Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 3, pp. 609–620, DOI 10.12694/scpe.v24i3.2307

APPLICATION OF IMPROVED APRIORI ALGORITHM IN INNOVATION AND ENTREPRENEURSHIP ENGINEERING EDUCATION PLATFORM

XUANYUAN WU*, YI XIAO†, AND ANHUA LIU‡

Abstract. The implementation of innovation and entrepreneurship education is inseparable from professional education, so it is important for the rich data in the education platform to mine the connection between professional courses and between grades and courses. The study of association rule algorithm based on education data mining improves the time performance efficiency and accuracy of Apriori algorithm. The study improves the time efficiencies of Apriori algorithm by maintaining Map table and splitting transaction database; the accuracy is improved by using mixed criteria to measure the accuracy and filtering deformation rules based on the inference of confidence. The results of the validation of the time efficiency of the algorithm show that the running time of the improved algorithm in solving frequent itemsets is improved by about 93.86%, 92.48% and 92.76%, respectively, compared with the other three algorithms. The running time of the algorithm for generating frequent itemsets of all orders is about 91.35 ms, which is 66.13% and 83.72% better than the Apriori algorithm and AprioriTid algorithm, respectively. The mining results of student examination data based on the education platform are reasonable and practical, which are of good practical significance for the innovation and entrepreneurship engineering education platform to develop training plans and improve teaching quality.is assumed.

Key words: Engineering education platform; innovation and entrepreneurship practice; association rules; data mining; Apriori algorithm

1. Introduction. With the advancement of computer technology and database technology, the need for techniques to perform data mining and provide information for decision making based on massive data has continued to grow in various industries [1]. Association rules are important indicators in data mining to reflect the implicit connection between multiple transactions and are being used in the retail industry, educational data mining, medical industry, and financial industry [2, 3]. Association rules are also presented in various forms in the education industry, including the connection between different course grades and the effect of course placement order on overall grades [4]. Among them, analyzing students' test scores and mining the inter-course correlations and learning status information reflected by students' grades can be of great help to students' course selection planning and teachers' teaching according to their needs. There is abundant data in the Innovation and Entrepreneurship Engineering Education Platform, including students' basic information, usual course performance, examination results, stage statistics, etc. These data contain some very valuable information, such as students' mastery of the content they have learned. Therefore, using association rule algorithms to mine the relationship between students' examination results of each course and the relationship between courses is important for activities such as the development of students' course plans and the scheduling of learning progress [5, 6]. However, most domestic universities currently lack a more scientific and in-depth detailed study of the massive data of students' performance. Therefore, the study makes improvements based on the Apriori algorithm and uses the improved algorithm to mine association rules on student data within the database of the education platform, aiming to mine the relationship between students' examination results and the relationship between courses. It helps teachers to develop a curriculum plan that meets students' needs according to their four-year mastery of university courses and the characteristics of engineering education, so as to facilitate the teaching of students' innovation and entrepreneurship practice from shallow to deep and step by step.

*Basic Department, Hunan Communication Polytechnic, Changsha, 410132, China (wuguiqui@163.com)
†Modern Education Technology Center, Hunan Communication Polytechnic, Changsha, 410132, China
‡College of Continuing Education, Hunan Vocational College of Science and Technology, Changsha, 410004, China

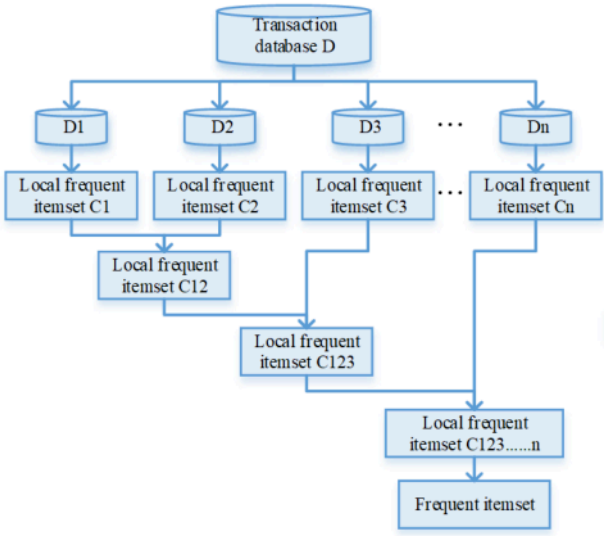
2. Related Work. With the increasing competition in various industries, data mining techniques are extensively used in diverse fields and have achieved good research success. Yang et al. addressed database technology in data mining, studied the underlying theory and methods of databases, and processed and analyzed data stored based on databases with the aim of improving the usability and popularity of database technology [7]. Amin et al. medical researchers developed prediction models using combination of distinct features and seven classification techniques for identifying salient features and data mining techniques that can enhance the success of predicting cardiovascular diseases and obtained 87.4% accuracy in predicting heart disease [8]. Yun et al. researchers proposed a visualization-aided decision-making system based on data mining techniques for industrial applications. and analyzed the architecture of the system in the context of a practical data mining technique case study verified its effectiveness and robustness [9]. Francis and Babu proposed a new prediction algorithm for predicting student exam results by combining clustering methods and data mining and evaluated and verified the accuracy of this hybrid data mining algorithm in real time using a dataset of students from various academic disciplines in a higher education institution [10]. Ayatollahi et al. compared the performance of artificial neural network and support vector machine-based data mining methods for cardiovascular disease prediction, the latter obtained 16.71 Hosmer-Lemehue test results and up to 92.23% sensitivity, while the area under their subjects' working characteristic curve was larger than that in the neural network model, indicating that support vector machines are more suitable for cardiovascular disease prediction [11].

