EasyChair Preprint

№ 4522

An Apriori-Algorithm-Based Analysis Method on Physical Fitness Test Data for College Students

Shupo Nan and Maojian Chen

November 7, 2020

An Apriori-Algorithm-Based Analysis Method on Physical Fitness Test Data for College Students

Shupo Nan [1] and Maojian Chen[2]

[1] Xinlian College of Henan Normal University, Henan 450000, China
[2] School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
2111008@stu.xlxy.edu.cn

Abstract. Since being required to carry out physical fitness tests for students, colleges and universities have accumulated a huge amount of data to deal with annually. However, it is almost impossible to discover the potential relationship among the indicators in physical fitness tests by adopting traditional data processing and analysis methods. Hence, how to identify potential information from the test data and seek corresponding solutions has become the key to improving students' physical fitness and teaching quality. The Apriori algorithm as a classic approach explores the relationship among a huge amount of data. For the purpose of improving computational performance in addressing a large number of redundant candidate sets through the use of the traditional Apriori algorithm, a method is accordingly developed in this paper. It first determines the attribute significance in test data by using decision tree classifier. Then, it deletes the attributes with lower significance in the test data, reduces the number of scans of data sets, decreases the number of candidate sets, and generates the corresponding association rule. Finally, an experiments conducted to verify the effectiveness of the proposed algorithm.

Keywords: Apriori Algorithm, Attribute Significance, Physical Fitness Test, Association Rule.

1 Introduction

Nowadays, a huge amount of physical fitness test data collected by colleges and universities has been accumulated in China. Faced with this mass of data, it is difficult to carry out scientific evaluation and analysis by using traditional data processing methods. In addition, traditional methods have limited ability to mine the potential knowledge in the data and difficulty in proposing more valuable solutions for improving the physical fitness of college students. Therefore, the value of the data is far from being exploited. After uploading the data to the physical fitness test platform, generally, colleges and universities only make data backup and simple queries instead of in-depth analysis to obtain valuable teaching information. Inevitably, this will cause a waste of information resources.

As a classic algorithm in the pattern mining field, the association analysis is mainly to explore useful association rules from a large number of multivariate data sets by

looking for rules that can best explain the relationship among data variables, so it is an approach of discovering relationship among multiple types of data. Common association algorithms include Apriori, FP-Growth, PrefixSpan, and SPADE. At present, research on the data relationship from the perspective of association rules is rarely found in literatures. Zhang et. al [1] analyzed the relationship among each individual test indicators directly with association rules, but no effective improvement measures were taken to address the disadvantages of association rules. Liu et. al [2] proposed an array-based Apriori algorithm to reduce the redundancy of candidate frequent item sets and analyze the data relationship. The Apriori algorithm was improved by compressing the transaction database and reducing the invalid comparisons in the join operation and analyzed the relationship among course scores through the improved algorithm [3]. Miao [4] conducted a correlation mining over the experiment scores and those factors by the Apriori algorithm, established strong association rules, and discovered key factors that affected teaching quality.

Apriori algorithm is a very classic algorithm for discovering association rules, with extensive practical application. Scholars have developed various optimization methods for the Apriori algorithm. a heterogeneous network model was proposed to examine the complex nonlinear relationship between drugs and diseases and to predict the correlation between them [5]. Han et. al [6] limited the sequence search and spatial expansion by means of the projected sequence data set, which greatly reduced the number of candidate sub-sequences generated by mining. The authors Han .et al narrowed the search space by continuously expanding the existing frequent sequences to obtain new frequent sequences [7]. Wang et. al [8] formalized the association rule mining with graph pattern and decomposed it into two sub-problems - frequent pattern mining and rule generation. A method was proposed to find rules describing abnormal event patterns from multivariate time series, which first converted the distribution of time series into a normal distribution, and then mined and found frequent rules among the abnormal event symbols by using association rules [9].

Based on the association rule of desert grassland data diversity, Du et. al proposed an improved Apriori algorithm, and enhanced its efficiency by increasing the judgment data set, decreasing the generation of candidate item sets and greatly reducing time consumption [10]. A parallelization improvement of Apriori algorithm based on Mapreduce was made to reduced the number of candidate sets in the iteration process and save storage space [11]. In order to improve algorithm efficiency, Xu et. al [12] proposed an improved Apriori algorithm for audit log association mining. Based on the existing optimization ideas of Apriori algorithm, Zeng et. al [13] proposed an improved association rule mining algorithm - MIFP-Apriori algorithm in combination of matrix, improvement of frequent pattern tree and frequency optimization strategy. By scanning the transaction database to establish a 0-1 transaction matrix, Li et. al [14] obtained the weighted support and confidence, which could effectively extract hidden valuable items. An association rule discovery algorithm was proposed based on fuzzy association rule mining [15]. Shao et. al [16] adopted the text classification algorithm in text analysis technology to replace manual classification of problems into concepts and generated the concept map in combination with the association rule mining method. The concept map expressed association rules and closeness degree among concepts

through association pairs and association degree, which could clearly show the structural association between concepts. A one-way tree-based method was proposed for extracting rare association rules when inserting new data into the original database. This method could generate complete frequent and rare patterns without re-scanning the updated database set [17].

There are two focuses in the current study on Aprior algorism. First, to study the generation of candidate item sets and frequent item sets, by which the researchers try to reduce the generation of the candidate sets so as to boost the implementation efficiency of Aprior algorism. Second, to conduct dimension-reduced processing of data, in the process of which high-dimension data is processed by two main methods, namely, Principal Component Analysis (PCA) and Multidimensional Scaling (MDS). PCA aims to get fewer data dimensions by mapping high-dimension data to low-dimension space by means of a certain linear projection and at the same time hoping that the data variance on the projected dimension is the biggest. MDS is purposed to construct proper low-dimension space through similarities of paired samples so that the distances between the samples in both low-dimension space and high-dimension space can be as consistent as possible[18].

