Sustainable Development of College and University Education by Use of Data Mining Methods

https://doi.org/10.3991/ijet.v16i05.20303

Liwen Wang
Zhejiang Industry Polytechnic College, Shaoxing, China

Soo-Jin Chung (✉)
Wonkwang University, Iksan-si, Korea
wangyeomun@gmail.com

Abstract—To improve the education efficiency of the students, the student-centered education plan is explored. First, the Apriori algorithm of association rules is used to mine the potential related patterns in the score data of college students and establish a reasonable teaching method. Second, aided by the decision tree model, the factors affecting students' academic performance are studied, and the potential relationship between different courses is studied. Finally, the Apriori algorithm of association rules combined with decision tree model is used to generate the early warning mechanism of students' achievement, and the course performance of college students is empirically analyzed. The results show that: C language has two sides of dependence on many subjects; higher mathematics linear algebra mathematical statistics computer composition principle computer network. The teaching scheme of C language C++ Java more conforms to the learning mechanism of college students. Through empirical analysis, the early warning mechanism of association rule Apriori algorithm and decision tree model can effectively analyze student's course and give student's achievement. It is found that the method proposed can provide theoretical basis for students, teachers, and university administrators to carry out education reform and education management decision-making, improve students' performance and education quality, and realize the "student-oriented" education concept, so it can be applied to the actual education management.

Keywords—"Students-oriented"; data mining; association rules; Apriori algorithm; sustainable development

1 Introduction

The concept of student-oriented is a people-oriented idea, which emphasizes respect and support for students. It is student-centered and applicable to all students. In school teaching management, the guiding principle of student-oriented must be adhered to, and corresponding reforms in school system, policy and environment should be implemented. As the "people-oriented" concept is continuously

implemented, it has positively deepening the reform of higher education management system. In order to grasp the scientific connotation of "people-oriented" in higher education, the necessity and internal requirements of "people-oriented" in higher education should be deeply understood, which will help further exploring the value orientation of "people-oriented" in higher education. The concept of human-oriented provides some sustainable development strategies for implementing higher education in university [1-3].

The school information management system stores much potentially valuable performance data information. For much achievement data, association rules in data mining technology can be used to obtain the association relationship between subjects and find the neglected content in learning, which can provide targeted help and academic early warning for each student, and provide teaching guidance for teachers and administrators. Data mining is regarded as a set of technologies that allow automatic or semi-automatic extraction of much useful information, models and trends from many datasets, such as "clustering", "classification", "association" and "regression"; intelligent artificial algorithms such as Apriori algorithm, Bayesian algorithm, and neural network will be used to extract patterns from data, and these patterns are realized to interpret and predict their behaviors [4-8]. Association rule is an important task to discover frequent patterns in data mining. It has been successfully used in computer network, recommendation system, and medical care [9-10]. Wang et al. (2020) studied an improved Apriori algorithm for time series of frequent itemsets, applied it to mining association rules based on time constraints, and concluded that this algorithm was superior to the traditional algorithm in storage space [11]. Sun et al. (2020) built 0-1 transaction matrix by scanning transaction database to gain weighted support and confidence. The method can shorten the running time, reduce the memory demand and the number of operations, and effectively extract the hidden and valuable items [12].

At present, due to the lack of data mining analysis of students' scores and related application research, the data is stored and processed by establishing a data warehouse of students' academic performance, and reliable experimental data is provided for the subsequent analysis of students' academic performance. The Apriori algorithm of association rules is used to fully mine the association between subjects and the hidden valuable information, which plays an early warning role in students' subject scores, and provides technical service support for teachers' teaching plan reform and university administrators.

2 Materials and Methods

2.1 Materials

1286 students majoring in computer science from 2016-2019 are selected from the academic affairs system of Zhejiang Industry Polytechnic College as research samples. Advanced mathematics, computer network, computer composition principle, C++ and other core courses are selected as experimental samples. Table 1 and Table 2

show the original data of different data sources, and Figure 1 is the flow chart of this experiment.