Finding frequent itemsets is the biggest step in the process of mining association, and one of the most classic algorithms is Apriori algorithm. This algorithm is mainly used for mining single-dimensional binary association rules and has obtained good research results in many fields. Xie X and other scholars proposed a risk prediction and factor risk analysis method fused with Drosophila optimization algorithm and general regression neural network algorithm for coal and gas protrusion accidents, and Apriori algorithm was used to mine the hazard data of accidents. The proposed algorithm was applied to a coal mine to obtain 100% accuracy of accident risk level prediction [12]. Ünvan successively used Apriori algorithm and FP Growth algorithm to analyze a dataset containing 225 products and used FP Growth algorithm to find the top ten rules based on confidence values, according to which supermarkets can make product The supermarkets can adjust the product location based on these rules to increase the product sales and supermarket revenue [13]. Zhan et al. proposed a systematic method for linking customer knowledge and innovative product development in a data-driven settings. The method was used for data mining to obtain customer requirements through Apriori algorithm, correlation rules and decision tree methods. The results proved the validity and usefulness of the proposed method [14]. Ali et al. used the Apriori algorithm in data mining to investigate recovery and mortality factors in a schistosomiasis pretreatment dataset collected from Hubei, China, and evaluated in different tools and models to obtain generative rules with minimum support and minimum confidence indicating higher than 90% and also identified properties indicative of individual recovery and mortality: body mass index, nutrition, degree of ascites, etc., to provide professionals with More accurate guidance [15]. Jing proposed a personalized tourist route intelligent recommendation method based on association rules, which determines the range of tourist attractions by attribute clustering, then extracts the features of attractions, tourists and tourists' interest points by using association rule algorithm to complete the personalized classification of tourist routes, and finally calculates the similarity of tourist routes by dynamic and static attributes and outputs the top attractions with the highest probability. The method obtains 98.5% accuracy, 97% recall and 6s running time in simulation experiments [16].

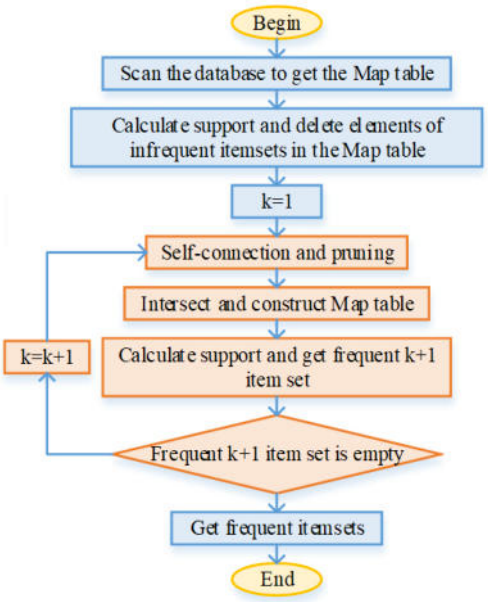
Comprehensive domestic and foreign research scholars through data mining and Apriori algorithm can be found that although Apriori algorithm has good performance in data mining, but there are also certain defects, to apply it to the analysis of student data in the innovation and entrepreneurship engineering education platform, the first need to make improvements for its deficiencies. Therefore, the study improves the time efficiency of Apriori algorithm by maintaining Map table and filtering deformation rules, and applies it to the association rule mining of students' examination results.

3. Improved Apriori algorithm for innovation and entrepreneurship engineering education platform.

3.1. Time efficiency study of improved Apriori algorithm. The common association rule algorithms used for educational data mining are Apriori algorithm, Partition algorithm, FP-Growth algorithm, etc [17].



(a) Flowchart for reducing the number of transaction database scans



(b) Flowchart for reducing the number of scanned transactions

Fig. 3.1: Time efficiency improvement of Apriori algorithm

In the generating process of frequent item set by Apriori algorithm, the candidate frequent item set decreases with the increase of order, and the scanning transaction database still needs to traverse all transaction sets and all transactions in the set, which can greatly slow down the algorithm's time efficiency [18]. The improved AprioriTid algorithm reduces the number of scanned transaction sets by constructing a Tid table, but the large quantity of item sets included in the Tid table transactions at the beginning of the algorithm still causes time wastage. Therefore, the study addresses the shortcomings of Apriori algorithm and AprioriTid algorithm and makes corresponding improvements to them in respect of time performance for the number of transactions in the transactional database and the frequency of transaction database scans. In order to reduce the quantity of transactions scanned in the database, the improved algorithm replaces the operation of scanning the database by maintaining a Map table each time when counting the number of supports for a certain frequent itemset. The number of scans of the database is reduced by splitting the transactional database into several disjoint parts. The exact process of improving time performance is shown Figure 3.1. In Figure 3.1(b), the transactions of each partition are scanned in turn and the local frequent set and its support number are obtained, and then the local frequent set is updated to the candidate frequent set. When all partitioned transactions are scanned, all candidate frequent sets are updated and the final frequent set is determined based on their support numbers and the total number of transactions. To verify the feasibility of the improved algorithm, suppose there is a transaction database with m records, where the number of transactions contained in the first i record is $N_{R[i]}$. When using the Apriori algorithm for frequent item set solving, the count of transactions $Num_{[j]}$ to be scanned for the item set of the j^{th} stratum is calculated as shown in Equation (1).

