Correlation analysis of College students' achievement based on improved Apriori algorithm

Hongqin Wang

Department of Software Engineering, Software Engineering institute of Guangzhou, Guangdong Province, 510990, China
Hongqin Wang’s E-mail: whq@sise.com.cn

Abstract. A large amount of student achievement information has been accumulated in the teaching information management system of colleges. It is of great significance to use data mining technology to analyze the data and find out the potential value of the data. The basic idea and defects of the Apriori algorithm are described in this paper, an improved Apriori algorithm is proposed, which is used to analyze the course scores of all students in the software engineering major of a university in 2018, and dig out the relationship between scores and courses, so as to provide learning guidance for students, provide teaching reference for teachers, and have certain guiding significance for the quality of talent cultivation in colleges and universities.

1. Introduction

With the continuous advancement of information construction[1] in colleges and universities, the data of student achievement in the teaching management system of colleges and universities is increasing day by day. These information are simply input, statistics, query and backup by users, but the value hidden behind the massive data can not be effectively used. Therefore, it is necessary to process and analyze these data, find useful rules, and provide effective support for teaching arrangement and teaching decision-making.

Association rule mining is a mining technology to find potential valuable information from massive data, so as to find the association relationship or correlation between the same data set. In this paper, the improved Apriori algorithm is used to mine students' achievement, which can find out the factors that affect students' achievement and the relationship between courses. These information can provide supporting suggestions for teaching and management, and also provide a favorable teaching basis for making a reasonable talent training plan.

2. The introduction of Apriori algorithm

2.1. Basic idea of Apriori algorithm

When using association rules for data mining and searching for frequent itemsets, the simplest and basic algorithm is the Apriori algorithm. Apriori uses an iterative method for searching layer-by-layer, k-item itemsets are used to explore (k+1) itemsets. First, by scanning the database, accumulating the count of each item, and collecting items that meet the minimum support, find the set of frequent 1-item sets[2], which is denoted as L1. Then, L1 is used to find the set L2 of frequent 2 itemsets, and L2 is used to find L3, and so on, until no more frequent K itemsets can be found. To find each Lk, a full scan of the database is required.
Using Apriori algorithm to find all frequent itemsets, it is mainly divided into two steps, joining and pruning.

The connection process refers to finding out the frequent item-set L_k, connecting with itself through L_{k-1} to generate a candidate K item set, this candidate item set is denoted as C_k, where the elements of L_{k-1} can be connected.

The process of pruning refers to that C_k is a superset of L_k, that is, its members may or may not be frequent, but all frequent itemsets are included in C_k. Scan the database to determine the count of each candidate in C_k. To determine L_k[2]. Using the Apriori property, any infrequent (k-1) itemsets cannot be a subset of frequent k itemsets. Therefore, if a candidate k item set is not in L_k, the candidate item may not be frequent, so it can be deleted from C_k[3].

After finding all frequent itemsets, strong association rules are generated from frequent itemsets. If you want to express the probability of Y appearing in the transaction of X, the strong association rule needs to be satisfied. For support(X=>Y)≥min_sup and confidence(X=>Y)≥min_conf, call the association rule X=>Y as strong association rules. Among them, min_sup is the minimum support threshold, and min_conf is the minimum confidence threshold.

2.2. Disadvantages of Apriori algorithm
During the execution of the Apriori algorithm, in order to generate C_k, the database needs to be scanned multiple times, and a large number of redundant candidate item sets will be generated during the connection process, which will increase the I/O burden and affect the efficiency of the algorithm.

3. Improved Apriori algorithm
In view of the shortcomings of the above-mentioned Apriori algorithm, the Apriori algorithm is improved[4][5]. First, taking the scores of 539 software engineering students from a university in 2018 as the research object, the courses they learned are divided into mathematics, language and specialty, and the scores of each subject are stored in the corresponding database. Then mining frequent itemsets in vertical data format for each sub database, and finally integrating the mining association rules to realize the mining of the whole database.

In the process of sub database mining, we first scan the database and create a new database. The transaction project database is transformed into the project transaction database, the number of projects is counted, and the unsatisfied itemsets are deleted according to the minimum support. Then we have to use the connection to find out the intersection of the records and regenerate the record database, so as to avoid scanning the database many times.

