An Improved Apriori Algorithm for Association Mining Between Physical Fitness Indices of College Students

https://doi.org/10.3991/ijet.v16i09.22747

Tao Pan
Guizhou University of Commerce, Guiyang, China
201510468@gzcc.edu.cn

Abstract—The physical fitness of college students can be evaluated scientifically based on the data of physical education (PE). This paper firstly relies on the Apriori algorithm to mine the hidden correlations between the physical fitness indices from the PE data on college students, and identify the indices closely associated with the physical fitness of college students. Then, the Apriori algorithm was improved to reduce the time complexity of association rule mining. Based on the improved algorithm, it was learned that the correlation coefficients of several indices surpassed the minimum support of 0.2 and minimum confidence of 0.7, reflecting their important impacts on physical fitness. Thus, physical fitness of college students is significantly influenced by speed, endurance, flexibility, and vital capacity, but not greatly affected by height and weight. The research results provide an important guide for the test and curriculum designs of PE for college students.

Keywords—Apriori algorithm, data mining, association rules, physical education (PE), physical fitness indices

1 Introduction

With the progress in data mining technology, there has been a steady growth in the volume of data on physical education (PE) of college students. This trend is expected to continue in the foreseeable future. The massive PE data suggest a continuous decline in physical fitness among college students in China [1, 2], posing a threat to their learning performance and daily life.

Against this backdrop, it is important for every college to step up the monitoring and performance evaluation of PE. However, the traditional data processing methods cannot effectively evaluate the physical fitness of college students, due to the sheer volume of the relevant data. As a result, PE experts are faced with the new task to mine the correlations between fitness indices of college PE.

If the above task is solved effectively, it will be possible to evaluate the physical fitness of college students based on the data of PE. Efficient and effective evaluation of the PE data on college students helps colleges to formulate better a PE curriculum to enhance the physical fitness of their students, and enables decision-makers to optimize their decisions concerning PE implementation.

To identify the key factors affecting the physical fitness of college students, this paper mines the correlations between physical fitness indices out of mass PE data, with the help of the association rule mining algorithm called the Apriori algorithm, and also identifies the most important indices of physical fitness. Besides, the Apriori algorithm was improved to reduce the time complexity by the divide and rule principle. Finally, the indices identified by the improved algorithm were compared with those obtained by the original algorithm. The research findings provide a good reference for improving the physical fitness of college students.

2 Literature Review

In the field of education, many colleges rely on data mining to assist students in course selection, evaluate their development, innovation ability, and entrepreneurship, and predict their post-graduation growth.

Joksimović et al. [3] applied data mining to manage student achievements, laying a solid basis for improving teaching quality. Sun and Bin [4] mined some hidden information that affect the computer grade examination results of a college, and provided an important guide for computer teaching. Jen et al. [5] improved the Apriori algorithm, and combined it with preference information to mine and analyze student scores. Kumar et al. [6] explored the correlations between academic performance and habits, and thereby prewarned the abnormal situation of students.

Through fuzzy k-means clustering, Zhen [7] captured deep knowledge from teaching data samples, and evaluated the academic performance of students. Ishman et al. [8] explored the cumulative data on teachers and student evaluations with the Apriori algorithm, and revealed the frequent problems in the teaching process. Drawing on the data mining theory of rough set, Tican and Taspinar [9] examined teacher behaviors and teaching effect of experimental courses in colleges, and identified the factors that constrain the development of learners. Liu et al. [10] probed deep into the problems in current teaching evaluation method, conducted an experimental analysis with association rule algorithm, rough set algorithm, etc., and constructed a multi-element teaching theory.

Based on data mining, Ji et al. [11] set up a personalized distance education system, and provided personalized distance education services, which cater to the needs of each student. Luo et al. [12] derived an overall development model from the historical performance of students, and created a mining model to predict their future performance and prewarn performance declines. With the aid of association rule algorithm, Yu [13] acquired the data related to student learning, identified the association rules that influence learning effect, and pinpointed the factors with a strong correlation with learning effect.