This paper tries to pre-process data by Decision Tree Classifier (DTC), which, as a case-based inductive classification algorithm, adopts the top-to-bottom incursive algorithm to compare the attribute values on the middle node of the decision tree. Then the downward branch is selected based on corresponding values so as to get the conclusion of the classification at the end of the decision tree. As one of the steps in the classification by DTC is to judge the importance degree of each attribute value and, at the same time, the attributes of the data of the physical fitness tests used in this paper are multiple, it is recommended to first judge the importance degree of the test data and then to analyze them.

This paper does not classify data by the decision tree. Rather, it judges the importance degrees of data attributes by DTC, deletes the attributes that are less important, and then conducts association rule analysis, resulting in higher generation efficiency of association rules.

The main contributions of this paper are as follows. First, to effectively pre-process data of the physical fitness tests in the generation of association rules, DTC is adopted to determine how important the data attributes are and to reduce candidate item sets, resulting in the improvement of the implement of Apriori algorithm. Second, the verifications are conducted in the data analysis of physical fitness tests of college students through the combination of theory and practice. Thus, the feasibility of the method developed in this paper is verified and the valuable information is extracted from the generated association rules, providing theoretical support for the improvement of fitness of college students.

2 Overview

2.1 Basic Idea

The basic idea of the Apriori algorithm was to first scan the data set, count the number of items in it and the occurrence times of each different item set, and then obtain all frequent 1 item sets according to the minimum support, namely, L1; secondly, identify frequent 2 item set L2 by using L1, and repeat the process until no more new frequent sets were identified. In summary, there were two steps: the connecting step and the pruning step.

Connecting Step: in order to find L_k (frequent k item set), connect L_{k-1} to itself and denote *k* item as C_k. Yet in pairwise self-connection, only item sets that differed from the last item could be connected.

Pruning Step: obviously, some of the members of the superset were frequent, and some were not. However, all frequent k item sets were included. If the data sets were scanned, and the item set of each member was counted to find the members whose value was not less than minimum support to form Lk (i.e., frequent k item set), the number of members in this calculation can be significant, which resulted in an increased amount of calculation. As a result, when the algorithm was put into application, the anti-monotonicity principle would be used. In the principle, whether k-1 item sets were frequent for all members of the candidate set was determined, and as long as there was a member which was not a k-1 frequent item set, this member could not become a k-item frequent set and could be deleted. The remaining members constituted the pruned candidate set, and then the data set scan method was applied to count the members to determine the members whose count value was not less than minimum support and to form the collection of frequent item sets [18].

2.2 Disadvantages

(1) Generation of a Large Number of Candidate Item Sets

In the process of generating candidate k item set C_k from *k*-1 item set L_{k-1}, many candidate item sets might be generated, and the speed was getting increasingly faster, which put pressure on the running time and memory space of the algorithm [18].

(2) Multiple Scans of the Transaction Database

Each time a candidate item set was obtained by the Apriori algorithm, the database was searched. Therefore, when a frequent k item set was generated, the database would be scanned k times [18].

(3) Hidden Useful Information in the Candidate Set with Lower Degree of Support

During the implementation of the Apriori algorithm, candidate sets below the minimum support would be deleted. But the deleted candidate sets might also contain interesting and meaningful information. Therefore, if the degree of support was used as the sole criterion, a lot of useful information and association rules could be lost, and it was impossible to fully discover useful information or the information that was of real interest.

3 Improvement of the Apriori Algorithm

The students need to take physical fitness tests every year, and each student shall test 7 items, which results in a huge amount of data. Due to the heavy workload of statistical data analysis, the test data will be sealed after being uploaded to the physical fitness test platform of the ministry. To what extent can test data reflect students' physical quality? Are there any associations among these test items? How to develop teaching plans for targeted training to improve students' physical fitness? In order to deal with these problems, an implementation plan was put forward by mining the inherent association rules with the Apriori algorithm. The complete algorithm flow chart is shown in **Fig. 1**.

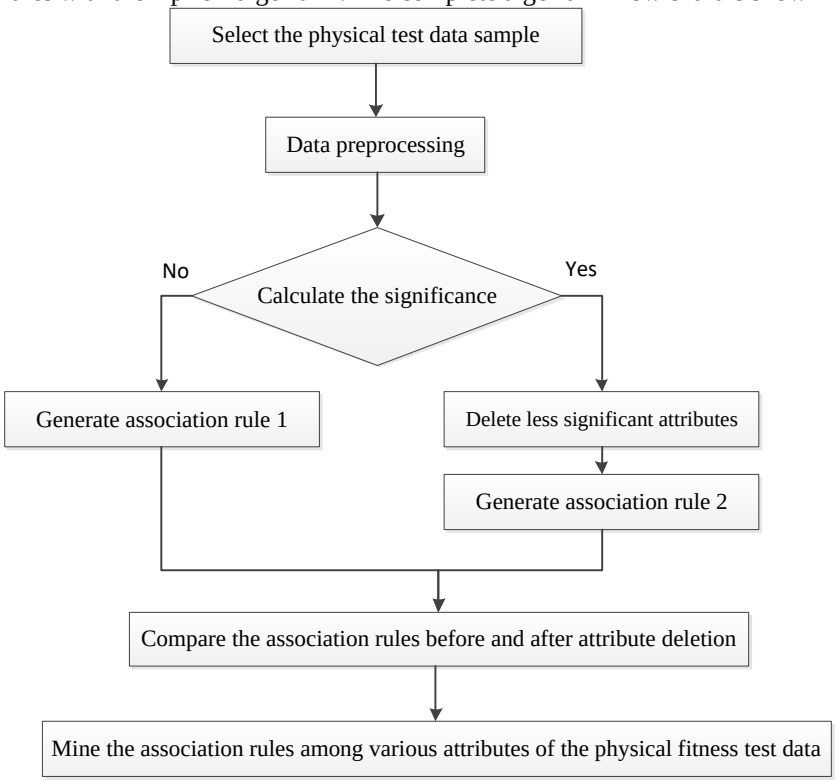

Fig. 1. The flowchart of Apriori algorithm

3.1 Selecting Data Sample for Preprocessing

The input format required by the Apriori algorithm is a 0,1 matrix, so the data needs to be preprocessed after the data samples are selected and determined, and the data sample should be converted into a 0,1 matrix.