The implementation process of the improved Apriori algorithm is described as follows.
(1) The student achievement database is compressed into achievement transaction database after data preprocessing.
(2) According to the course content, all subjects and course scores of software engineering are divided into three sub databases: mathematics, language and specialty. Each sub database contains only one kind of course and the course scores of all such courses.
(3) For each individual sub database, a set of potential frequent itemsets is generated.
(4) The frequent itemsets in each database are integrated, and the confidence of the association rules generated by them is calculated. By comparing with the minimum confidence threshold, all the strong association rules are screened out, and a global candidate frequent itemset is obtained. The association rules obtained from the global candidate frequent itemsets are updated to the system association rule base.

4. Using the improved Apriori algorithm to analyze the students' achievement

4.1. data cleaning and replacement
The original educational administration information management system contains the scores of various majors, grades and all courses. Some courses have different errors in the records of some data, and even
some courses are not associated with each other. In order to facilitate mining, these redundant data are removed. Nine courses of advanced mathematics, Java and database principle of software engineering students of grade 2018 are selected as the mining objects, and some of the cleaned data are shown in Table 1.

Table 1. part of the original score data information

| Student ID | advance d-math | linear algebra | probability theory | Java | Android | JAVA EE | data base | design pattern | algorithm analysis |
|------------|----------------|----------------|--------------------|------|---------|---------|-----------|----------------|--------------------|
| 1840128101 | 67             | 71             | 72                 | 90   | 84      | 71      | 85        | 78            | 80                 |
| 1840128102 | 79             | 78             | 85                 | 68   | 75      | 62      | 90        | 83            | 81                 |
| 1840128103 | 82             | 88             | 86                 | 90   | 94      | 91      | 92        | 82            | 79                 |
| 1840128104 | 75             | 77             | 70                 | 68   | 66      | 69      | 81        | 80            | 78                 |
| 1840128105 | 92             | 89             | 95                 | 88   | 80      | 79      | 84        | 83            | 90                 |
| 1840128106 | 83             | 80             | 90                 | 79   | 83      | 84      | 77        | 67            | 85                 |
| 1840128107 | 62             | 69             | 71                 | 87   | 81      | 70      | 65        | 72            | 73                 |
| 1840128108 | 87             | 85             | 90                 | 91   | 82      | 91      | 75        | 81            | 81                 |
| 1840128109 | 90             | 81             | 79                 | 80   | 81      | 86      | 78        | 85            | 86                 |
| 1840128110 | 87             | 90             | 90                 | 81   | 78      | 83      | 90        | 78            | 81                 |

The continuous numerical data are transformed, and the failed data are removed. Then the scores are divided into four grades: excellent, good, medium and passing, which are represented by A, B, C and D respectively. 90 or above is A grade, 80 to 89 is B grade, 70 to 79 is C grade, 60 to 69 is D grade. In order to facilitate statistics, the nine courses are also represented by the letters K1-K9, and their one-to-one correspondence is: Higher Mathematics--K1, linear algebra--K2, probability and statistics--K3, Java programming--K4, Android Application Development--K5, Java EE technology--K6, database principle--K7, design pattern--K8, algorithm analysis--K9. The data after replacement are shown in Table 2.

Table 2. students' scores after pretreatment

| Student ID | K1 | K2 | K3 | K4 | K5 | K6 | K7 | K8 | K9 |
|------------|----|----|----|----|----|----|----|----|----|
| 1840128101 | D  | C  | D  | A  | B  | C  | B  | C  | B  |
| 1840128102 | C  | C  | B  | D  | C  | D  | A  | B  | B  |
| 1840128103 | B  | B  | B  | A  | A  | A  | A  | B  | C  |
| 1840128104 | C  | C  | C  | D  | D  | D  | B  | B  | C  |
| 1840128105 | A  | B  | A  | B  | B  | C  | B  | B  | A  |
| 1840128106 | B  | B  | A  | C  | B  | B  | C  | D  | B  |
| 1840128107 | D  | D  | C  | B  | B  | C  | D  | B  | B  |
| 1840128108 | B  | B  | A  | A  | B  | A  | C  | B  | B  |
| 1840128109 | A  | B  | C  | B  | B  | C  | B  | B  | B  |
| 1840128110 | B  | A  | A  | B  | C  | B  | A  | C  | B  |

4.2. analysis of association rules

Now, the improved Apriori algorithm is used to mine the preprocessed performance data in the Matlab environment, and different association rules are mined by changing different support degrees and confidence degrees. The association rules after mining are shown in Table 3.