$$Num_{[j]} = N_{I[j]} \times \sum_{i=0}^{m-1} N_{R[i]} \tag{3.1}$$

In Equation 3.1, assume that the solved frequent itemset has a total of k order and the sample set of frequent terms contained in the j^{th} order is $N_{I[j]}$. The total quantity of transaction comparisons N_{TOTAL} is calculated

as shown in Equation 3.2.

$$Num_{Total1} = \sum_{j=1}^{k} Num_{[j]} \tag{3.2}$$

When using the AprioriTid algorithm for support figure calculation of candidate frequent item sets, the corresponding Tid table also has k order. Assume that there are $n_{[j]}$ records in the Tid table of order j and the number of itemsets contained in the i record is $N_{TID\_I[i]}$. Therefore, the number of itemsets in the Tid table scanned by the candidate frequent itemsets at the j level $Num_{Tid\_Scan[j]}$ is calculated as shown in Equation 3.3.

$$Num_{Tid_Scan[j]} = N_{I[j]} \times \sum_{i=0}^{n_{[j]}} N_{Tid_I[i]} \tag{3.3}$$

According to equation 3.3, the corresponding total number of item set comparisons can be obtained from N_{TOTAL2}, and its computational expression is shown in equation 3.4.

$$Num_{Total2} = \sum_{j=1}^{k} Num_{Tid_Scan[j]} \tag{3.4}$$

Instead of scanning the original transaction database, the Map table which corresponds to the candidate itemsets is scanned when the support figure is calculated using the improved Apriori algorithm. The candidate frequent item set number in all strata is shown in Equation 3.5.

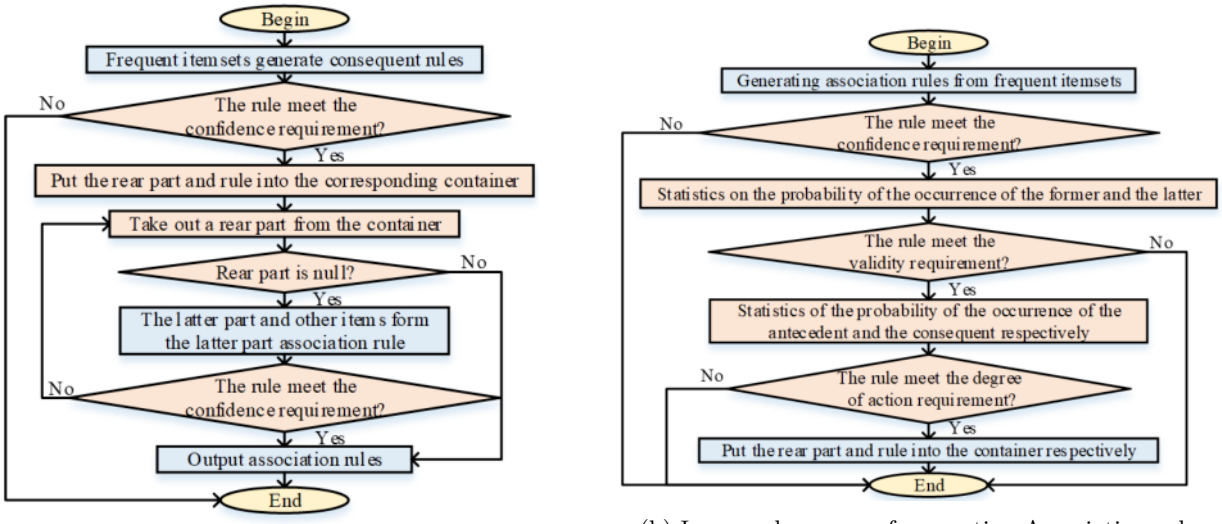
$$N_{Tatal_I} = \sum_{j=1}^{k} N_{I[j]} \tag{3.5}$$

Since a candidate frequent item set of order k consists of two frequent item sets of order $k-1$, and assuming that the number of IDs contained in the first frequent item set of i is $N_{Tid[i]}$ and that candidate frequent item set consists of the first and the first $Num_{Tid\_Scan[j]}$ item sets, the number of comparisons of the support number of a candidate frequent item set and the total number of comparisons are calculated as shown in Equation 3.6.

$$\begin{cases} N_{Com[i]} = max\big(N_{Tid[i1]}, N_{Tid[i2]}\big) \\ Num_{Total3} = N_{Com[i]} \times N_{Total_I} \end{cases} \tag{3.6}$$

Thus the total number of comparisons of the three algorithms can be approximated by the comparisons of $\sum_{i=0}^{m-1} N_{R[i]}$, $\sum_{j=0}^{n_{[k]}-1} N_{Tid_I[j]}$ and $max\big(N_{Tid[i1]}, N_{Tid[i2]}\big)$. In the case of a given transaction database, $\sum_{i=0}^{m-1} N_{R[i]}$ is a constant value and the AprioriTid algorithm shows that $n_{[k]}$ is a value less than m and decreases as the stratum k increases. Therefore, Apriori algorithm is suitable when there are few records in the database and AprioriTid algorithm is suitable when there are more records. For, $max\big(N_{Tid[i1]}, N_{Tid[i2]}\big)$ also decreases as the hierarchy k increases, and $N_{Tid[i]}$ is always smaller than m, so $max\big(N_{Tid[i1]}, N_{Tid[i2]}\big)i$ is smaller than $\sum_{i=0}^{m-1} N_{R[i]}$. When the data is more random, each candidate frequent item set is present in fewer records, i.e., $max\big(N_{Tid[i1]}, N_{Tid[i2]}\big)$ is smaller. When the data are more similar, it is not possible to visually compare the sizes of $\sum_{j=0}^{n_{[k]}-1} N_{Tid_I[j]}$ because the size of the two cannot be determined. The comprehensive appeal discussion shows that the improved algorithm is feasible and will have some improvement in time efficiency compared to Apriori algorithm and AprioriTid algorithm.

