed after data collection, including removal of redundant data, processing of missing data, and numerical conversion. Educational data mining model for web-based learning process supervision is shown in Figure 1.

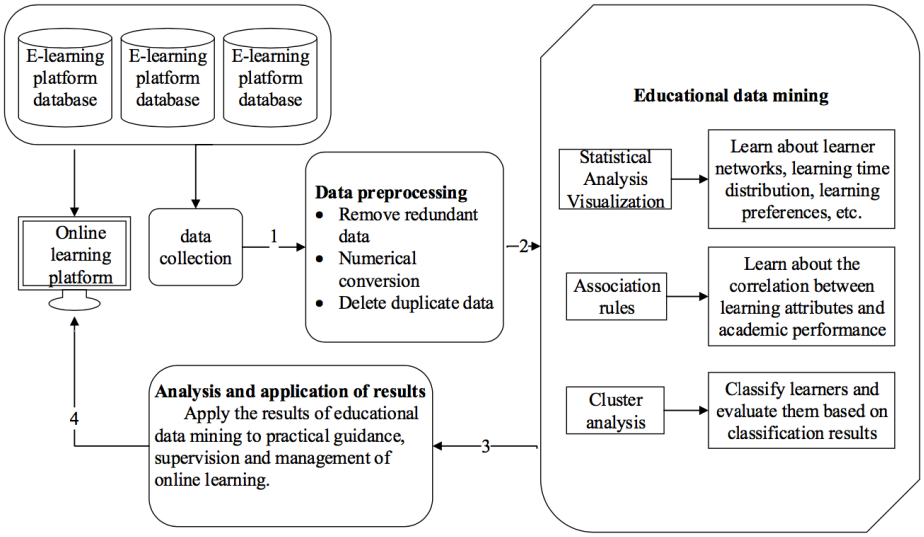

Fig. 1. Educational data mining model for web-based learning process supervision

After data preprocessing is completed, the core part of educational data mining is to select mining methods to analyze the data and obtain the results. For the learning process supervision of the e-learning platform, statistical analysis and visualization methods are used to understand the learner's e-learning time distribution and preference page. Association rules are used to analyze the relationship between e-learning attributes and academic performance. Cluster analysis is used to classify learners. Teachers can conduct different forms of supervision on various types of students according to the classification results and can also give corresponding network learning effect evaluation according to the classification results. Finally, the results of educational data mining are applied to the supervision of the online learning process. Students conduct a new round of network learning, generate new network learning data, and continue to analyze the new data generated. The online learning process is adjusted and optimized to achieve sustainable development towards the goals of research learning and autonomous learning.

3.2 Research object

The FLTRP (foreign language teaching and research press) university English teaching management platform is the research object. More than 500,000 online learning

data of 5,200 freshmen in the second semester of 2016-2017 were selected. It includes page view information, online learning time length, online score test and other data sheets. The education data mining is carried out in conjunction with the student information table and the score sheet in the educational management platform database. The students' online learning status is analyzed, and the student learning rules are summarized. It provides a way for teachers to supervise students' online learning process and evaluate the effectiveness of online learning.

3.3 Statistical analysis and visualization

According to the 16-week complete learning data and related information records of the students, the registration rate of the students logging into the online learning platform is calculated. The weekly registration rate is calculated by dividing the number of registered students per week by the total number of students, and statistics are performed by gender, as shown in Figure 2. In the process of online learning, the number of student logins does not reflect the student's activity correctly, but the number of logins can more accurately reflect the student's login status.

It can be seen that 60% of the students entered the online learning platform in the first week of school, and the network learning showed a good momentum, but since then the total login rate has been maintained at 50% to 60%. In the final phase of the semester, the login rate has dropped. The overall login rate of girls is high and higher than the total login rate. The participation of boys' online learning platform is far lower than that of girls, which needs to be paid attention to and supervised. The overall turnover rate of boys is large, which indicates that boys' participation in online learning is relatively scattered and their persistence is not strong. Weekly login rate chart is shown in Figure 2. College weekly registration rate chart is shown in Figure 3.