Table 3. association rules of some students' grades

| Order Number | Association Rules | Support | Confidence |
|--------------|-------------------|---------|------------|
| 1            | K1(A)=>K2(C)      | 0.38    | 0.70       |
| 2            | K2(B)=>K3(B)      | 0.38    | 0.72       |
| 3            | K1(C)=>K4(D)      | 0.40    | 0.55       |
From the mining association rules, from 1 and 2, we can see that the score of advanced mathematics affects the score of linear algebra, and linear algebra affects the score of probability theory, students with high number score above middle level have a low probability of passing the Java program. From 4 and 5, we can see that the score of Java has a great impact on Android programming and Java EE technology. Because Java is the foundation of Android programming and Java EE technology, if you learn the foundation well, it will be easier to learn the follow-up courses. It can be seen from 6 and 7 that the scores of database principles and design patterns affect the course of algorithm analysis. It can be seen from 3 and 8 that the learning quality of mathematics courses and language courses has a great impact on professional courses, because mathematics courses and language courses are basic courses, which can exercise students' logical thinking ability. The learning of professional courses requires not only logical thinking ability, but also strong coding ability. If mathematics courses are not well learned, it will affect a series of subsequent courses, and even may fail. It can be seen from 9 that probability theory and design pattern course have great influence on algorithm analysis course.

To sum up, schools should strengthen the guidance of basic courses, implement different teaching tasks according to different students, and try to teach students in accordance with their aptitude. More attention should be paid to students. For courses with weak academic performance, guidance should be strengthened to improve students' learning efficiency, which is also beneficial to the improvement of teaching quality.

4.3. algorithm performance analysis

The time performance comparison of the improved and improved algorithms based on the same minimum support is shown in Figure 1. Because the improved algorithm only needs to scan the database once to find the frequent itemsets, the database is reduced by a part every time the frequent itemsets are generated, so the memory space is saved and the time efficiency is improved. When the minimum support is high, the time cost of the two algorithms is similar, because the higher support will reduce the candidate set, thus reducing the time of scanning the database, but the minimum support is low. The time cost of the traditional Apriori algorithm and the improved Apriori algorithm[4] is gradually increasing. Because the traditional Apriori algorithm will produce a large number of redundant candidate sets, which increases the mining time.

Figure 1. performance comparison of two Apriori algorithms
5. Conclusion
Data mining of association rules is widely used. This paper analyzes the score data of 2018 students majoring in software engineering in a university. After unifying and standardizing the data, we use the original Apriori algorithm and the improved Apriori algorithm to mine the database data. From the mining time, we can see the improved Apriori algorithm, the efficiency of mining[6] is improved. Through mining, the implicit relationship between the scores of various subjects is mined, and some strong association rules are obtained. Association rule discovery data courses have a great impact on subsequent learning, so students must learn it well. We should strengthen the guidance for the students with middle grades. Using the useful information hidden in these data can help teachers better guide the teaching work, understand the learning situation of students, do a good job in supervision, and make a more reasonable talent training plan for the school, which has practical significance.

Acknowledgments
Fund Project Number: Key Scientific Research Projects of Universities in Guangdong Province (Guangdong Education Science Letters [2019]68. project number: 2018KQNCX39). 2020 Scientific Research project of South china institute of Software Engineering. Gu (project number:ky202003).

Author's brief introduction: Hongqin Wang, female, lecturer, postgraduate, the research direction is big data application technology.

References
[1] Fatos Xhafa,Srikanta Patnaik.(2019)Advances in Intelligent,Interactive Systems and Applications Springer Science and Business Media LLC.
[2] John D. Kelleher,Brendan Tierney.(2018)Data Science. Springer Science and Business Media LLC.
[3] Sun, Xingming, Jinwei, et al.(2020)Artificial Intelligence and Security, Springer Science and Business Media LLC, Hohhot, China.
[4] Yan Guo, Minxi Wang, Xin Li.(2017)Application of an improved Apriori algorithm in a mobile e-commerce recommendation system. Industrial Management & Data Systems.
[5] Hongqin Wang, Hongxia Wang,Huiyong Jiang,et al.(2020)Research on an improved algorithm of Apriori based on Hadoop.In:2020 International Conference on Information Science, Parallel and Distributed Systems.XI’an,China.pp.242-245.
[6] Hongqin Wang,Hongxia Wang, Li Jiang,et al. (2020)Research and application of improved K-means based on MapReduce.In:International Conference on Computer Vision and Data Mining.XI’an,China.pp.1-6.