Based on the consumption data of campus smart card, Liu et al. [14] adopted statistics and social network research methods to obtain the features of student communication behavior. By association rule algorithm, Scherer et al. [15] dug into the information of student achievements, quantified the association rules between courses, and forecasted the number of students who could not graduate normally. Fan et al. [16]

explored the techniques of association rule mining, proposed the optimized Apriori algorithm, and applied it to find the correlations between the main factors affecting student performance.

Aher and Lobo [17] combined clustering algorithm and association rule algorithm to mine and analyze course scores, and thus discover the factors affecting student performance. Kotsiantis et al. [18] improved the incremental mining algorithm based on the degree of learning interest, and then obtained reliable and reasonable association rules of course structure, providing a theoretical basis for course recommendation. Zhou et al. [19] presented a personalized course recommendation algorithm, and employed it to recommend computer courses; the results show that their algorithm can recommend courses as per the needs of different students, after analyzing the courses associated with the promotion of student scores.

3 Correlation Analysis Based on Apriori Algorithm

One of the latest trends in PE research is to mine the hidden relations between physical fitness indices out of a huge number of PE data on college students. This paper firstly constructs and tests a correlation analysis model of PE data based on the Apriori algorithm.

The Apriori algorithm aims to mine association rules with the help of frequent itemsets [20-24]. The basic idea of the algorithm is: first, find all frequent itemsets, whose support is greater than or equal to the predefined minimum support; then, identify the strong association rules from the frequent itemsets, which must satisfy both minimum support and minimum confidence; next, generate the rules that only contain the items in the corresponding set; after that, retain only the rules whose confidence is greater than the user-defined minimum confidence. The specific flow of the Apriori algorithm is shown in Figure 1.

Frequent itemsets refer to the itemsets with a support higher than the minimum support threshold. The mining of frequent itemsets is the cornerstone of data mining. Association rules can be mined from these itemsets.

Association rule mining, a rule-based machine learning approach, seeks for relations of interest in a large database. The purpose of association rule mining is to identify the strong rules in the database, with the aid of some indices.

The strength of an association rule can be measured by its support and confidence. Support means the frequency of an itemset or rule appearing in all items. The support $\sigma(A)$ of itemset A can be calculated by:

$$s(A) = \sigma(A)/N \tag{1}$$

where, N is the total number of itemsets.

The support of rule $A \rightarrow B$ can be defined as:

$$s(A \rightarrow B) = \sigma(A \cup B)/N \tag{2}$$

Thus, the support of A ==> B indicates the probability that itemsets A and B appear at the same time:

$$\text{support}(A ==> B) = P(A \cup B) \tag{3}$$

Confidence refers to the probability of itemset B appearing at the same time with itemset A in a transaction T:

$$\text{confidence}(A ==> B) = P(A \cap B) \tag{4}$$

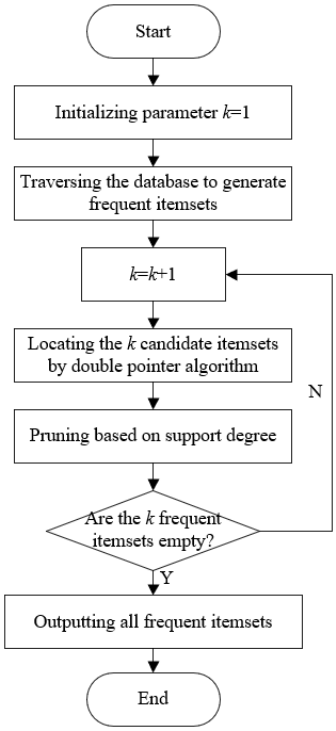

Fig. 1. Flow of the Apriori algorithm

The original data are the PE information of 3,787 college students of class 2016, collected from the PE test centre of a college. The dataset covers the gender, height, weight, and vital capacity of each student, as well as several PE indices required by the National Student Health Standards, namely, push-up test result, sit-up test result, sit-and-reach test result, standing long jump test result, 50m dash test result, and 1,000m/800m running test result.