Table 1. Original data of some courses

Student ID	Course number	Semester	Score	Instructor
001	011001	1	80.0	Professor Li
002	011001	1	75.0	Professor Li
003	011002	2	79.0	Professor Wang
004	011002	2	68.0	Professor Wang

Table 2. Original data of some course scores

Advanced mathematics	Linear algebra	Computer network	Computer compose principle	Mathematical statistics	C language	Java	C++
80.0	73.0	78.0	82.0	77.0	81.0	78.0	75.0
75.0	68.0	73.0	71.0	66.0	71.0	69.0	70.0
62.0	65.0	69.0	62.0	66.0	60.0	63.0	62.0
96.0	93.0	90.0	87.0	91.0	95.0	92.0	93.0

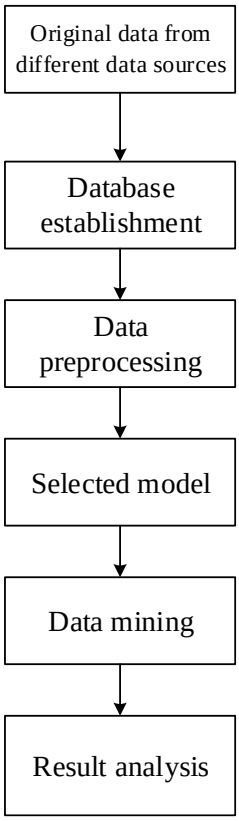

Fig. 1. Flow chart of this experiment

2.2 Methods

1. Establishment of data warehouse: The data warehouse is subject-oriented. The experimental data warehouse takes the analysis of students' scores as the theme, and adopts the top-down implementation mode and logic structure mode of centralized management to ensure its high operation efficiency in the rapidly increasing data related to students' scores. Before the establishment, the abstract subject is determined according to the needs and the data are extracted, transformed, and loaded under its control. Figure 2 is the data warehouse structure. The construction of data warehouse is realized by SQL Server 2019 in the actual design.