According to the statistics of the online platform registration rate of each student in each college, the four most representative colleges are selected for specific analysis, as shown in Figure 3. There are more girls in the liberal arts college. The login rate trend is basically the same as the total login rate in Figure 2, but the login rate declines significantly at the end of the semester. The art college registration rate fluctuates greatly. In the 5th week to the 11th week, no student was logged in for 7 consecutive weeks. The registration rate of the electrical engineering institute is almost the same as the total registration rate, and the trend is very flat. It shows that the stability and continuity of the students' learning is good. The computer colleges of the same engineering department are fluctuating. Figure 3 shows that there is a large difference in the login rates between colleges. There are significant differences in English learning habits between liberal arts students. English teachers should regularly exchange online teaching experience to improve the effectiveness of online teaching.

In the English e-learning system platform, there are four main pages for students: home page, reading and writing page, listening and speaking page, and online testing. Among them, the home page is the information selection page, and there is no substantive learning content. According to the duration of each student's stay on 4 pages, the page with the most time is selected as its preference page, and the number of people on each page is counted. The conclusion is that 20.36% of students prefer the home page.

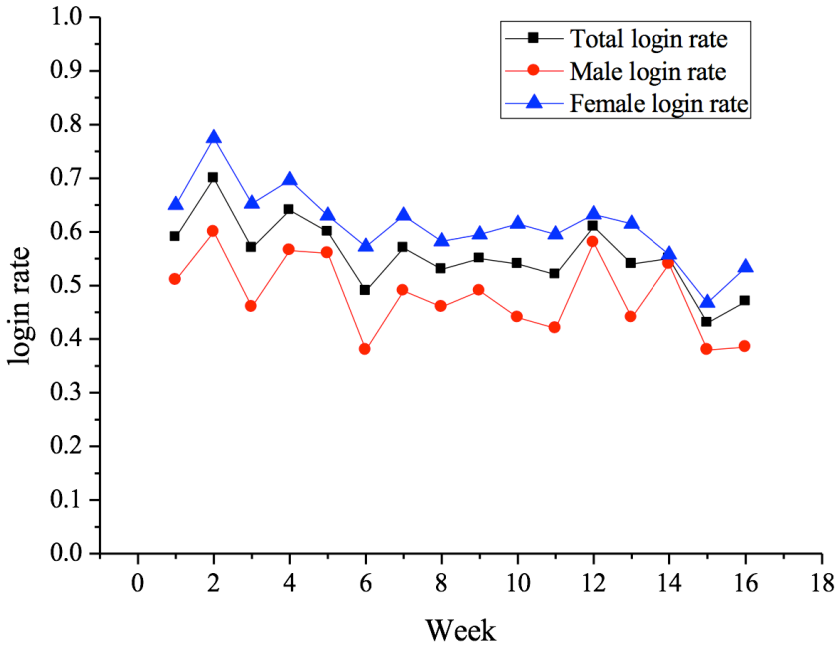

Fig. 2. Weekly login rate chart

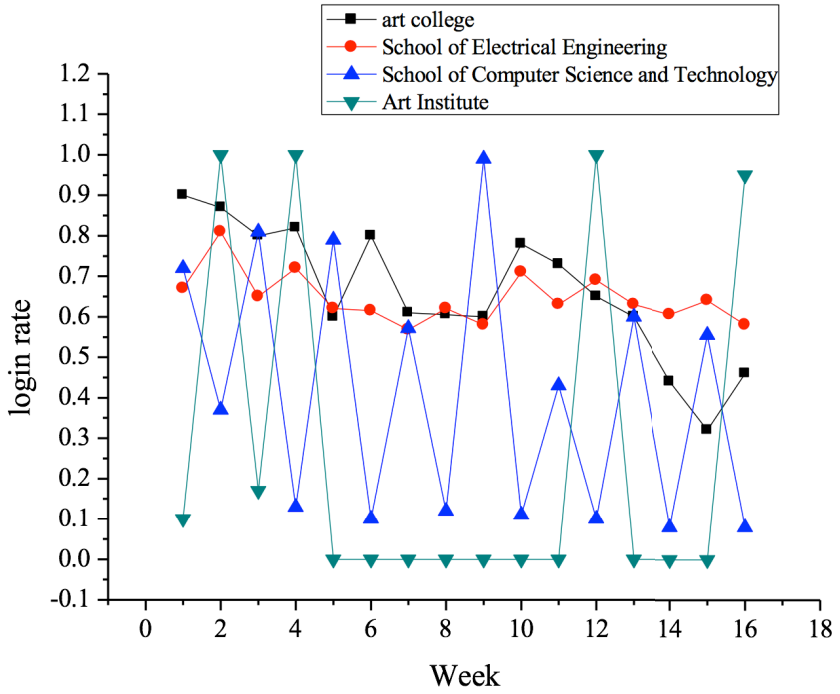

Fig. 3. College weekly registration rate chart

11.88% of students prefer the listening and speaking page. 33.62% of students prefer the reading and writing page. 34.14% of students prefer the online testing. Positive guidance is very important. It gives students an interest in and training in speaking and listening. English should really become a language skill for students, not just to cope with exams.

3.4 Association rule mining

Association rule mining is a rule-based machine learning algorithm. The algorithm can find relationships of interest in large databases. Its purpose is to use some metrics to distinguish strong rules that exist in the database. That is to say, association rule mining is used for knowledge discovery, not prediction, so it is an unsupervised machine learning method. Association rules are used to discover associations between attributes. By mining frequent itemsets, the links between attributes are obtained. The association rule mining process mainly consists of two phases: In the first phase, all the high frequency project groups are found from the data set. In the second phase, association rules are generated from these high frequency project groups. The general rules are described as a pair containing the left item set (condition) and the right item set (conclusion), and the importance and credibility of the rule are measured by confidence and support. Microsoft association rules are selected. Confidence is an attribute of an association rule. The predictability of the rules is determined by the usefulness and importance of confidence and importance judgment rules. Importance is also known as interest score or gain, which is used to measure item sets and rules and test the validity of the rules. The higher the importance score, the better the quality of the rule.