Based on the physical indices and the principle of the Apriori algorithm, the minimum support and minimum confidence were set to 20% and 70%, respectively. After data pre-processing, the Apriori algorithm was adopted to handle the data. Firstly, the list of itemsets was scanned through to remove the itemsets that violate the minimum support threshold. The remaining itemsets were merged into a two-element set of items. Next, the transaction records were scanned again to delete the itemsets that violate the minimum support threshold. The above operations were repeated until all itemsets were removed.

Table 1 shows the correlations between the indices obtained by the model from the actual data. The following conclusions can be drawn from the table:

1. The support and confidence of the correlations between index 2 and indices 6 and 7 were greater than 0.2 and 0.7, respectively, indicating that weight is correlated with push-up test result and sit-up test result.
2. Index 2 is not significantly correlated with indices 8 or 9.
3. The support and confidence of the correlations between index 3 and indices 4-8 were above the minimums of 0.2 and 0.7, respectively, suggesting that vital capacity is correlated with push-up test result, sit-up test result, sit-and-reach test result, standing long jump test result, and 50m dash test result.
4. Indices 3 and 9 are closely correlated, i.e., vital capacity is strongly correlated with 1,000m/800m running test result. This is in line with the actual data on fitness.
5. Index 4 is not necessarily correlated with index 5.
6. Index 4 is strongly correlated with indices 6-9, that is, sit-and-reach test result is or related with push-up test result, sit-up test result, 50m dash test result, and 1,000m/800m running test result.

Table 1. Correlations between indices obtained by the Apriori algorithm

No.	Index 1	Index 2	Minimum confidence	Minimum support
1	4	6	0.912	0.49
2	4	7	0.925	0.39
3	4	8	0.896	0.31
4	4	9	1	0.35
5	5	6	0.91	0.34
6	5	7	0.922	0.41
7	5	8	1	0.32
8	5	9	1	0.26
9	6	7	0.916	0.38
10	6	8	0.917	0.38
11	6	9	1	0.32
12	7	8	1	0.38
13	7	9	0.912	0.38
14	8	9	0.965	0.32
15	2	6	0.865	0.28
16	2	7	0.912	0.31
17	2	8	0.723	0.18
18	2	9	0.825	0.16
19	3	4	0.762	0.26
20	3	5	0.798	0.25
21	3	6	0.912	0.32
22	3	7	0.887	0.36
23	3	8	0.962	0.34
24	3	9	1	0.45
25	4	5	0.554	0.16

The above results demonstrate the importance of endurance in the PE, for speed, flexibility, and vital capacity are important indices of physical fitness.

4 Correlation Analysis Based on Improved Apriori Algorithm

Despite its good mining effect, the Apriori algorithm has a low efficiency. It consumes too much time to process a large dataset, owing to the need to traverse the data multiple times. To lower the time complexity, this section improves the Apriori algorithm, and applies the improved method to verify the factors affecting the physical fitness of college students.

The Apriori algorithm was improved by adopting the divide and rule principle. Following this principle, the transaction dataset was recursively split into several small conditional transaction datasets, facilitating the mining of frequent itemsets. After the improvement, the algorithm only needs to traverse the transaction dataset twice, and eliminates the need for generating candidate sets. In this way, it is one order of magnitude faster than the original algorithm.

The improved algorithm first sorts the items in the transaction dataset by support, inserts the items of each transaction in descending order to a frequent pattern (FP) tree, with null as the root node, and records the support at each node. Figure 2 provides an example of the FP-tree.