3.2. Accuracy study of improved Apriori algorithm. Commonly used accuracy measures for association rule algorithms include support-confidence, usefulness and validity. Using only two criteria, support and confidence, to measure the results of association mining often generates a large number of useless rules, wrong rules and redundant rules [19]. Considering the advantages and disadvantages of different measures, the study



(a) Process of generating Association rules

(b) Improved process of generating Association rules

Fig. 3.2: The process of generating affiliation rules before and after improvement

uses a mixture of measures for association rule metrics. The degree of action can be used to measure the degree of influence of the occurrence of the antecedent piece in the association rule on the subsequent piece, and is therefore also known as the degree of relevance, which is calculated as shown in Equation 3.7.

$$\textbf{Importance}(A \Rightarrow B) = \frac{Support(A \cup B)}{Support(A) * Support(B)} \tag{3.7}$$

In Equation 3.7, $\textbf{Importance}(A \Rightarrow B)$, represents the relevance of the affiliation rule $(A \Rightarrow B)$ and $Support(A)$ stands for $A^{'}s$ support level. $\textbf{Importance}(A \Rightarrow B) = 1$ denotes that A and B are mutually independent, and the association rule $A \Rightarrow B$ is called irrelevant rule. $\textbf{Importance}(A \Rightarrow B) > 1$ Indicates a positive correlation between A and B, i.e., the possibility of A increases the likelihood of B occurring, and the association rule $A \Rightarrow B$ is referred to as a positive correlation rule. However, the correlation still has some limitations, so in order to have a more comprehensive understanding of the impact caused by the occurrence of the item set A on the occurrence of the item set B , the concept of effective degree emerged, which is calculated as shown in Equation 3.8.

$$\textbf{Validity}\,(A \Rightarrow B) = P\,(A \cup B) - P(\overline{A} \cup B) \tag{3.8}$$

In Equation 3.8, $P\,(A \cup B)$ represents the probability that both the item set A and the item set B appear in the transaction database, while $P(\overline{A} \cup B)$ represents the probability that the item set A does not appear in the database and the item set B appears in the database. To improve the accuracy of the association rules, improvements are made to address the shortcomings of the support-confidence measurement process. The process of generating association rules before and after the improvement is shown in Figure 3.2. 3.2

As can be seen in Figure 3.2(b), the specific improvement steps are: firstly, we judge the support and confidence level of the generated affiliation rules, and those that meet the confidence requirements enter the next process; then we judge the validity, and delete the association rules that are more likely to appear in the absence of the antecedent; finally, we judge the usefulness, and further delete those association rules whose appearance of the antecedent will lead to a lower probability of the appearance of the consequent. The second improvement is carried out for the deformation rules of association rules. The confidence degree is calculated

as shown in Equation 3.9.

$$\text{Confidence}\,(A \Rightarrow B) = \frac{\text{Support}\,(A \cup B)}{\text{Support}\,(A)} \tag{3.9}$$

In Equation 3.9, Confidence $(A \Rightarrow B)$ represents the confidence of the association rule $(A \Rightarrow B)$. According to Equation 3.7, we can know that assuming the existence of the association rule $R_5 \Rightarrow R_1 \wedge R_2 \wedge R_3 \wedge R_4$, its confidence level is shown in Equation 3.10.

$$\textbf{Confidence}(R_5 \Rightarrow R_1 \wedge R_2 \wedge R_3 \wedge R_4) = \frac{Support(R_5 \cup R_1 \wedge R_2 \wedge R_3 \wedge R_4)}{Support(R_5)} \tag{3.10}$$

In Equation 3.10, $Confidence(R_5 \Rightarrow R_1 \wedge R_2 \wedge R_3 \wedge R_4)$ is the confidence level of the association rule $R_5 \Rightarrow R_1 \wedge R_2 \wedge R_3 \wedge R_4$. And the confidence of the deformation rule $R_5 \wedge R_4 \Rightarrow R_1 \wedge R_2 \wedge R_3$ of this rule Confidence $(R_5 \wedge R_4 \Rightarrow R_1 \wedge R_2 \wedge R_3)$ is shown in Equation 3.11.

$$\text{Confidence}\,(R_5 \wedge R_4 \Rightarrow R_1 \wedge R_2 \wedge R_3) = \frac{Support\,(R_5 \wedge R_4 \cup R_1 \wedge R_2 \wedge R_3)}{Support\,(R_5 \wedge R_4)} \tag{3.11}$$

$R_5 \wedge R_3 \wedge R_4 \Rightarrow R_1 \wedge R_2$ The confidence level of Confidence $(R_5 \wedge R_3 \wedge R_4 \Rightarrow R_1 \wedge R_2)$, another deformation rule of the association rule $R_5 \Rightarrow R_1 \wedge R_2 \wedge R_3 \wedge R_4$, is shown in Equation 3.12.

$$\text{Confidence}\,(R_5 \wedge R_3 \wedge R_4 \Rightarrow R_1 \wedge R_2) \tag{3.12}$$

The confidence level of for , another deformation rule of the association rule , is shown in Equation 3.13.

$$\text{Confidence}\,(R_5 \wedge R_2 \wedge R_3 \wedge R_4 \Rightarrow R_1) = \frac{Support\,(R_5 \wedge R_2 \wedge R_3 \wedge R_4 \cup R_1)}{Support\,(R_5 \wedge R_2 \wedge R_3 \wedge R_4)} \tag{3.13}$$