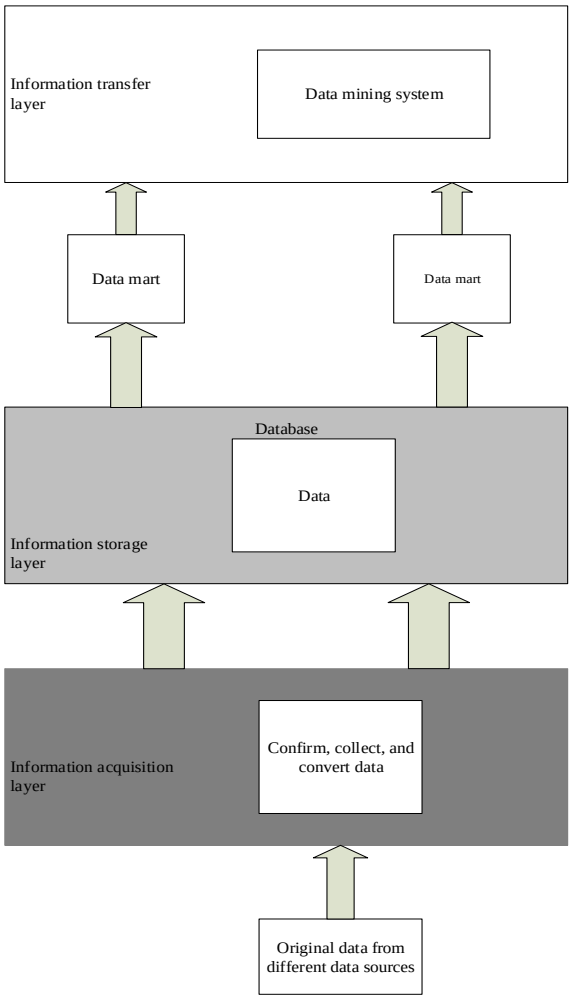

Fig. 2. Structure chart of data warehouse

a) Data confirmation and extraction: Data used for mining analysis from different data sources is confirmed and extracted.
b) Data preprocessing: Due to the different types and systems of the original data, there may be some problems in the data. Before loading the data into the data warehouse, it is necessary to process the original data to obtain more accurate data. The steps are as follows. Error data processing: in data collection, transfer, and loading, there will be various errors, such as the loss and dislocation of some data fields. These records need to be processed in advance. In the design of the system, in this case, the missing or misplaced fields in the data will be re-associated and matched by other means. If the corresponding matching cannot be achieved, it will be treated as missing value. Format correction: to ensure data normalization, data with the same attribute in different tables need to be corrected. Separation and combination of fields: in order to conduct more efficient data mining analysis, some fields in the original data need to be separated in advance. Data discretization: the student's score is a continuous measurement parameter, which is divided into five standards and simply discretized. Noise: it refers to the deviation or random error of a certain variable in the data. If it is not handled, it will affect the accuracy and quality of the data. Smoothing technology is used to deal with the noise.
c) Loading and storing data*:* After transformation, the original data will be loaded into data warehouse and stored, and OLAP tools will be used to extract decision-oriented data, such as reports and various views.

2. Data mining model:

a) Data mining: The experimental data mining model uses the Apriori algorithm to explore the relationship between students' courses and scores. The frequent itemset that meets the minimum support is extracted from the data, and the strong association rules that meet the preset minimum confidence and support are generated from it. The goal, data collection, and preprocessing of data mining model have been processed in the above data warehouse.
b) Data transformation: Acquired data are transformed into the data form suitable for data mining of this experiment. The Apriori algorithm is a Boolean algorithm, so the continuity of students' scores will lead to a lot of troubles in the operation process of the algorithm and often prolong the work. The continuous score data are transformed into Boolean data (0,1). Due to the differences in the evaluation standards of different courses, the conversion standard of the experimental data is the course average score. 1 means that students' score is higher than the course average score, and 0 means that students' score is below the course average. Table 3 is the results of partial data conversion.

Table 3. Data conversion results

Advanced mathematics	Linear algebra	Computer network	Computer Compose Principle	Mathematical statistics	C language	Java	C++
1	1	1	1	1	1	1	1
1	0	1	0	1	1	0	1
0	0	1	0	0	0	0	0
1	1	1	1	1	1	1	1

c) Definition of association rule: $I = \{i1, \ldots, ik\}$ is the itemset in the data warehouse DB (transaction set of data warehouse). If the number of items in I is k, I is set into ak- itemset, the association rule is $X \to Y$, X, Y is the true subset of I, and $X \cap Y \neq \emptyset$. The association intensity of association rules can be expressed by support and confidence that are defined as follows.

$$support(X \to Y) = \frac{\sigma(X \cup Y)}{N} \tag{1}$$

$$confidence(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \tag{2}$$

Where: σ indicates the support count of the itemset, and N indicates the number of total transaction sets.

If $support(X \to Y) \geq \min\_sup$ and $confidence(X \to Y) \geq \min\_conf$, association rule $X \to Y$ will be strong association rule, and min_conf and min_sup indicate the minimum confidence and support.

According to the data analysis and actual needs, the minimum support and the minimum confidence are set as 0.2 and 0.4, respectively. This data mining model compares the set minimum support and confidence with the confidence and support of each frequent itemset after preprocessing to get the corresponding association rules.

d) Decision-tree model of association rule merging: There is not only valuable information between different courses, but also hidden information between students' course scores and course information, such as time attribute information. Therefore, decision-tree is used to distinguish the course related attributes of different years and semesters. For the factors related to the early warning of students' scores, the influence of the correlation between courses, teachers, students, courses and other factors should be considered.

From the above association rules, the strong association rule with high credibility is obtained as the preselected new attribute, and it will be distinguished through the information gain. Finally, the new attribute is generated and it will be combined with the original data attribute to construct the decision-tree. Figure 3 is the flow chart of association rule merging decision-tree model.