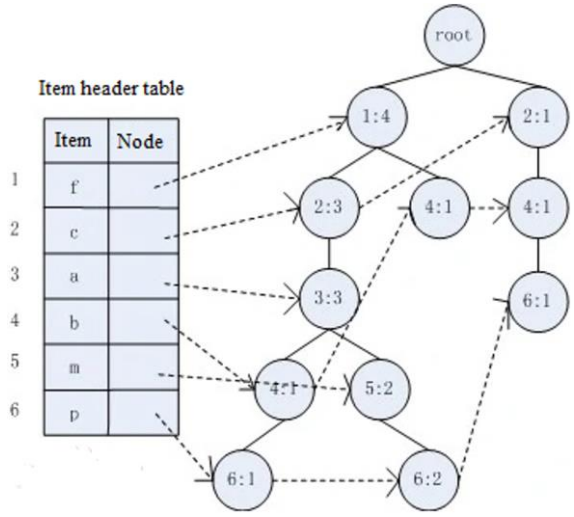

Fig. 2. An example of the FP-tree

The FP-tree was constructed and projected iteratively. For each frequent item, a conditional projection database and an FP-tree were constructed. This process was repeated for every newly constructed FP-tree, until the tree was empty or contained only one path. If the tree was empty, its prefix was taken as the FP; if the three contained only one path, the FP was obtained by enumerating all possible combinations and connections with the prefix of the tree. Figure 3 gives the flow of the improved Apriori algorithm.

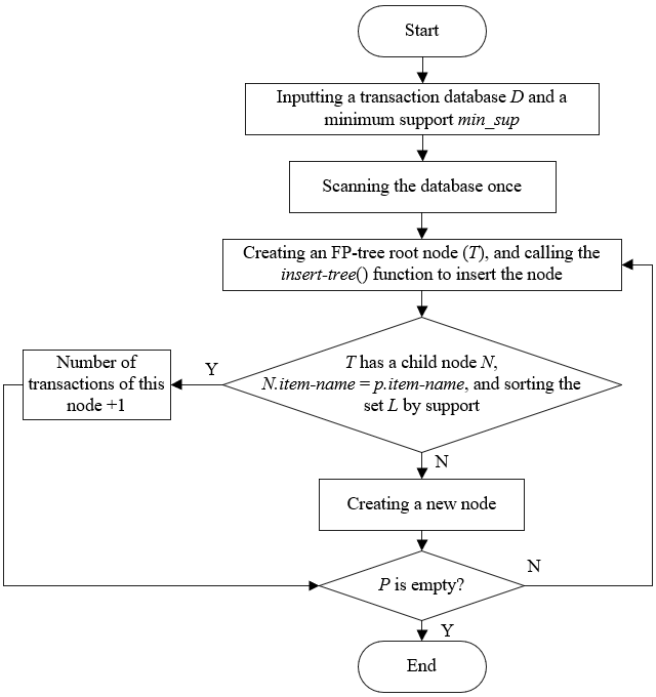

Fig. 3. Flow of the improved Apriori algorithm

Next, the improved Apriori algorithm was applied to filter the original data and analyze the correlations between physical fitness indices.

The correlations of the nine rules between all physical fitness indices are presented in Table 2. The following can be inferred from the data in this table:

1. Gender (index 1) was associated with most of the other indices. The support and confidence of the correlation between this index and other indices were greater than 0.2 and 0.7, respectively. The association rules indicate that the correlations are not very strong.
2. The support and confidence of the correlation between index 2 and indices 1 and 3-7. Thus, height is correlated with weight, vital capacity, push-up test result, sit-up test result, sit-and-reach test result, and standing long jump test result. The support between index 2 and index 8 was greater than 0.2, but the confidence between them was smaller than 0.7. This means height is not strongly correlated with 50m dash test result. Moreover, indices 2 and 9 have a certain correlation, suggesting that height is associated with 1,000m/800m running test result.
3. The support and confidence of the correlation between index 3 and indices 4 and 5 were greater than 0.2 and 0.7, respectively. Hence, weight is correlated with vital capacity and sit-and-reach test result. Meanwhile, support and confidence of the correlation between index 3 and indices 7 and 8 were also greater than 0.2 and 0.7, respectively. This means weight is correlated with push-up test result, and sit-up test result.