According to the related concept of support, it is known that the association rule has equal support with its three deformation rules, and there is a relationship as shown in Equation 3.14.

$$\begin{aligned} Support(R_5) &\geq Support(R_5 \wedge R_4) \\ &\geq Support(R_5 \wedge R_4 \wedge R_3) \\ &\geq Support(R_5 \wedge R_4 \wedge R_3 \wedge R_2) \end{aligned} \tag{3.14}$$

The confidence relationship of association rules and their deformation rules can be obtained by combining Equation 3.10 to Equation 3.14, as shown in Equation 3.15.

$$\begin{aligned} &\text{Confidence}(R_5 \Rightarrow R_1 \wedge R_2 \wedge R_3 \wedge R_4)_{\min} \\ &\quad \leq \text{Confidence}(R_5 \wedge R_4 \Rightarrow R_1 \wedge R_2 \wedge R_3) \\ &\quad \leq \text{Confidence}(R_5 \wedge R_4 \wedge R_3 \Rightarrow R_1 \wedge R_2) \\ &\quad \leq \text{Confidence}(R_5 \wedge R_4 \wedge R_3 \wedge R_2 \Rightarrow R_1) \end{aligned} \tag{3.15}$$

In Equation 3.15, Con_{min} is the minimum confidence level. Equation 3.15shows that the morphing rules of an association rule are all strong correlation rules that satisfy the confidence measure[20]. When a powerful association rule is obtained, its similar deformation rules must be powerful association rules, but this part of rules is not our target rules, so we need to filter out this part of deformation rules in the process of producing association rules.

3.3. An Improved Apriori algorithm based on the education platform student performance analysis method. In the innovation and entrepreneurship engineering education platform, Students' performance in examinations is not only an overwhelming measure of teaching quality, but also an important basis for guiding students to properly select and complete the corresponding innovation and entrepreneurship courses [21]. The structure of student performance analysis application system is shown in Figure 3.3.

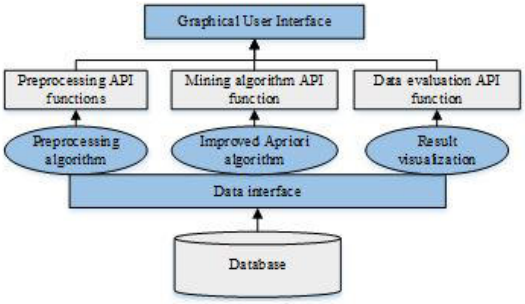


Fig. 3.3: Structure chart of student performance analysis application

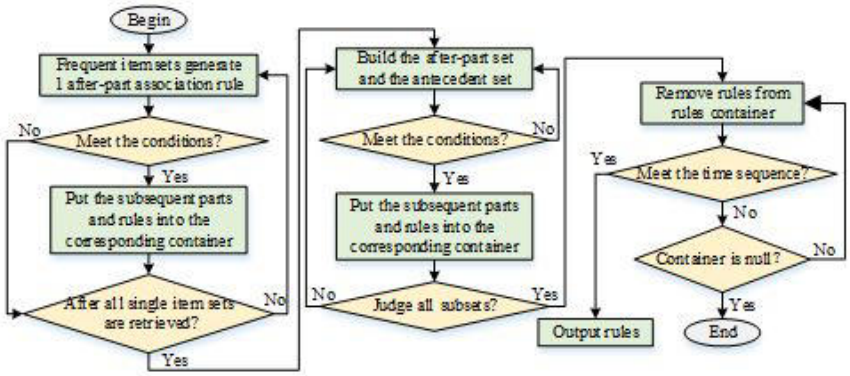


Fig. 3.4: The process of generating association rules according to the improved Apriori algorithm

In Figure 3.3, the study divides the student performance analysis application into three modules according to different user needs: student performance analysis module, individual student performance analysis module, and teacher course analysis module. The student grade analysis module is mainly used to mine the correlations among students' course grades in the four years of college and store the obtained rules in the rule base. The process of generating association rules according to the improved Apriori algorithm is as shown in Figure 3.4.

As shown in Figure 3.4, a single item is selected as a posterior in the input frequent item set at a time, and it is judged whether the association rule satisfies the relevant conditions, and then the posterior items that satisfy the conditions are put into the posterior container, and the formed association rule is put into the corresponding rule container. After that, the items in the post-item container are taken in turn to form the post-item set with other items in the frequent item set, and the other items in the item set form the pre-item set. Finally, take the elements from the rule container in turn and judge whether they satisfy the temporal order, and print the rules that satisfy the condition, otherwise discard the rule. When the container of association rules is empty it means the generation of association rules is completed. The student performance analysis module is an analysis of all students' performance. Students can view the relationship between the grades of each course, but it is difficult to visually identify the rules related to them from the results. Therefore, the study uses the course exam results entered by students as the antecedent of the correlation rules in the individual student performance analysis module, and predicts students' performance in subsequent courses based on the generated rules. Based on the prediction results, students can have a clearer idea of which courses they need to focus on in their subsequent courses, so that they can make scientific and effective study plans in advance. In university teaching, the

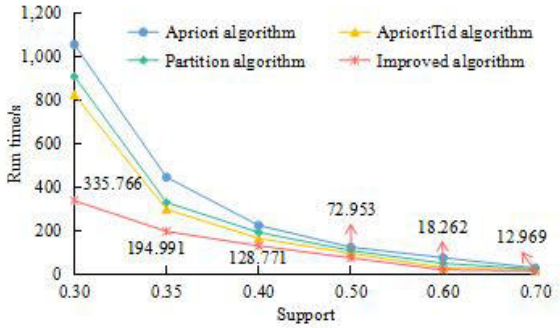


Fig. 4.1: The time of solving frequent itemsets by different algorithms with small transaction differences

study of public basic courses is often carried out through large classes, while some subjects with stronger specialization, such as those in the innovation and entrepreneurship engineering education platform, are mostly taught in small classes. The teacher curriculum analysis module can analyze the teacher curriculum according to the connection between students' performance in each subject and carry out scientific class teaching, which is more conducive to teachers' teaching according to their ability.