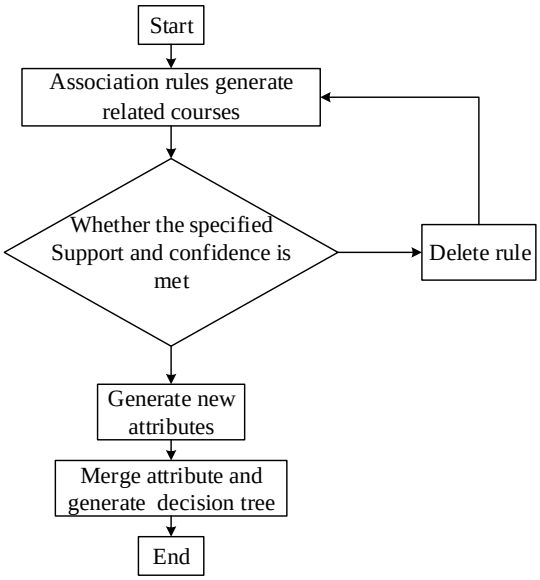

Fig. 3. Flow chart of association rule merging decision-tree model

According to the needs of mining, the scores of computer-majoring students are obtained from the educational administration system. After preprocessing, a score group (class) is added to the students' curriculum score table. All the scores are divided into five sections, namely ≥ 90, $90 > b \geq 80$, $80 > c \geq 70$, $70 > d \geq 60$, and $e < 60$. In the production process of the above association rules, the performance judgment of students in semesters and academic years is added, and Figure 4 is the flow chart. In order to narrow the mining scope and improve the mining quality, after the original data of different data sources are selected and processed, the analysis attributes set in this experiment are: course category, student source, gender, professional direction, absence of examination and course attributes. The minimum confidence and support of association rules are 0.5 and 0.2, respectively. Precise rules are generated, association rules that cannot meet the teaching plan are deleted, and the new attributes are evaluated. The attributes with high reliability are merged into the original attributes and the decision tree is constructed.

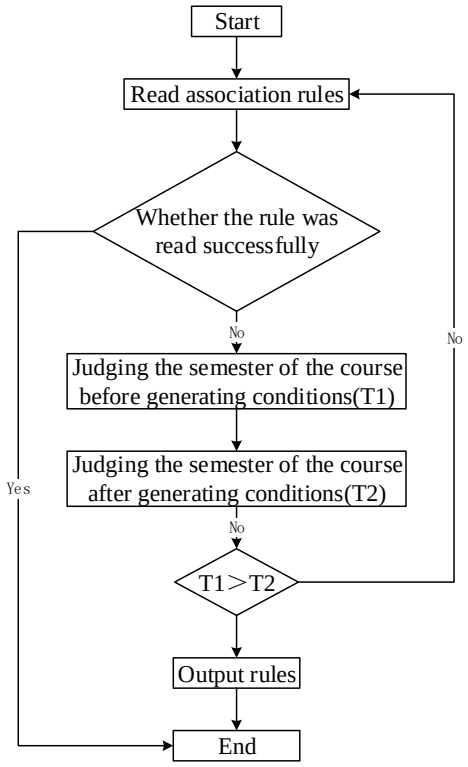

Fig. 4. Flow chart of student's semester score determination

3 Results

3.1 Analysis on the results of association rules of different courses

Table 4 shows that learning advanced mathematics first is conducive to learning linear algebra, with a confidence of 0.92 and a support of 0.35; learning linear algebra first benefits studying mathematical statistics, with a confidence of 0.86 and a support of 0.47; learning mathematical statistics first contributes to the study of computer composition principle, with a confidence of 0.87 and a support of 0.45; meanwhile, learning the principle of computer composition first conduces to the learning of computer network, with a confidence of 0.88 and a support of 0.52. Therefore, according to the experimental results, the scientific and reasonable teaching plan is obtained: higher mathematics → linear algebra → mathematical statistics → principles of computer composition → computer network. Additionally, learning C language first helps learning C++; learning C++ first is conducive to the learning of the Java, and the confidence and support are 0.85, 0.48, 0.87, and 0.34, respectively. Therefore, as a basic course of computer major, C language should be set up earlier.