3.2 Determining the Attribute Significance

Before the association rules were calculated, this study first determined the attribute significance of the test data by Decision Tree Classifier, and entered the scores of 7

test items and the final test scores through the fit function for training. In the training model, the attribute significance was obtained through the feature importance method. The higher the value, the more significant the attribute was. In order to improve the operating efficiency of the Apriori algorithm, the less significant attributes need to be deleted.

3.3 Conducting Comparative Analysis Before and After Attribute Deletion

For the two data sets obtained before and after the attribute deletion, this paper first set different minimum support and minimum confidence, and compared the obtained operating results, so as to determine the minimum support and minimum confidence of the test data., Then the paper compared the number of elements in the frequent item sets of the two data sets. Finally, it analyzed the advantages and feasibility of the proposed algorithm.

3.4 Generating the Association Rules

After minimum support and minimum confidence were determined, for the data set before and after deleting the attribute, L1, Ck-1, and Lk could be obtained, and finally the corresponding association rules in the form of A—>B were obtained.

3.5 Analyzing the Association Rule Obtained

The association rules obtained before and after the attribute deletion were compared, and the reasonableness of test data was verified and analyzed by using the attribute significance.

3.6 Mining the Roles of Association Rules in the Analysis of Test Scores

Based on the association rule generated by deleting the test data with less significant attributes, this paper first analyzed the intrinsic relationship among the data to find the association among the test data. Secondly, it analyzed the hidden information in the deleted attributes, obtained valuable information from the deleted attributes, and summarized the impact and enlightenment on PE and training.

4 Application of Apriori in Physical Fitness Test Data of College Students

4.1 Data Preparation

In this paper, a total of 475 students majoring in mechanical design and manufacturing, material forming and control engineering, musicology, visual communication design and exhibition art design of the class of 2018 in the chosen university were selected as research objects. Each student’s physical fitness test included 7 items, namely, height and weight, vital capacity, 50m running, sit and reach, standing long jump, pull-up (male)/1 minute sit-up (female) and 1000m (male)/800m (female).

There was a lot of redundancy in the original test data. Some of these attributes, such as grade number, class number, student ID, ethnic code, name, date of birth, and

home address, were not used in data mining. In order to reduce the dimension of the data, these useless data attributes could be deleted. According to the individual item standards and weights of national students' fitness published by the Ministry of Education, the test data in the original data table was converted into the form of scores. The specific method was to first combine height and weight into BMI, and then convert the test data into the percentage system. Part of the conversion results are as shown in **Table 1.**

Table 1. Physical Fitness Test Form (Percentage System)

BMI	Vital Capacity	50m Running	Standing Long Jump	Sit and Reach	1000m (Male)/800m (Female)	Pull-up (Male) /Sit-up (Female)	Final Score
100	76	100	78	66	64	64	80
80	90	66	50	80	64	10	65.5
100	72	78	68	70	68	20	70.8
100	95	100	76	70	72	40	82.25
100	72	85	78	72	74	64	79
100	90	100	72	100	76	64	87.3
100	85	100	70	80	68	68	83.15
100	70	100	68	76	50	68	76.7
100	70	95	76	85	70	64	81
100	70	100	80	68	76	40	79.5
60	40	30	50	68	62	30	48.2
80	85	100	90	100	68	50	82.35
100	80	85	78	76	66	30	75.6
80	0	74	64	70	64	20	55
80	60	85	64	76	66	40	69.2
100	74	100	68	76	62	40	76.9
80	66	90	74	85	66	50	74
100	95	100	76	70	72	40	82.25
100	72	85	78	72	74	64	79

The purpose of this experiment was to first analyze whether the individual scores of 80 and above in the fitness test data obtained were related and mutually reinforcing, and the extent to which these individual item scores influenced the final overall grade of "good" or "excellent". Therefore, this experiment was converted according to whether the test data were 80 scores or above, where "1" means Yes and "0" means No. Part of the conversion results are as shown in **Table 2**.

After analyzing the data relationship, teachers may develop targeted teaching plans and programs to continuously increase the students' test scores. The ultimate goal is to improve the physical quality of college students and lay a solid physical foundation

for cultivating college students to develop all-round ability in areas such as morals, intelligence, physical fitness, work and aesthetics.