Table 2. . Correlations between indices obtained by the improved Apriori algorithm

No.	Index 1	Index 2	Minimum confidence	Minimum support
1	1	2	< 0.7	0.32
2	1	3	< 0.7	0.25
3	1	4	< 0.7	0.31
4	1	5	< 0.7	0.32
5	1	6	< 0.7	0.26
6	1	7	< 0.7	0.41
7	1	8	< 0.7	0.32
8	1	9	< 0.7	0.32
9	2	3	0.756	0.26
10	2	4	1	0.45
11	2	5	1	0.52
12	2	6	1	0.42
13	2	7	0.952	0.52
14	2	8	< 0.7	0.22
15	2	9	1	0.85
16	3	4	0.925	0.32
17	3	5	0.946	0.36
18	3	6	0.925	0.44
19	3	7	0.902	0.46
20	3	8	0.875	0.45
21	3	9	0.796	0.26
22	4	5	0.966	0.34
23	4	6	0.903	0.46
24	4	7	0.924	0.38
25	4	8	0.875	0.36
26	4	9	0.795	0.30
27	5	6	0.912	0.35
28	5	7	0.917	0.34
29	5	8	1	0.32
30	5	9	1	0.26
31	6	7	0.916	0.44
32	6	8	0.912	0.36
33	6	9	1	0.32
34	7	8	0.906	0.36
35	7	9	1	0.36
36	8	9	0.986	0.32

4. The support and confidence of the correlation between index 4 and indices 5-8 were greater than the minimum support and confidence of 0.2 and 0.7, respectively. As a result, vital capacity is correlated with push-up test result, sit-up test result, sit-and-reach test result, standing long jump test result, and 50m dash test result. Hence, the physical fitness of college student is greatly affected by speed, flexibility, and vital capacity.

In the analysis of association rules, if the number of variables in the analysis problem is large and there is a complex relationship between variables, principal component analysis can be used to simplify the data set, which will help to analyse and model the problem.

Principal component analysis (PCA) is a data dimension reduction method based on sample statistics and covariance minimization theory. Its basic idea is to simplify the complex relationship between independent variables. Its method is to transform multiple variables into a few comprehensive variables through matrix decomposition. Each principal component is a linear combination of the original variables, and each principal component is not related to each other. Therefore, these principal components can reflect most of the information of the original variables, and the information contained is not overlapping.

Assuming that there are T original independent variables, PCA extracts information from all independent variables and integrates them into T variables with different importance. Suppose that the original sample has T attribute values, and there are n samples in total, then we can get a $n \times T$ matrix as follows:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1T} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nT} \end{bmatrix}$$

After standardization, the correlation coefficient matrix of the data is:

$$R = \begin{bmatrix} r_{11} & \cdots & r_{1T} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nT} \end{bmatrix}$$

The eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_T)$ and corresponding eigenvectors of the correlation coefficient matrix R are calculated. The important principal component is chosen and the principal component expression is written. Finally, the original data after standardization is replaced by the expression of principal component, and the scores of each principal component can be obtained. The specific forms are as follows:

$$\begin{bmatrix} F_{11} & \cdots & F_{1k} \\ \vdots & \ddots & \vdots \\ F_{n1} & \cdots & F_{nk} \end{bmatrix} F_{ij} = a_{j1}x_{i1} + a_{j2}x_{i2} + \cdots . + a_{jT}x_{iT}$$

After principal component analysis of our dataset, according to the score of principal component, the indexes 1, 2, 5, 6, 7, 8 and 9 are reserved.

To sum up, the physical fitness of college students is significantly influenced by speed, endurance, flexibility, and vital capacity, but not greatly affected by height and weight.

Further, the improved Apriori algorithm was compared with the original algorithm in terms of performance. Figure 4 contrasts the execution time of the two algorithms under different minimum supports. It can be seen that the improved Apriori algorithm had a much lower time complexity than the traditional algorithm. Therefore, the proposed algorithm is more suitable for analysing the mass PE data on college students.