The teaching plan recommended according to the experimental results is: C language, C++, Java.

Table 4. Association rules of some courses

Course	Course	Confidence	Support
Advanced mathematics	Linear algebra	0.92	0.35
Computer Compose Principle	Computer network	0.88	0.52
Mathematical statistics	Computer Compose Principle	0.87	0.45
Linear algebra	Mathematical statistics	0.86	0.47
C language	C++	0.85	0.48
C++	Java	0.87	0.34

3.2 Dependency network of some courses

Figure 5 shows that many courses have a two-side dependence on C language, which is consistent with the school's setting of this course as a basic professional course for computer-major students. Therefore, this course needs to be set up in the early learning of computer-major students and can enhance students' understanding and teachers' teaching quality. Moreover, advanced mathematics and the principles of computer composition depend on many computer professional courses, which shows that advanced mathematics and the principles of computer composition are the basic theoretical reserve courses for computer-students. Therefore, these two courses are particularly important for the follow-up course learning of students in this major. They should be learned and mastered as soon as possible, and teachers should also focus on the key explanation and management of the course.

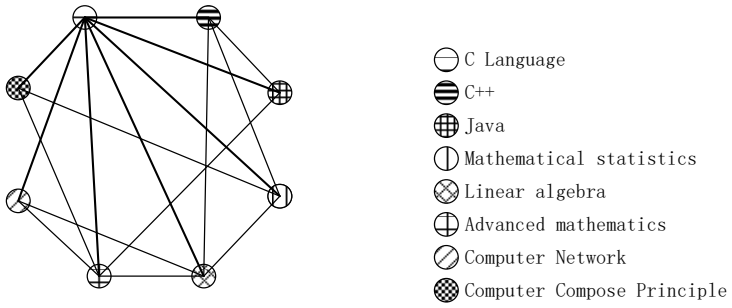

Fig. 5. Course dependency network

The above analysis results show that the courses of computer-major students are certainly dependent, and the order of courses also affects students' scores. According to the analysis of correlation rules, it can be concluded that the results of the previous courses will directly affect the results of the follow-up courses. Therefore, in order to improve students' scores and teachers' teaching quality, colleges and universities

should follow the scientific and reasonable theoretical basis when setting up students' courses.

3.3 High reliability association rule results

According to the Apriori algorithm and evaluation criteria, high confidence rules are obtained and processed by association rules. Then, strong association rules between students' course scores with time attribute are obtained, as shown in Table 5. The confidence of computer application foundation, electrical and electronic technology and computer assembly and maintenance is 0.9782; the confidence of local area network establishment, enterprise network security advanced technology, and enterprise network integrated management is 0.9571; the confidence of e-commerce process, e-commerce website design and production, and network marketing practice is 0.8902; the confidence of Linux server operating system, e-commerce process, and enterprise network integrated management is 0.8634; for the items mentioned above, their support are all 1.0000; the above courses are highly relevant, which is conducive to the analysis of early warning factors of students' scores.

Table 5. High reliability association rules

Rules	Confidence	Support
Computer Application Foundation, Electrical and Electronics $\dashrightarrow$ Computer Assembly and Repair	0.9782	1.0000
LAN Formation, Advanced Technology of Enterprise Network Security $\dashrightarrow$ Enterprise Network Comprehensive Management	0.9571	1.0000
E-commerce process, E-commerce Website Design and Production $\dashrightarrow$ Internet Marketing Practice	0.8902	1.0000
Linux Server Operating System, E-commerce process $\dashrightarrow$ Enterprise Network Comprehensive Management	0.8634	1.0000

3.4 Results of merging decision-tree of some association rules

The combination decision-tree analysis of strong association rules selected by Apriori algorithm can enhance the integrity of students' score early warning and excavate the hidden reasons of students' failure in subjects. Table 5 shows the more valuable hidden rules mined. The course category is one of the important influencing factors of students' score early warning, which is consistent with the actual experience evaluation results. Its branch attributes include basic courses, applied courses, and professional courses.