Table 2. Physical Fitness Test Form (80 Scores or Above)

BMI	Vital Capacity	50m Running	Standing Long Jump	Sit and Reach	1000m (Male)/800m (Female)	Pull-up (Male) /Sit-up (Female)	Final Score
1	0	1	0	0	0	0	1
1	1	0	0	1	0	0	0
1	0	0	0	0	0	0	0
1	1	1	0	0	0	0	1
1	0	1	0	0	0	0	0
1	1	1	0	1	0	0	1
1	1	1	0	1	0	0	1
1	0	1	0	0	0	0	0
1	0	1	0	1	0	0	1
1	0	1	1	0	0	0	0
0	0	0	0	0	0	0	0
1	1	1	1	1	0	0	1
1	1	1	0	0	0	0	0
1	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0
1	0	1	0	0	0	0	0
1	0	1	0	0	0	0	0

4.2 Algorithm Implementation

4.2.1 Calculating the Attribute Significance and Getting a New Data Table

In this study, Decision Tree Classifier was used to calculate the attribute significance. The results of attribute significance are shown in Table 3. Based on the results, it was found that the significance level of three indicators of BMI, 1000m (male)/800m (female), and pull-up (male)/1 minute sit-up (female) was 0, and that of standing long jump was 0.0046874, which indicated that the scores of these four items had little impact on the student's overall grade of "good" or "excellent". Accordingly, the four items can be deleted during data analysis.

Table 3. The results of Attribute Significance

BMI	Vital Capacity	50m Running	Standing Long Jump	Sit and Reach	1000m (Male)/800m	Pull-up (Male)

					(Female)	/Sit-up (Female)
0	0.3345	0.07778	0.0046874	0.58302928	0	0

4.2.2 Conducting Comparative Analysis Before and After Attribute Deletion

4.2.2.1. Comparison of Different Minimum Support and Confidence Levels

In order to verify the analysis, this study took different support and confidence value for the test data before and after attribute deletion through the Apriori algorithm, and compared the results of the two obtained association rules.

When the confidence value are 0.3, 0.4, and 0.5 respectively and the support value were different, the number of association rules before and after deleting the attribute value was compared. The comparison results are shown in **Fig. 2**, **Fig. 3** and **Fig. 4**.

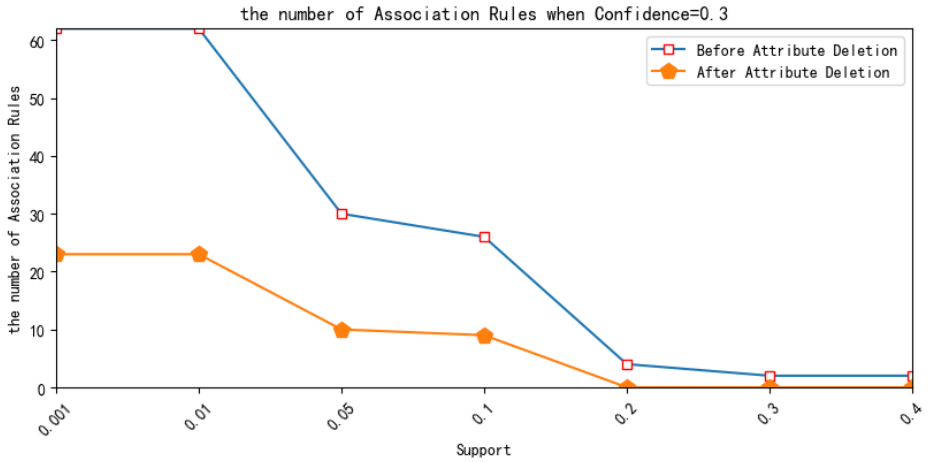

Fig. 2. The number of Association Rules when Confidence = 0.3

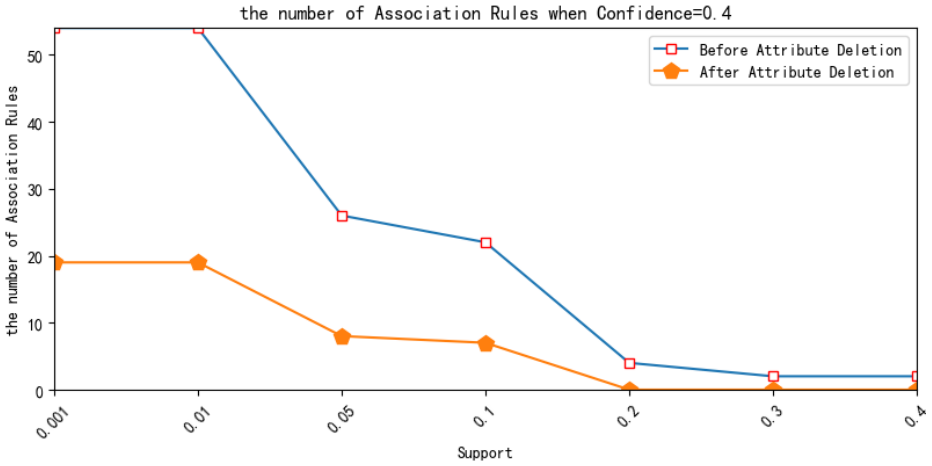

Fig. 3. The number of Association Rules when Confidence = 0.4

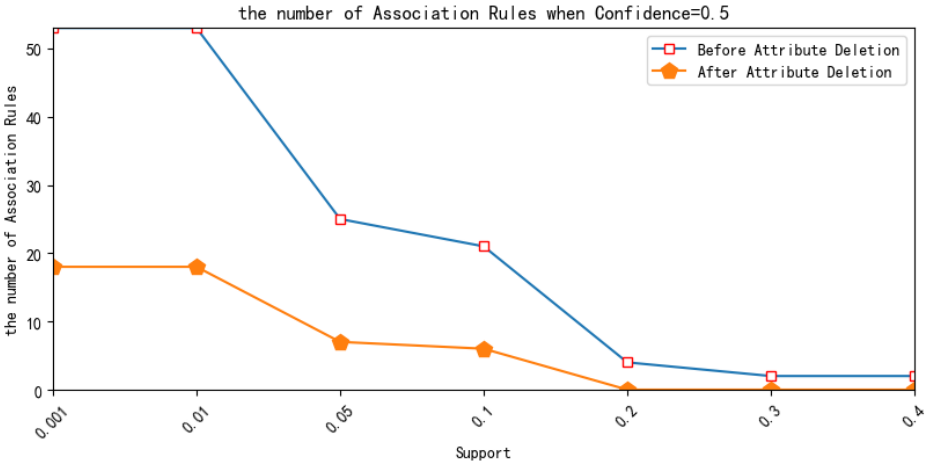

Fig. 4. The number of Association Rules when Confidence = 0.5

When the confidence value are 0.001, 0.01, 0.05 and 0.1 respectively and the support value were different, the number of association rules before and after deleting the attribute value was compared. The comparison results are shown in **Fig. 5**, **Fig. 6**, **Fig. 7** and **Fig. 8**.