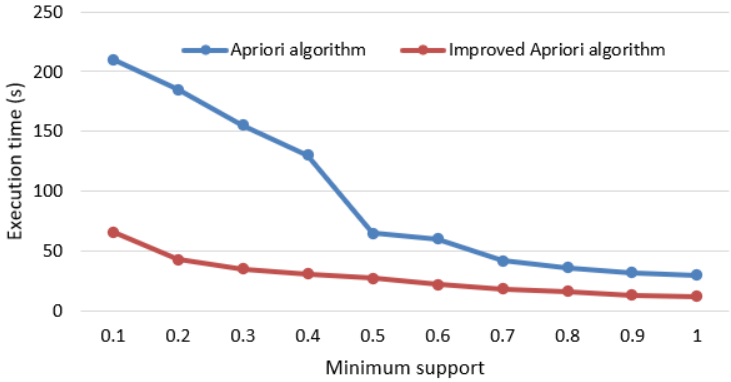

Fig. 4. The execution time of the two algorithms under different minimum supports

5 Conclusion

This paper mainly analyses the PE data on college students with the original and improved Apriori algorithms, and mined the correlations between the physical fitness indices. The results show that the height and weight of college students in China are increasing, but the overall physical fitness is on the decline; the physical fitness of college students is significantly influenced by speed, endurance, flexibility, and vital capacity, but not greatly affected by height and weight; the physical fitness of college students can be improved by providing scientific trainings on speed, endurance, flexibility, and vital capacity. The research results provide an important guide for PE teachers and designers of PE curriculum.

6 References

[1] Wang, Y., Sun, C., Guo, Y. (2020). A Multi-Attribute Fuzzy Evaluation Model for the Teaching Quality of Physical Education in Colleges and Its Implementation Strategies, International Journal of Emerging Technologies in Learning, 16(2): 159-172. https://doi.org/10.3991/ijet.v16i02.19725

[2] Yang, B. (2020). Training Model of Innovative Talents in Physical Education Major, International Journal of Emerging Technologies in Learning, 15(24): 176-190. https://doi.org/10.3991/ijet.v15i24.19035

[3] Joksimović, S., Gašević, D., Loughin, T. M., Kovanović, V., Hatala, M. (2015). Learning at distance: Effects of interaction traces on academic achievement. Computers & Education, 87: 204-217. https://doi.org/10.1016/j.compedu.2015.07.002

[4] Sun, G., Bin, S. (2018). Topic Interaction Model Based on Local Community Detection in MOOC Discussion Forums and its Teaching. Educational Sciences: Theory & Practice, 18(6). 2922-2931. https://doi.org/10.12738/estp.2018.6.191

[5] Jen, A., Webb, E.M., Ahearn, B., Naeger, D.M. (2016). Lecture evaluations by medical students: concepts that correlate with scores. Journal of the American College of Radiology, 13(1): 72-76. https://doi.org/10.1016/j.jacr.2015.06.025

[6] Kumar, A., Puranik, M.P., Sowmya, K.R. (2016). Association between dental students' emotional intelligence and academic performance: A study at six dental colleges in India. Journal of Dental Education, 80(5): 526-532. https://doi.org/10.1002/j.0022-0337.2016.80.5.tb06112.x

[7] Zhen, C. (2021). Using big data fuzzy K-means clustering and information fusion algorithm in English teaching ability evaluation. Complexity. https://doi.org/10.1155/2021/5554444

[8] Ishman, S.L., Stewart, C.M., Senser, E., Stewart, R.W., Stanley, J., Stierer, K.D., Kern, D.E. (2015). Qualitative synthesis and systematic review of otolaryngology in undergraduate medical education. The Laryngoscope, 125(12): 2695-2708. https://doi.org/10.1002/lary.25350

[9] Tican, C., Taspinar, M. (2015). The effects of reflective thinking-based teaching activities on pre-service teachers' reflective thinking skills, critical thinking skills, democratic attitudes, and academic achievement. The Anthropologist, 20(1-2): 111-120. https://doi.org/10.1080/09720073.2015.11891730

[10] Liu Y., Zhou X., Zhang Z., Xu X., Xu X. (2020). Analysis of Multi-Element Blended Course Teaching and Learning Mode Based on Student-Centered Concept under the Perspective of "Internet+". Journal of Business Theory and Practice, 8(1): 13-18.