Figure 6 shows that when a girl fails in computer application foundation and electrical and electronic technology courses, she may also fail in computer assembly and maintenance to a large extent, which is consistent with the understanding that boys are better than girls in engineering learning and understanding. Through the analysis of the above results, the relevant factors influencing the students' academic

scores are obtained. Teachers and university administrators should pay attention to the phenomenon of students' failure of different courses, and put forward corresponding teaching reform programs and measures to avoid the students' failure of relevant courses, and improve the level of talents in colleges and universities. Moreover, it can avoid the phenomenon that the students who have failed in the course are constantly failing due to their carelessness and thus affect their studies.

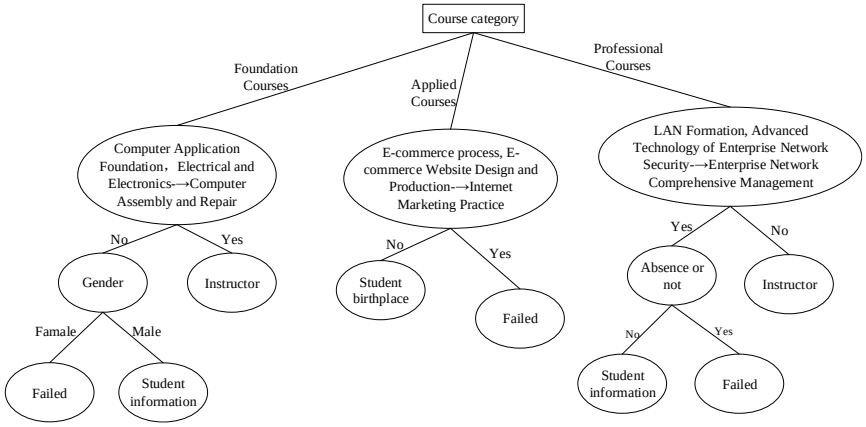

Fig. 6. Results of merging decision tree of some association rules

4 Discussion

The concept of "student-oriented" means that teachers and administrators regard students as the main part of learning activities. Teaching objectives, teaching environment, teaching materials, teaching process organization, teaching methods and other factors should be student-centered. All are designed and served for students' personal development and all-round development, so as to promote students' active and full participation in the process of teaching planning, continuously improve their scores and promote their healthy development physically and mentally [13-15]. The education industry is developing towards a direction of information and intelligence. Data mining technology has been applied to the field of education, such as teaching information management, teachers' teaching evaluation, analysis of students' psychological characteristics, formulation of scientific and reasonable teaching programs, and analysis of students' examination results to find problems and effectively strengthen teaching. Among them, the most commonly used data mining technology for predicting and classifying the factors affecting students' scores is decision-tree, Bayesian classifier and artificial neural network (ANN) [16-18]. Therefore, based on theoretical basis and previous experimental summary, association rule is built to mine the value data hidden in students' score data, and the decision-tree model is combined to analyze the factors that affect students' scores.

According to the test results of the proposed model, the deep learning model can be combined with the potential value information related to students to provide a basis