Association rules are applied to the supervision of the online learning process. The relationship between online learning attributes and academic achievement is found, which helps teachers to understand the students' online learning status, so as to better supervise the students' learning situation, give feedback in time, improve students' learning efficiency and increase teaching quality. Association rules require data attributes to be discrete data. The equal-frequency binning method is used to discretize the relevant data. The discretization coding table for each attribute is shown in Table 1. The attributes of other association rules include student preference page and gender.

Table 1. Discretization coding table for each attribute

Attributes	Value	Code	Attributes	Value	Code	Attributes	Value	Code
Online time	<10h	T1	***Home time***	<10h	F1	***Final English score***	<72	S1
	45~64h	T2		10~18h	F2		72~80	S2
	>64h	T3		>18h	F3		>80	S3
Learning interface	<20h	X1	***Test page time***	<15h	E1			
	20~32h	X2		15~22h	E2			
	>32h	X3		>22h	E3			

Based on the importance score, the following five important rules are selected: Rule 1: Preference page = "online test", learning page time = "X2", final English score = "S3". The confidence level is 0.831 and the importance is 0.421. The test page takes the most time, the learning page time is at a medium level, and the final English score is good. Rule 2: Online time = "T1", learning page time = "X1", ending English score = "S1". The confidence level is 0.546 and the importance is 0.382. If there is less online time, the study page takes less time and the English score at the end of the period is bad. Rule 3: Preference page = "Home page", online time = "T1", final English score = "S1". The confidence level is 0.667 and the importance is 0.305. The homepage takes the most time. If the online time is less, the English score at the end of the period is bad. Rule 4: Test page time = "E1", learning page time = "X1", final English score = "S1". The confidence level is 0.526 and the importance is 0.267. The test page takes less time, the learning page takes less time, and the English score at the end of the period is poor. Rule 5: Learning page time = "X3", test page time = "E3", final English score = "S3". The confidence level is 0.459 and the importance is 0.222. The learning page takes a lot of time, the test page takes a lot of time, and the English score at the end of the period is good. According to the above five rules, if the students' online learning habits are better, that is, they spend more time on online learning and spend more time in the learning module, the final English can achieve good results. However, students' online learning habits are poor, that is, they spend a small amount of time on online learning, and spend more time on the homepage, so the English scores at the end of the period tend to be poor. This shows that students' good online learning habits are closely related to academic achievement. Therefore, teachers can promptly monitor and remind students who have poor online learning habits.