4. Results of the improved Apriori algorithm in the innovation and entrepreneurship engineering education platform.

4.1. Performance analysis of Apriori algorithm before and after optimization. The performance comparison experiments before and after the optimization of Apriori algorithm are based on the verification of time efficiency improvement, firstly, by conducting time complexity comparison experiments on Apriori algorithm, AprioriTid algorithm, Partition algorithm and the improved algorithm. The time complexity comparison experiments are conducted in three times, including time-efficient effectiveness in the situation of small differences in transactions, time-efficient effectiveness in the situation of large differences in transactions, and time-efficient effectiveness in the generation of frequent itemsets of each order. The data in Experiment 1 are 1000 data excerpted from the UCI dataset and the mushroom.dat data in PUMSB, where the differences between each transaction are small. By varying support settings, the time required for each algorithm to solve the set of frequent items under different thresholds was counted, and the experimental results obtained are shown in 4.1.

In Figure 4.1, both the AprioriTid algorithm and the Partition algorithm, which are derived from the improved Apriori algorithm, have a certain degree of reduction in the time needed to solve the frequent itemset compared to the Apriori algorithm, and the algorithms are more time efficient. The improved algorithm has less running time compared to other algorithms, especially in the range of support threshold of 0.30 0.40, and this time efficiency improvement is more obvious. Experiment 2 selected 1000 data from the T10I4D100K dataset to verify the time-efficient of the improved algorithm in the presence of large transaction differences. Due to the significant differences between transactions in the experimental data, a smaller minimum support needs to be set in the setup, otherwise the frequent item set cannot be obtained. The time required by the four algorithms to solve the frequent itemset under different thresholds is shown in Figure 4.2.

As shown in Figure 4.2, the improved algorithm improves the running time by about 93.86%, 92.48% and 92.76% compared to the Apriori algorithm, AprioriTid algorithm and Partition algorithm when the support threshold is 0.010. The improved algorithm improved the average running time by about 91.93%, 81.43% and 88.88% compared to the Apriori algorithm, AprioriTid algorithm and Partition algorithm for different support thresholds, respectively. The runtime ratio of the improved algorithm to the other three algorithms in Experiment 2 is larger compared to Experiment 1. The reason for this analysis is that with significant differences between transactions, the number of transaction IDs contained in the Map table corresponding to

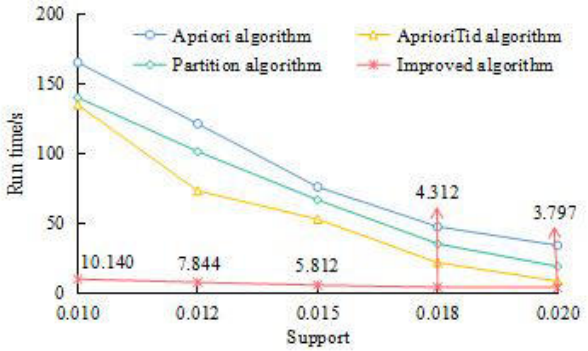
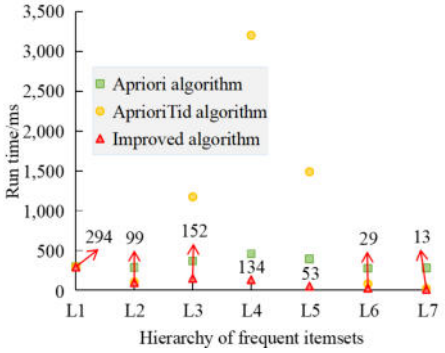
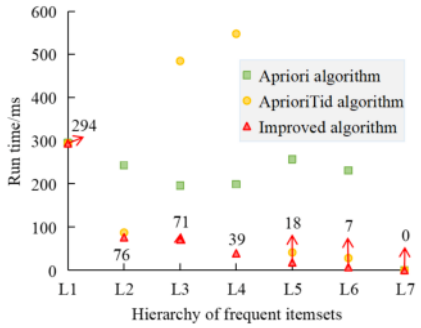


Fig. 4.2: The time of solving frequent itemsets with different algorithms in the case of large transaction differences



(a) Time to generate frequent itemsets of each order when the support is 0.3



(b) Time to generate frequent itemsets of each order when the support is 0.5

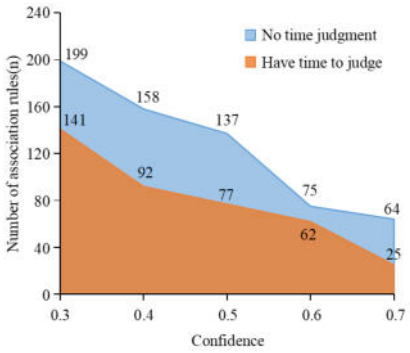
Fig. 4.3: Time to generate frequent itemsets of each order when support is 0.3 and 0.5

each item set is relatively small, so the comparison times are decreased and the effectiveness of solving frequent item sets is improved. Experiment 3 is to count the running time of different algorithms during the generation of frequent itemsets of each order under the support threshold of 0.3 and 0.5, and the results are presented in Figure 4.3.