for the analysis of factors affecting students' scores, which is consistent with the research results of López-Zambrano et al. (2020) and David et al. (2020) [19, 20]. Based on the deep learning prediction model, the big data in MOOC (massive open online courses) and the information provided by counselors are used to predict and research the students' scores, and the prediction ability of the model based on the user's video viewing behavior is tested. It can be found that the frequency of watching video every week is better than the single watching function in predicting students' score. The model of association rules combined with decision-tree can effectively mine the dependence relationship between students' courses and predict the influencing factors of students' scores through the order of course scores, which is consistent with the research results of Preet et al. (2019) [21]. By investigating the demographic, social, academic and behavioral factors that affect students' performance, three accuracy-based technologies are used to build an integrated model. The 10 times cross validation technology is applied to access the suitability of the results obtained in the integrated model and students' performance is accurately predicted. The results show that the academic performance of last semester and other factors have a significant impact on the current academic performance; any serious accident in the past year will also affect academic score, and this model has high accuracy. Early identification of influencing factors on academic performance is conducive to early detection of high-risk students; thus, preventive and corrective measures can be put forward to improve students' overall academic performance.

Decision-tree is used to analyze the influence of new attributes on students' curriculum scores, to achieve an effective early warning of students' score, and provide a basis for the formulation of teachers' teaching plan. This result is consistent with the research results of Concepción et al. (2018) [22], who used the logistic regression model to classify the data experiments of students for distance learning courses, and effectively predicted the dropout risk of students.

To sum up, data mining technology based on the concept of "student-oriented" can be applied to education field, including the early warning of students' curriculum performance, teaching plan reform, management decision-making and other applications, and provide effective supervision and early prevention in the process of students' course learning.

5 Conclusion

Based on the concept of "student-oriented", association rule is used to mine the value data hidden in the students' score data, and the decision-tree algorithm is combined to analyze the factors affecting students' scores, thus providing theoretical support in the early warning of students' scores, teaching planning guidance, and teaching management mode. However, only the decision-tree and association rules are analyzed, and the analysis method is not enough. In the later stage, more methods such as ANN and cluster analysis can be included in the research of students' scores.

6 References

[1] Wang, S. Z. Tang, Q. (2018). Construction of Guiding System for Growth and Development of College Students under the Student-oriented Concept. Asian Agricultural Research,10(05): 89-91

[2] Kasai, H. Ito, S. Tajima, H. et al. (2020). The positive effect of student-oriented clinical clerkship rounds employing role-play and peer review on the clinical performance and professionalism of clerkship students. Medical teacher, 42(1): 73-78. https://doi.org/10.1080/0142159x.2019.1656330

[3] Sumit, G. Neena, S. Anurag, C. et al. (2017). Reforming pathology teaching in medical college by peer-assisted learning and student-oriented interest building activities: A pilot study, 30(2): 126-132. https://doi.org/10.4103/efh.efh_267_16

[4] Altinay F, Beyatli O, Dagli G, et al. (2020). The Role of Edmodo Model for Professional Development: The Uses of Blockchain in School Management. International Journal of Emerging Technologies in Learning (iJET), 15(12): 256-270. https://doi.org/10.3991/ijet.v15i12.13571

[5] Khlaisang J, Koraneekij P. (2019). Open online assessment management system platform and instrument to enhance the information, media, and ICT literacy skills of 21st century learners. International Journal of Emerging Technologies in Learning (iJET), 14(07): 111-127. https://doi.org/10.3991/ijet.v14i07.9953

[6] Sun, X. R. (2019). Research on time series data mining algorithm based on Bayesian node incremental decision tree. Cluster Computing: The Journal of Networks, Software Tools and Applications, 22 (12): 10361-10370. https://doi.org/10.1007/s10586-017-1358-6

[7] Muhisn Z, Ahmad M, Omar M, et al. (2019). The Impact of Socialization on Collaborative Learning Method in E-Learning Management System (eLMS). International Journal of Emerging Technologies in Learning (iJET), 14(20): 137-148. https://doi.org/10.3991/ijet.v14i20.10992

[8] Wang, X. L. Su, K. Su, L. R. et al. (2019). Research on Improved Apriori Algorithm Based on Data Mining in Electronic Cases. International Journal of Healthcare Information Systems and Informatics, 14(3): 16-28. https://doi.org/10.4018/ijhisi.2019070102

[9] Akbar, T. Gandomi, A. H. Asadollah, S. (2020). A survey of evolutionary computation for association rule mining. Information Sciences, 524: 318-352. https://doi.org/10.1016/j.ins.2020.02.073