[11] Ji C., Fan M. (2011). The application of web data mining in personalized modern distance education. Energy Procedia, (13): 714-720.

[12] Luo W., Paris S.G., Hogan D., Luo Z. (2011). Do performance goals promote learning? A pattern analysis of Singapore students' achievement goals. Contemporary educational psychology, 36(2): 165-176. https://doi.org/10.1016/j.cedpsych.2011.02.003

[13] Yu H. (2020). Online teaching quality evaluation based on emotion recognition and improved AprioriTid algorithm. Journal of Intelligent & Fuzzy Systems. https://doi.org/10.3233/JIFS-189534

[14] Liu Y., Yang H., Sun G.X., Bin S. (2020). Collaborative filtering recommendation algorithm based on multi-relationship social network. Ingénierie des Systèmes d'Information, 25(3): 359-364. https://doi.org/10.18280/isi.250310

[15] Scherer, R., Rohatgi A., Hatlevik O. E. (2017). Students' profiles of ICT use: Identification, determinants, and relations to achievement in a computer and information literacy test. Computers in Human Behavior, 70: 486-499. https://doi.org/10.1016/j.chb.2017.01.034

[16] Fan H., Xu J., Cai Z., He J., Fan X. (2017). Homework and students' achievement in math and science: A 30-year meta-analysis, 1986–2015. Educational Research Review, 20: 35-54. https://doi.org/10.1016/j.edurev.2016.11.003

[17] Aher S.B., Lobo L.M.R.J. (2012). A comparative study for selecting the best unsupervised learning algorithm in e-learning system. International Journal of Computer Applications, 41(3), 27-34.

[18] Kotsiantis S., Patriarcheas K., Xenos M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. Knowledge-Based Systems, 23(6): 529-535. https://doi.org/10.1016/j.knosys.2010.03.010

[19] Zhou W. Han W. (2019). Personalized recommendation via user preference matching. Information Processing & Management, 56(3): 955-968. https://doi.org/10.1016/j.ipm.2019.02.002

[20] Bin S., Sun G. (2020). Optimal energy resources allocation method of wireless sensor networks for intelligent railway systems. Sensors, 20(2): 482. https://doi.org/10.3390/s20020482

[21] Li Z.Q., Xu C.J., Liu C. (2019). Frequent subtree mining algorithm for ribonucleic acid topological pattern. Revue d'Intelligence Artificielle, 33(1): 75-80. https://doi.org/10.18280/ria.330113

[22] Song Y.F. (2020). A correlation analysis model of human factors in mine accidents based on apriori algorithm. International Journal of Safety and Security Engineering, 10(3): 409-415. https://doi.org/10.18280/ijsse.100314

[23] Zhou N., Zhang Z.F., Li J. (2020). Analysis on course scores of learners of online teaching platforms based on data mining. Ingénierie des Systèmes d'Information, 25(5): 609-617. https://doi.org/10.18280/isi.250508

[24] Mu W.Z. (2019). A big data-based prediction model for purchase decisions of consumers on cross-border e-commerce platforms. Journal Européen des Systèmes Automatisés, 52(4): 363-368. https://doi.org/10.18280/jesa.520405

7 Author

Tao Pan was graduated from School of Physical Education, Guizhou Normal University, he is a lecture in Guizhou University of Commerce, and his main research direction is youth sports. E-mail: 201510468@gzcc.edu.cn

Article submitted 2021-01-30. Resubmitted 2021-03-01. Final acceptance 2021-03-10. Final version published as submitted by the authors.