In Figure 4.3(a) and Figure 4.3(b), the difference between the running time of the Apriori algorithm for generating frequent itemsets of each order is not significant, while the runtime of the AprioriTid algorithm for generating L3 itemsets and L4 itemsets is much larger than that of the Apriori algorithm. However, as the class of frequent itemsets increases, the transactions in the AprioriTid algorithm gradually decrease and the algorithm running time is significantly lower than that of the Apriori algorithm. Regardless of the support threshold of 0.3 or 0.5, the running time efficiency of the improved algorithm is always significantly higher than that of the Apriori algorithm and the AprioriTid algorithm throughout the generation of frequent itemsets. The average running times of the improved algorithm, Apriori algorithm and AprioriTid algorithm for each order of frequent item set generation were about 91.35 ms, 269.72 ms and 561.00 ms.

4.2. Effectiveness of Apriori algorithm for mining student achievement data in educational platform. The study selected all student grades of three graduated grades of software engineering majors in the

(a) The number of rules generated in two cases when the support is 0.2



(b) The number of rules generated in two cases when the support is 0.3

Fig. 4.4: Number of rules generated by time decision and no time decision under different support

database of the innovation and entrepreneurship engineering education platform as sample data, and conducted data preprocessing operations such as checking, filling, cleaning, and conversion. In the data preprocessing, considering the excessive variability of each student's choice of elective courses, only the students' compulsory course grades were selected and only the grades of students' first exams were used. Considering that string matching and storing the Chinese names of courses would cause a large memory overhead, numeric codes are used instead of the Chinese names of courses. The probability of having duplicate data in the resulting set of frequent items is too low due to too large a span of student grades, so five grades of A, B, C, D, and E are used instead of percent test scores, and the weighted average score is denoted by the letter Z. When a student misses an exam in a course, if the student has a make-up exam grade, the make-up exam grade is used as the exam grade, otherwise it is automatically classified as an E grade. To analyze the effect of the academic year taken in the course on the mining results, the study collected the result data obtained without considering the academic year taken in the course and considering the academic year taken in the course, as shown in Figure 4.4.

As shown in Figure 4.4, with support degrees of 0.2 and 0.3, the number of association rules generated when considering the academic year taken in the course is much less than that generated when not considering the academic year taken in the course, and the ratio of the number of association rules is about 0.63. Therefore, the study adds the judgment of time to the algorithm for grade analysis, so that some redundant association rules can be removed. The minimum degree of support is set to 0.198 and the minimum confidence level is set to 0.657. The number of frequent item sets of each order generated by the improved algorithm for data mining is shown in Figure 4.5.

As shown in Figure 4.5, invalid association rules and uninteresting association rules account for a large proportion of the results. Some of the invalid association rules can be removed by the validity and usefulness, and the uninteresting rules can be removed by the inference formula and course time judgment. Removing these wrong rules can better select the rules that are more beneficial to users for popularization. Some of the strong association rules in the mining results are presented in Table 4.1.

A higher confidence level for the association rule in Table 1 represents a greater degree of association between the preceding and following pieces of the rule, that is, a higher degree of importance of the rule. The study divided the strong association rules into the rules of course scheduling rationality and the rules of subject's influence on overall performance. According to the rules $2C \Rightarrow 5C$ and $5C \Rightarrow 7C$ in Table 1, it can be found that if a software engineering student gets a poor grade in course 2, there is a high probability that it will lead to a poor grade in course 5, which will eventually affect the grade in course 7. Therefore, when arranging courses in the education platform, we need to arrange course 2 in the front and focus on improving teaching level to prepare the basis for students to take course 5 and course 7 next. According to the three rules in Table

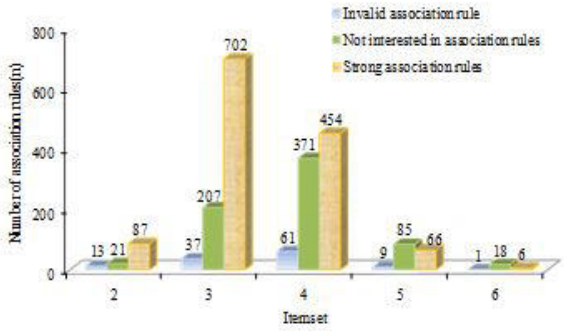


Fig. 4.5: The quantity of frequent itemsets of each order generated by the improved algorithm for data mining

Table 4.1: Some Intense Association Rules in Mining Results

Rule Antecedent	**Rule Aftereffect**	**Support of Front Parts**	**Confidence**
2C	5C	0.3333	0.9259
6A	4A	0.2469	0.8250
1A	11A	0.4383	0.8169
3A	ZA	0.3333	0.7593
5C	7C	0.2346	0.7556
10B	1A	0.5370	0.7287
9A	ZA	0.4198	0.7260
4A	1A	0.3148	0.7255
8C	9C	0.4136	0.7075
8A	ZA	0.4198	0.7059

1 $3A \Rightarrow ZA$, $8A \Rightarrow ZA$ and $9A \Rightarrow ZA$, it can be seen that Course 3, Course 8 and Course 9 have a greater impact on students' overall performance, i.e. students with good performance in these subjects have mostly higher overall performance. Therefore, the credit weights of these courses can be appropriately adjusted in the innovation and entrepreneurship engineering education platform as a way to balance the professional level of students.