[10] Shingo, M. Takuro, H. Takashi, K. (2020). SemiSupervised Learning for Class Association Rule Mining Using Genetic Network Programming. IEEJ Transactions on Electrical and Electronic Engineering, 15(5): 733-740. https://doi.org/10.1002/tee.23109

[11] Wang, C. X. Zheng, X. Y. (2020). Application of improved time series Apriori algorithm by frequent itemsets in association rule data mining based on temporal constraint. Evolutionary Intelligence, 13 (1): 39-49. https://doi.org/10.1007/s12065-019-00234-5

[12] Sun, L. N. (2020). An improved apriori algorithm based on support weight matrix for data mining in transaction database. Journal of Ambient Intelligence and Humanized Computing, 11 (2): 495-501. https://doi.org/10.1007/s12652-019-01222-4

[13] Zhang, Y. Q. Li, Y. B. (2019). The Implementation Strategies of Individualized Education in the Field of Basic Education. Open Journal of Social Sciences, 7(6): 223-232

[14] Vivek, P. K. Remya, T. P. Veenus, P. K. et al. (2019). Student-Centered Pedagogy – Lessons from DPEP. International Journal of Smart Education and Urban Society, 10(4): 17-29. https://doi.org/10.4018/ijseus.2019100102

[15] Yang, D. Wu, S. X. Y. Wang, W. S. et al. (2019). Is the Student-Centered Learning Style More Effective Than the Teacher-Student Double-Centered Learning Style in Improving

Reading Performance? Frontiers in psychology, 10: 2630. https://doi.org/10.3389/fpsyg.2019.02630

[16] Hooshyar. D. Pedaste, M. Yang, Y. et al. (2020). Mining Educational Data to Predict Students' Performance through Procrastination Behavior. Entropy, 2020, 22(1): 12. https://doi.org/10.3390/e22010012

[17] Amjed, A. S. Mostafa, A. E. Khaled, S. et al. (2019). Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques. Technology, Knowledge and Learning, 24 (4): 567-598. https://doi.org/10.1007/s10758-019-09408-7

[18] Qu, S. J. Li, K. Wu, B. et al. (2019). Predicting Student Achievement Based on Temporal Learning Behavior in MOOCs. Appl. Sci. 9(24): 5539. https://doi.org/10.3390/app9245539

[19] López-Zambrano, J. Lara, J. A. Romero, C. et al. (2020). Towards Portability of Models for Predicting Students' Final Performance in University Courses Starting from Moodle Logs. Appl. Sci. 10(1): 354. https://doi.org/10.3390/app10010354

[20] David, J. L. Tenzin, D. et al. (2020). Grade prediction of weekly assignments in MOOCS: mining video-viewing behavior. Education and Information Technologies: The Official Journal of the IFIP Technical Committee on Education, 25 (c): 1333-1342. https://doi.org/10.1007/s10639-019-10022-4

[21] Preet, K. Sachin, A. et al. (2019). An ensemble-based model for prediction of academic performance of students in undergrad professional course. Journal of Engineering, Design and Technology, 17(4): 769-781. https://doi.org/10.1108/jedt-11-2018-0204

[22] Concepción, B. Campanario, M. L. David, L. et al. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. Computers and Electrical Engineering, 66: 541-556. https://doi.org/10.1016/j.compeleceng.2017.03.005

7 Authors

Liwen Wang was born in shaoxing, Zhejiang, P.R. China,in 1982 .She received the Master degree from the university of electronic science and technology.Now, she works in Zhejiang Industry Polytechnic College.Her research interests include educational management, software engineering. Liwenwanghaha199@yeah.net

Soo-Jin Chung is the Professor of the International Exchange Department, Wonkwang University, Iksan-si, Korea.

Article submitted 2020-12-05. Resubmitted 2021-01-12. Final acceptance 2021-01-13. Final version published as submitted by the authors.