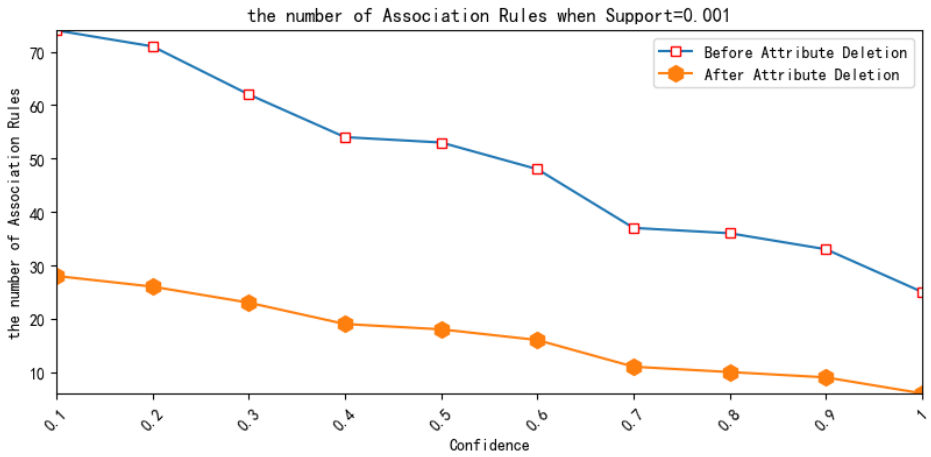

Fig. 5. The number of Association Rules when Confidence = 0.001

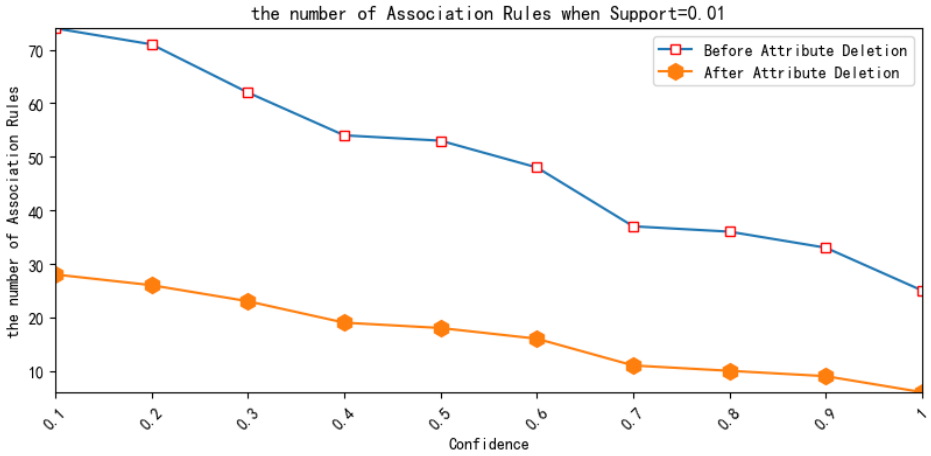

Fig. 6. The number of Association Rules when Confidence = 0.01

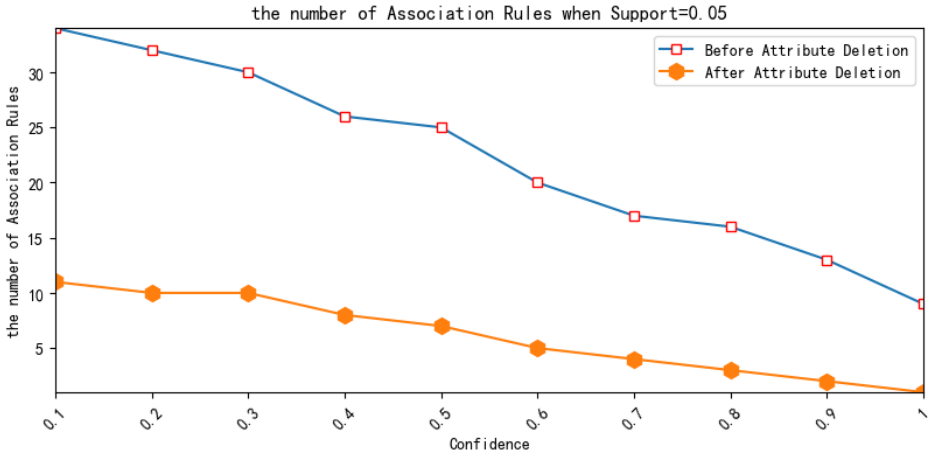

Fig. 7. The number of Association Rules when Confidence = 0.05

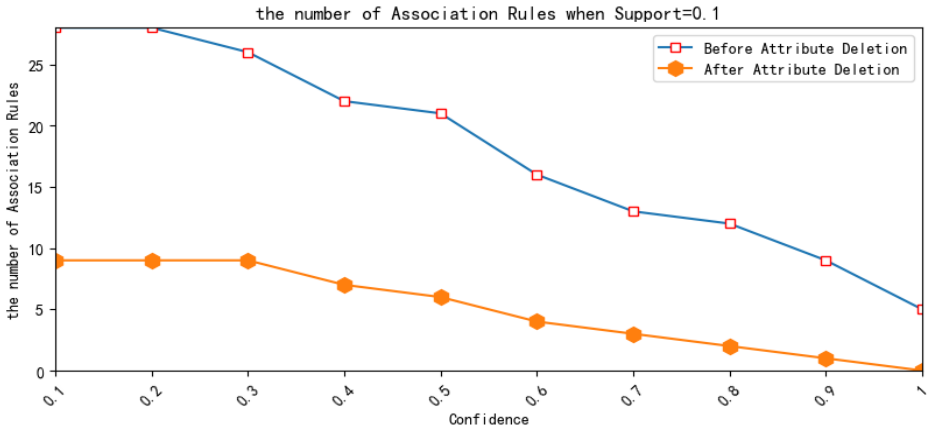

Fig. 8. The number of Association Rules when Confidence = 0.1

Through repeated comparisons and experiments, it is found that when the value of support and confidence are different, the number of association rules generated after attribute deletion is less than that of association rules generated before deletion.

4.2.2.2 Comparison of Frequent Item Sets

Given the minimum support value was 0.05, and the minimum confidence value was 0.5, after determining the minimum support and minimum confidence, this study compared the frequent 1 item set, frequent 2 item set, frequent 3 item set, and frequent

4 item set before and after the attribute deletion. The comparison results of the frequent sets before and after the attribute deletion are shown in **Fig. 9**.

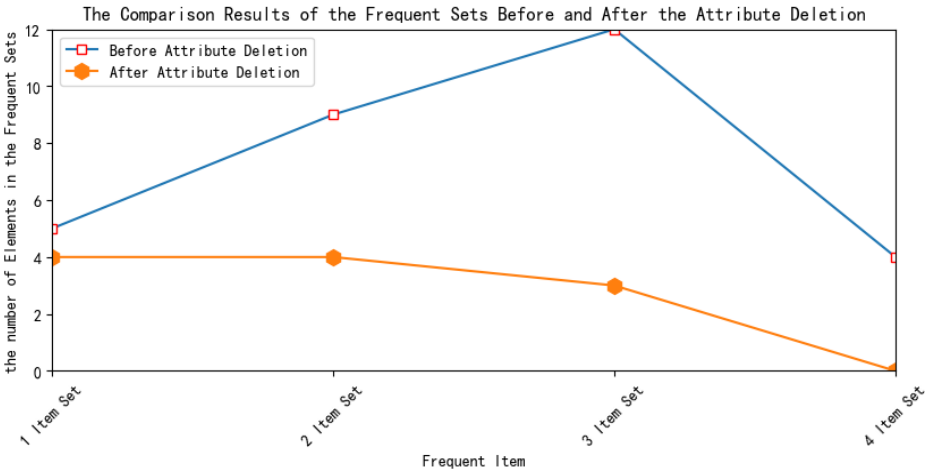

Fig. 9. The Comparison Results of the Frequent Sets Before and After the Attribute Deletion

After comparing the number of elements in the frequent sets before and after the attribute deletion, it is found that the number of elements in the frequent 1 item set, frequent 2 item set, frequent 3 item set, and frequent 4 item set after the attribute deletion is less than that of the elements before attribute deletion.

In short, the frequent k item set is generated by deleting the minimum support from the candidate k item set, and the candidate k item set is generated by connecting the frequent k-1 item set. As a result, the generation of the frequent k item set is closely correlated with the frequent k-1 item set. After the three less significant attributes are