3.5 Cluster analysis

Cluster analysis refers to the process of grouping a collection of physical or abstract objects into multiple classes of similar objects. Clustering is a process of classifying data into different classes or clusters, so objects in the same cluster have great similarities, and objects between different clusters have great dissimilarity. The principle of division is that there is a high degree of similarity between objects in the same cluster. There are large differences between objects of different clusters. Clustering analysis is simple and intuitive, which is mainly applied to exploratory research. The results of the analysis can provide multiple possible solutions, and the selection of the final solution requires the subjective judgment of the researcher and subsequent analysis. There are many clustering algorithms, such as EM clustering and K-means clustering. Among them, the K-means clustering algorithm is characterized in that each object can only be assigned to one cluster. Clusters are not connected to each other and do not overlap each other. Here, clustering analysis is used to classify learners, and the K-means algorithm is more suitable.

The purpose of cluster analysis is to classify students according to their learning behavior characteristics, which facilitates teachers to manage and evaluate students in a targeted manner, and provides timely feedback to students. Students adjust their study plans according to their own classification to achieve better learning results. The data

is still discretized (similar to the discretization principle of Table 1), including online time attribute T, reading and writing page time attribute R, listening and speaking page time attribute L, and online test score attribute E. Among them, the attribute E is discretized from E to E (E0, E1, E2, E3) according to the score. E0 means that no online test has been performed. The other three time attributes are discretized from T to T (T1, T2, T3), R (R1, R2, R3), L (L0, L1, L2, L3). L0 indicates that the listening and speaking page learning has not been performed. Cluster analysis is performed based on these four attributes, and the number of clusters is set to 3. The clustering results are shown in Table 2.

Table 2. Clustering group table

		Online time			**Listening and speaking interface time**			
		T1	***T2***	***T3***	***L0***	***L1***	***L2***	***L3***
Moderate	Category 1	46.43%	33.06%	0%	38.23%	39.45%	33.41%	5.77%
Self-conscious	Category 2	0.04%	47.95%	100%	21.52%	22.31%	44.35%	89.46%
To be adjusted	Category 3	53.53%	18.99%	0%	40.25%	38.24%	22.24%	4.77%
		Reading and writing interface time			**Online test results**			
		R1	***R2***	***R3***	***E0***	***E1***	***E2***	***E3***
Moderate	Category 1	0%	92.13%	9.11%	35.59%	32.19%	37.47%	28.68%
Self-conscious	Category 2	5.69%	7.87%	90.89%	23.72%	44%	39.05%	59.80%
To be adjusted	Category 3	94.31%	0%	0%	40.69%	23.81%	23.48%	11.52%

According to Table 2, the student learning situation of Category 1 is at a medium level, and the values of each learning data attribute are basically in the middle segment, so such students are defined as moderate. The learning situation of students in category 2 is quite good. The login time and the listening page are longer, and it takes a lot of time to learn on the reading/writing page. The final test scores are also good, so these students are defined as self-conscious. The students in category 3 have a poor learning situation. The data in the table reflects that the students' online learning time is small. Almost every page is in the range with less time. The test scores are mostly unsatisfactory, so it is defined as the type to be adjusted. Teachers should always pay attention to and remind such students. The clustering result can also provide teachers with certain learning effect evaluation reference. According to the clustering results, students are provided with e-learning process scores. Teachers can take different educational measures according to different categories of students to ensure learning efficiency.

4 Result analysis and discussion

The main content and purpose of educational data mining is to mine and analyze educational data. Its sublimation is based on the analysis results to provide decision support opinions, thereby improving the teaching process and teaching effects. Based on the analysis results, the following conclusions and recommendations are drawn:

The online time of online learning cannot be an absolute factor in judging the effectiveness of student network learning. The evaluation and supervision of e-learning is a difficult task. Many teachers mainly evaluate and consider the effects of online learning based on their online learning time. This evaluation method is not comprehensive. The study found that although some students' online time is longer, the learning effect is not good. Therefore, the supervision of online learning content should be strengthened.

Online learning habits are associated with academic performance. The results of association rules show that students have good online learning habits, and their examination results are often excellent. Therefore, it is very important to cultivate students' good habits of online learning. When teachers supervise students' online learning, if they find that students' online learning habits are not good, they should give timely reminders and urge students to adjust in time.