5. Conclusion. The cultivation of innovation and entrepreneurship is an important part of the reform of student cultivation mode in China's universities, and there are high requirements for students' capability and innovation skills in engineering education. Therefore, it is necessary to arrange courses and innovation and entrepreneurship practice activities in a targeted way based on students' ability to master the subject theory. The Apriori algorithm of association rule mining is studied to mine the implicit information in students' performance. Improvements to the Apriori algorithm are made in two aspects: time efficiency and accuracy, specifically by reducing the number of scanned databases and transactions to improve the operational efficiency of the algorithm, and by filtering deformation rules to improve the accuracy. In the validation experiments on the effectiveness of the improved performance, the time required for the improved algorithm to resolve the frequent itemsets is improved by about 93.03% compared with the other three algorithms. The operation time of the improved algorithm to generate frequent itemsets of each order is about 91.35 ms, which is 66.13% and 83.72% better than the Apriori algorithm and AprioriTid algorithm, respectively. However, the algorithm proposed in the study only refers to the existing support and confidence levels in relevant research when generating

association rules, and does not delve into the most suitable selection method for support and confidence levels. In the future, further optimization of the algorithm should be carried out for the selection of minimum support and minimum confidence levels, and further research is needed on how to select the best rule among the generated association rules for popularization.

Fundings. The research is supported by: 2023 Hunan Provincial Social Science Achievement Evaluation Committee member general funding project "Research on Innovation of Professional Innovation Integration Path under Employment Priority Strategy" (No.: XSP2023JYZ022).

REFERENCES

[1] Sunhare, P., Chowdhary, R. & Chattopadhyay, M. Internet of things and data mining: an application-oriented survey. journal of King Saud University- Computer and Information Sciences. (2022)
[2] Menon, S., Ghoshal, A. & Sarkar, S. Modifying transactional databases to hide sensitive association rules. *Information Systems Research*. **33**, 152-178 (2022)
[3] Thurachon, W. & Kreesuradej, W. Incremental association rule mining with a fast incremental updating frequent pattern growth algorithm. *IEEE Access*. **9** pp. 55726-55741 (2021)
[4] Aldowah, H., Al-Samarraie, H. & Fauzy, W. Educational data mining and learning analytics for 21st century higher education: a review and synthesis. *Telematics And Informatics*. **37** pp. 13-49 (2019)
[5] Ghafar, A. Convergence between 21st century skills and entrepreneurship education in higher education institutes. *Education*. **9**, 218-229 (2020)
[6] Duan, N., Duan, L. & Lu, H. The current situation of innovation and entrepreneurship education and its optimization countermeasures: A case study of International Journal of Social Science and Education Research. (2021)
[7] Yang, J., Li, Y., Liu, Q., Li, L., Feng, A., Wang, T., Zheng, X. & Lyu, J. Brief introduction of medical database and data mining technology in big data era. journal of Evidence-Based Medicine. (2020)
[8] Amin, M., Chiam, Y. & Varathan, K. Identification of significant features and data mining techniques in predicting heart disease. telematics and Informatics. (2019)
[9] Yun, Y., Ma, D. & Yang, M. Human-computer interaction-based decision support system with applications in data mining. future Generation Computer Systems. (2021)
[10] Francis, B. & Babu, S. Predicting academic performance of students using a hybrid data mining approach. journal of medical systems. (2019)
[11] Ayatollahi, H., Gholamhosseini, L. & Salehi, M. Predicting coronary artery disease: a comparison between two data mining algorithms. bmc public health. (2019)
[12] Xie, X., Fu, G., Xue, Y., Zhao, Z., Chen, P., Lu, B. & And, J. and factors risk analysis based on IFOA-GRNN and apriori algorithms: application of artificial intelligence in accident prevention. *Process Safety And Environmental Protection*. **122** pp. 169-184 (2019)
[13] Ünvan, Y. Market basket analysis with association rules. *Communications In Statistics-Theory And Methods*. **50**, 1615-1628 (2021)
[14] Zhan, Y., Tan, K. & Huo, B. Bridging customer knowledge to innovative product development: a data mining approach. *International Journal Of Production Research*. **57**, 6335-6350 (2019)
[15] Ali, Y., Farooq, A., Alam, T., Farooq, M., Awan, M. & Baig, T. Detection of schistosomiasis factors using association rule mining. *IEEE Access*. **7**, 86108-18611 (2019)
[16] Jing, Y. An intelligent recommendation method of personalised tour route based on association rules. international Journal of Reasoning-based Intelligent Systems. (2023)
[17] Cantabella, M. Martínez-Espa na R, Ayuso B, Yá nez J A. *Mu Noz A. Analysis Of Student Behavior In Learning Management Systems Through A Big Data Framework*. **90** pp. 262-272 (2019)
[18] Zhu, S. Research on data mining of education technical ability training for physical education students based on Apriori algorithm. *Cluster Computing*. **22** pp. 14811-14818 (2019)
[19] Wang, Z., Tian, Q. & Duan, X. Research on the evaluation index system of college students' class teaching quality based on association algorithm. *Cluster Computing*. **22** pp. 13797-13803 (2019)
[20] Wang, H. & Gao, Y. Research on parallelization of Apriori algorithm in association rule mining. *Procedia Computer Science*. **183** pp. 641-647 (2021)
[21] Khan, A. & Ghosh, S. Student performance analysis and prediction in classroom learning: a review of educational data mining studies. education and information technologies. (2021)

Edited by: Mudasir Mohd
Special Issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: May 18, 2023
Accepted: Jul 18, 2023