deleted and when the data set is scanned in the experiment, as the number of elements in the frequent K-1 set decreases, the time per scan of the data set declines as well. Therefore, the implementation efficiency of the program is improved, and meanwhile the feasibility of the proposed algorithm is verified by the experimental results.

4.2.3 Generating Association Rules

The minimum support value was 0.05 and the minimum confidence value was 0.5. The association rules obtained by using Apriori algorithm are as follows:

4.2.3.1 Before Deletion

In this study, t1 represents BMI, t2 vital capacity, t3 50m running, t4 standing long jump, t5 sit and reach, t6 1000m (male)/800m (female), and t7 pull-up (male)/1 minute sit-up (female). The association rule before deleting the attribute was obtained, and the results of association rules are shown in **Table 4**.

The analysis of the operation results is as follows:

First, a total of 25 association rules were obtained, among which BMI, vital capacity, standing long jump, 50m running, and sit and reach had a significant correlation. Second, 1000m (male)/800m (female) and pull-up (male)/1 minute sit-up (female) did not appear in the association rule, which indicated that there was less significant correlation between these two test scores and other scores.

4.2.3.2 After Deletion

The association rule after deleting the attribute was obtained, and the results of association rules are shown in **Table 5**.

The following conclusions can be drawn from the operation results:

First, a total of 7 association rules were obtained, among which 50m running and vital capacity, standing long jump, and sit and reach had a significant correlation. Second, BMI, 1000m (male)/800m (female) and pull-up (male)/1 minute sit-up (female) did not appear in the association rule, which indicated that there was less significant correlation between these three test scores and other scores.