According to the characteristics of online learning behavior, students are clustered, which provides an important reference for teachers. Teachers can arrange different learning tasks for students of different categories, and focus on the classification of students with poor performance. According to the clustering results, students with different classifications are given different learning effects.

Teachers play an important role in online learning. Online learning is not for students to study alone. Teacher supervision and teacher-student communication are very important in the process of online learning. Adequate communication can promote students' interest in learning and motivation. As a guide and supervisor, teachers should enhance supervision and communication with students, giving timely feedback and guidance.

Rules and regulations enable the supervision of web-based learning process to follow. Managers should develop a practical online learning process supervision system. On the one hand, teachers have rules to follow in the process of supervision; on the other hand, the power, compulsory and rationality of supervision can be guaranteed. According to the continuous iteration of the educational data mining process, managers should adjust and optimize the network learning mode in time. Autonomous learning and research learning are realized.

5 Conclusions

The supervision of the online learning process, the evaluation of learning effects, and the optimization of the network learning structure are issues that educators need to constantly study. The aim is to achieve research learning and autonomous learning. The educational data mining model of e-learning process supervision is designed and applied to practice. The student's online learning data and related information are analyzed from multiple perspectives. However, due to the limited level of personal skills, the analytical methods used for student data excavation analysis are not comprehensive enough. The process of sampling and processing data also has certain drawbacks. Although the conclusion cannot fully reflect the relationship between student online learning data and student behavior and performance, it can provide certain reference value for the process supervision and evaluation system of online learning. In addition, it

provides methods and basis for the application of data mining in network learning. The subject still needs to be further studied. The model and data source of data mining still need to be enriched and perfected. In the future, based on this, a set of timely analysis software system will be generated to provide real-time analysis and feedback for teachers and students to promote the development of online learning.

6 References

[1] Akhtar, S., Warburton, S., Xu, W. (2017). The use of an online learning and teaching system for monitoring computer aided design student participation and predicting student success. International Journal of Technology and Design Education, 27(2): 251-270 https://doi.org/10.1007/s10798-015-9346-8

[2] Aali, M., Bhaskaran, S. S., Lu, K. (2017). Student performance and time-to-degree analysis by the study of course-taking patterns using J48 decision tree algorithm. International Journal of Modelling in Operations Management, 6(3): 194 https://doi.org/10.1504/IJMOM.2017.10005808

[3] Hahsler, M., & Karpienko, R. (2017). Visualizing association rules in hierarchical groups. Journal of Business Economics, 87(3): 317-335 https://doi.org/10.1007/s11573-016-0822-8

[4] Luna, J. M., Castro, C., Romero, C. (2017). MDM tool: A data mining framework integrated into Moodle. Computer Applications in Engineering Education, 25(1): 90–102 https://doi.org/10.1002/cae.21782

[5] Natek, S., & Zwilling, M. (2014). Student data mining solution–knowledge management system related to higher education institutions. Expert systems with applications, 41(14): 6400-6407 https://doi.org/10.1016/j.eswa.2014.04.024

[6] Pardo, A., Han, F., Ellis, R. A. (2017). Combining university student self-regulated learning indicators and engagement with online learning events to predict academic performance. IEEE Transactions on Learning Technologies, 10(1): 82-92 https://doi.org/10.1109/TLT.2016.2639508

[7] Verma, S. K., Thakur, R. S., Jaloree, S. (2017). Fuzzy association rule mining based model to predict students' performance. International Journal of Electrical and Computer Engineering (IJECE), 7(4): 2223-2231 https://doi.org/10.11591/ijece.v7i4.pp2223-2231

[8] Yukselturk, E., Ozekes, S., Türel, Y. K. (2014). Predicting dropout student: an application of data mining methods in an online education program. European Journal of Open, Distance and E-learning, 17(1): 118-133 https://doi.org/10.2478/eurodl-2014-0008

7 Authors

Jianhui Chen and **Jing Zhao** are with the School of Computer, Zhengzhou University of Aeronautics, Zhengzhou Henan 450015, China (jianhuichen10293@163.com).

Article submitted 27 September 2018. Resubmitted 18 October 2018. Final acceptance 27 October 2018. Final version published as submitted by the authors.