Table 4. The Association Rule Before Deleting the Attribute

Association Rule	Support	Confidence
t3-->t1	0.587368	1.000000
t3-->t5-->t1	0.191579	1.000000
t2-->t3-->t1	0.176842	1.000000
t2-->t5-->t1	0.113684	1.000000
t2-->t3-->t5-->t1	0.111579	1.000000
t4-->t1	0.071579	1.000000
t4-->t3	0.071579	1.000000
t3-->t4-->t1	0.071579	1.000000
t1-->t4-->t3	0.071579	1.000000
t2-->t5-->t3	0.111579	0.981481
t1-->t2-->t5-->t3	0.111579	0.981481
t5-->t1	0.214737	0.980769
t2-->t1	0.216842	0.927928
t1-->t5-->t3	0.191579	0.892157
t5-->t3	0.191579	0.875000
t1-->t2-->t3	0.176842	0.815534
t2-->t3	0.176842	0.756757
t2-->t3-->t5	0.111579	0.630952
t1-->t2-->t3-->t5	0.111579	0.630952
t1-->t3	0.587368	0.626966
t3-->t5-->t2	0.111579	0.582418
t1-->t3-->t5-->t2	0.111579	0.582418

t1-->t5-->t2	0.113684	0.529412
t1-->t2-->t5	0.113684	0.524272
t5-->t2	0.113684	0.519231

Table 5. The Association Rule After Deleting the Attribute

Association Rule	Support	Confidence
t2—>t5—>t3	0.111579	0.981481
t5—>t3	0.191579	0.875000
t2—>t3	0.176842	0.756757
t2—>t3—>t5	0.111579	0.630952
t3—>t5—>t2	0.111579	0.582418
t5—>t2	0.113684	0.519231
t4-->t3	0.071579	1.000000

4.2.4 Conducting Comparative Analysis of Operation Results

Firstly, under the known minimum support and minimum confidence value, compared with the original Apriori algorithm, association rules with practical significance obtained by using the improved Apriori algorithm were identical, and the insignificant attributes had been removed at the same time.

Secondly, the number of valid elements in the frequent 1 item set, frequent 2 item set, frequent 3 item set, and frequent 4 item set after attribute deletion was less than that of valid elements before the attribute deletion, which reduced the operation time from L_{k-1} to L_k.

Thirdly, the association rule obtained was used to explore the connections among various test scores. For example, there was a significant correlation among the three test data indicators of vital capacity, sit and reach, and 50m running.

Fourthly, some association rules with no practical significance were removed. In the association rules obtained by the original Apriori algorithm, some of them contained the attribute BMI, yet the association rule containing BMI had no practical meaning. In contrast, in the improved Apriori algorithm, the BMI had been deleted before the association rule was generated according to the attribute significance, so all the association rules obtained were meaningful.

4.3 Experiment

In order to enhance the readability and re-verify the effectiveness of the proposed algorithm, a simulation experiment was designed in this study. In this simulation experiment, the test data could be imported on the terminal interface to reset the minimum support and minimum confidence, and the various parameters could also be reset by pressing the reset button.

Before the simulation was carried out, the data set was imported in the first place, the minimum support was set to 0.05, and the minimum confidence was 0.5. Then, the

experiment was conducted by clicking the run button. The result of simulation experiments is shown in **Fig. 10**.

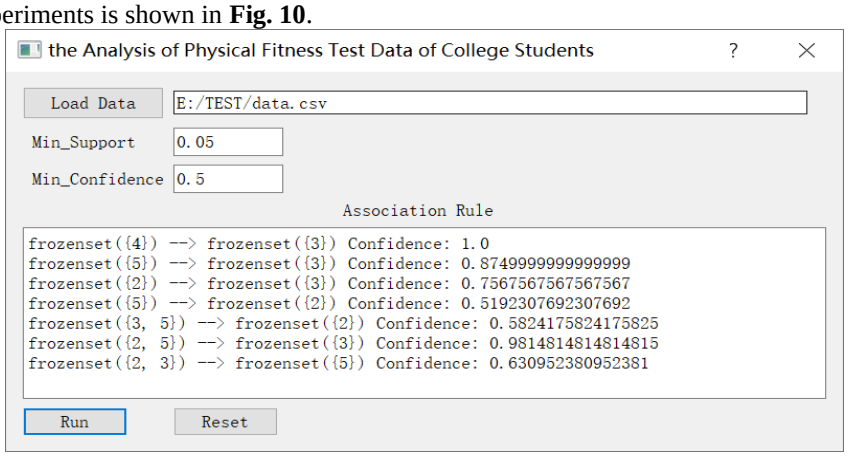

Fig. 10. The Experimental results

4.4 The Role of Association Rule in the Analysis of Physical Fitness Test Scores

Most of the students had ideal BMI, and few students were overweight or underweight, which indicates that most students have a good foundation of physical fitness. Through investigation, it is found that this is closely related to the setting of various sports courses and activities. At present, in addition to traditional sports courses such as basketball, football, volleyball, table tennis, badminton, and tai chi, the university offers aerobics and martial arts for girls, and taekwondo and other forms of second class activities for boys, which stimulates the enthusiasm of students to participate in sports activities. Besides, the university has also selected students with good physical quality, and assigned coaches to build football, table tennis, martial arts, taekwondo and aerobics teams. They have achieved many excellent results in municipal, provincial and national competitions. Due to the diverse forms of sports classes and activities, more students are motivated to participate in sports activities, which lays a foundation for maintaining good BMI.

The support of frequent 1 item set, 2 item set, and 3 item set were relatively low, which suggests that students' "good" and "excellent" rates of the test scores as well as the level of their physical fitness are relatively low. Individual test scores are far from "good" or "excellent". Therefore, the university needs to strengthen and encourage students to actively participate in physical exercise, comprehensively increase the test scores, and ultimately improve students' physical quality.

Standing long jump, 1000m (male)/800m (female) and pull-up (male)/1min sit-up (female) had the lowest support, especially the last two items whose "good" or "excellent" rate was almost 0. Therefore, the top priority in the subsequent teaching plan is to strengthen the training of those three items, which is crucial for improving the overall physical fitness score and physical quality of students.

The association rule was used to examine the relationship among the scores of various tests. In the association rule obtained by the Apriori algorithm, it is found that there is a significant correlation among three indicators of vital capacity, sit and reach and 50m running. For example, the association rule “vital capacity—>sit and reach—>50m running” represents that 11.1579% of the students’ grades are “good” or “excellent”, and in the case where the grades of vital capacity and sit and reach are “good” or “excellent”, 98.1481% of the students’ 50m running scores are also “good” or “excellent”. In addition, the association rule “vital capacity—>50m running” means that 17.6842% of the students’ scores are “good” or “excellent”, and in the case where the vital capacity scores are “good” or “excellent”, 75.6757% of the students’ 50m running scores are “good” or “excellent”. At the same time, the association rule “sit and reach—>50m running” means that 19.1579% of the students’ scores are “good” or “excellent”, and in the case where sit and reach grades are “good” or “excellent”, 87.5% of the students’ 50m running scores are “good” or “excellent”.

The association rule was used to discover intrinsic connections among individual test scores. For students with low individual test scores, in addition to the key training, the university may also offer training of related test items to improve the low individual test scores. For example, assuming that a student has low score of 50m running, the university may effectively improve the performance by training vital capacity and sit and reach.

5 Conclusion

In the process of analyzing the test data of students by the Apriori algorithm, this paper determined the attribute significance through Decision Tree Classifier, and obtained the corresponding association rules by analyzing the attributes with higher significance level. Due to the reduced number of data attributes, the number of scans of the data set and the number of candidate set generated were reduced when the Apriori algorithm was adopted. In this way, the operation efficiency of the program was improved, and the interesting and meaningful association rules were obtained.

The analysis of these association rules has the following effects. First, the intrinsic correlations that exist among test data can be discovered, which is crucial for effectively improving test scores and physical quality of students. Second, traditional strong association rules are based on the criteria of support and confidence, but the reality is that not all strong association rules are of interest or practical significance to researchers. Under such circumstances, the attribute significance can be used to screen out those strong association rules to find interesting and meaningful ones. Third, according to the attribute significance, hidden valuable information can also be found in the candidate sets with lower degree of support.

Some improvements can be achieved along this direction in the future, while using DTC. First, how to handle the vacant data and noise data in the physical fitness tests has a direct influence on the implementation of DTC. Thus, the data pre-processing can be further optimized. Second, the pruning is needed in the set-up of the decision tree to handle over-fitting caused by noise. It is significant in practice.

References

1. Chonglin Zhang, Lijuan Yu, Weibing Wu. Application of association rule data mining technology in the analysis of physical fitness test. Journal of Shanghai University of Sport 36(2), 42-44 (2012).
2. Xin Liu, Sujin Yang. Application of array-based Apriori algorithm in the analysis of physical fitness test data. Journal of Shandong University of Technology (Natural Science Edition) 25(5), 55-58 (2011).
3. Luyan Yuan, Feng Li. Application of improved association rule Apriori algorithm in course grade analysis. Chinese Journal of ICT in Education 17, 62-65 (2017).
4. Weicheng Miao, Wenjie Zhu. Analysis and research of physical experiment results based on association rule Apriori algorithm. Journal of Chifeng University (Natural Science Edition) 35(1), 14-16 (2019).
5. Xuan Ping, Hui Cui, Tonghui Shen, Nan Sheng, Tiangang Zhang. HeteroDualNet: A Dual Convolutional Neural Network With Heterogeneous Layers for Drug-Disease Association Prediction via Chou's Five-Step Rule. Frontiers in pharmacology 10, 1301 (2019).
6. J Han, J Pei, Mortazavi-Asl B. FreeSpan: Frequent pattern-projected Sequential pattern mining. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 355-359 (2000).
7. J Han, J Pei, Mortazavi-Asl B. PrefixSpan: Mining Squential Patterns Efficiently by Prefix-Projected Pattern Growth. Proc int conf Data Engineering, 215-224 (2001).
8. Xin Wang, Xu Yang, Huayi Zhan. Extending association rules with graph patterns. Expert Systems with Applications, 141 (2019).
9. Park Hoonseok, Jung Jae-Yoon. Deviant event pattern discovery from multivariate time series using symbolic aggregate approximation and association rule mining. Expert Systems with Applications 141, (2019).
10. Yongxing Du, Di Gao, Baoshan Li, et al. Application of improved apriori algorithm in desert grassland. Computer Engineering and Design 40(7), 2082-2086 (2019).
11. Jun Qin, Tianshu Hao, Qianqian Dong. Improvement of parallelization of apriori algorithm based on MapReduce. Computer Technology and Development 27(4), 64-68 (2017).
12. Kaiyong Xu, Xuerong Gong, Maocai Cheng. Audit rule association mining based on improved Apriori algorithm. Journal of Computer Applications 36(7), 1847-1851 (2016).
13. Zixian Zeng, Qingge Gong, Jun Zhang. Improved association rule mining algorithm - MIFP-Apriori algorithm. Science Technology and Engineering 19(16), 216-220 (2019).
14. Li-na Sun. 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 (2020).
15. Zhongjie Zhang, Jian Huang, Jianguo Hao, Jianxing Gong, Hao Chen.Extracting relations of crime rates through fuzzy association rules mining. Applied Intelligence 50(2), 448-467 (2020).
16. Zengzhen Shao, Yancong Li, Xiao Wang. Research on a new automatic generation algorithm of concept map based on text analysis and association rules mining. Journal of Ambient Intelligence and Humanized Computing 11(2), 539-551 (2020).
17. Borah Anindita, Nath Bhabesh. Rare association rule mining from incremental databases. Pattern Analysis and Applications 23(1), 113-134 (2020).
18. Guoyin Wan, Qun Liu, Hong Yu, et al. Big Data Mining and Application. Beijing: Tsinghua University Press, Beijing (2017).