Research on Audit Application Based on Apriori Algorithm

Honglei Chu#, Daji Ergu*

Key Laboratory of Electronic and Information Engineering, Southwest Minzu University, Chengdu, China
Email: #2512456066@qq.com, *335160863@qq.com

Abstract. The data analysis is very important in the audit field since there are plenty of audit big data, which cannot be identify easily. This paper applies one of the big data techniques to the audit work and solves the problem that is hard to be found in the hidden information data. Specifically, the Apriori algorithm is used to mine and analyze the audited personnel's learning test data in the audit data to form the association rules, and the results can provide guidance for the next audited personnel training and testing. Firstly, the test data is preprocessed and the data is discretized; then the Apriori algorithm is used in the Python language environment, and the processed data is used to mine the association rules, and the impact of the audited personnel is found in the formed association rules. The rules of the results are analyzed.

Keywords: Audit work, data mining, Apriori algorithm, test data.

1 Introduction

In the era of big data, various industries have accumulated a huge amount of data. It becomes an important that how to mine the hidden information in the massive data and use the corresponding hidden information to create value and improve working and living conditions. In the auditing department, there are a large amount of accumulated data that is heterogeneous and various types. It is difficult and cumbersome to find important information from the data itself and at the surface level manually[1,2]. Traditional manual auditing is far from meeting the auditing requirements of the audit bureau for individuals and businesses. Data mining technology can make better use of data and mine potentially important information to improve the quality of audit work. Therefore, it is very meaningful to apply data mining technology to audit work.

Data mining ideas and techniques are widely used in related research in many fields. For applying the Apriori algorithm, Wu Xiaodong applied Apriori algorithm to college students' performance data mining, mining related factors related to test scores, and providing reference for teaching management [1]; Ji Lianen designed a visual analysis system for student achievement in multi-view solid collaborative interaction based on the characteristics of the performance data [2]; Hu Zuhui used decision tree, association rules and logistic regression to study the relationship between students' online behavior related attributes and students' learning quality [3]; He Chu proposed a curriculum association classification model based on frequent pattern spectral clustering and student performance prediction algorithm [4]; Peng Hailei studied the relationship between the feedback speed of online course teachers and the degree of feedback on student participation and learning outcomes [5]; Cao Yuyu studied the prediction of college students' performance against incomplete data sets [6]; Du Qiu studied the application of association rule algorithm in enterprise personnel management system with Apriori algorithm to explore the factors affecting employee performance [7]; Sun Ting studied the use of association rules algorithm to explore Chinese medicine for the treatment of insomnia, summed up and analyzed the common drugs and their medication rules [8]; Jia Kebin selected the Apriori algorithm to analyze the current obstetric dataset and studied the application of data mining in mobile medical systems [9]; Zhao Hongli proposed a matrix-based Apriori improved algorithm, and applied it to college students' psychoanalysis, and explored potential valuable information, which has important guiding significance for college students' mental health [10]; Wang Hua used the improved Apriori algorithm to analyze student performance data [11].

Although various algorithms of data mining have been used to different fields and made great achievements, few have addressed the applications in the audit industry. This paper refers to the data

* Corresponding author

mining ideas in the auditing industry, and puts the objectives of the audit on the work and learning attitudes of public officials, and mainly applies to the analysis of the test situation of the "One learn and one test" of the personnel of the audited entity, and finds that it is audited. The results of the personnel are related to the factors, which reflect the learning style and other issues, and provide relevant improvement suggestions to relevant departments. Among them, "One learn and one test" is an examination that is performed by auditors online. They study the relevant laws and regulations and other related regulations, and conduct related tests so as to improve the working ability and working attitude of the audited personnel. The traditional method of auditing simply deals with people with unsatisfactory results. Such an approach cannot fully identify those who have not studied seriously, and cannot find people who are not serious about learning and work. For example, a person's performance is very good, but his performance may be plagiarized. In order to solve these problems, the Apriori algorithm is applied to mine the potential useful value of the audited data of the audited personnel so as to assist the auditor comprehensively analyze the auditing data.

The remainder of this paper is organized as follows. Section 2 reviews the Apriori algorithm. In Section 3, the mining test scores and framework of data mining are proposed. Then, an experiment is conducted in Section 4. The paper is summarized and concluded in Section 5.

2 Apriori Algorithm

The Apriori algorithm is based on statistics[3]. By scanning the data set, each item set is formed and its support is calculated. According to the support degree of the item set, it is determined whether each item set is a frequent item set, and then the confidence between each frequent item set is calculated, and the association rule between them is determined. The Apriori algorithm uses an iterative idea of layer-by-layer search traversal, counts the probability of occurrence of each transaction and the probability of combination between things, forms a set of items containing n transactions, and calculates the case where item set A appears. The probability that set B appears.

There are several related definitions in the Apriori algorithm including the support of the item set, the frequent item sets, the support of the association rules, and the confidence.

Support of item sets: one item set A and the whole database D, the support degree support (A) is:

$$\text{support(A)} = \frac{\text{D contains the number of item sets A}}{\text{The total number of items in D}} \tag{1}$$

Frequent item sets: The minimum support threshold (min_sup) is set in the Apriori algorithm. When the support of the item set is greater than the minimum threshold, the item set is called a frequent item set. If this item set contains k items, this frequent item set is called a frequent k item set.

Supporting degree of association rules: Association rules refer to the probability of occurrence of item set B under the condition that item set A appears. Association rule A => B, then support (A => B) is:

$$\text{support(A => B)} = \frac{\text{D contains the number of A} \cup \text{B}}{\text{The total number of items in D}} \tag{2}$$

Confidence: For the association rule A => B, the confidence of the association rule is the percentage of the transaction that contains the B transaction in the transaction containing A in D, then the confidence (A => B) is defined as,

$$\text{confidence}(\text{A} \;=>\; B) = \frac{\text{P(A} \cup \text{B)}}{\text{P(A)}} = \text{P(B} \mid \text{A)} \tag{3}$$

The minimum support threshold (min_sup) and the minimum confidence threshold (min_conf) need to be set first. The useful association rule is then selected in terms of these two conditions. The algorithm flow is as follows, as shown in Figure 1:

1) Scan the database and count each item to form a candidate 1- item set;
2) Determine whether the item set formed by step one is greater than the support number, and form a frequent 1-item set;
3) Apply the Apriori_Gen algorithm to form a candidate 2-item setbased on the frequent 1 item set formed in step two;
4) Determie whether the item set formed by step three is greater than the support number and form a frequent 2-item set;

5) Repeat the second, third, and fourth steps to form a frequent k-item set;
6) When an empty set of frequent item sets occurs, the iteration ends;
7) Calculate support and confidence, and satisfy support (A => B) ⩾ min_sup and confidence (A => B) ⩾ min_conf. Generate association rules from frequent item sets.

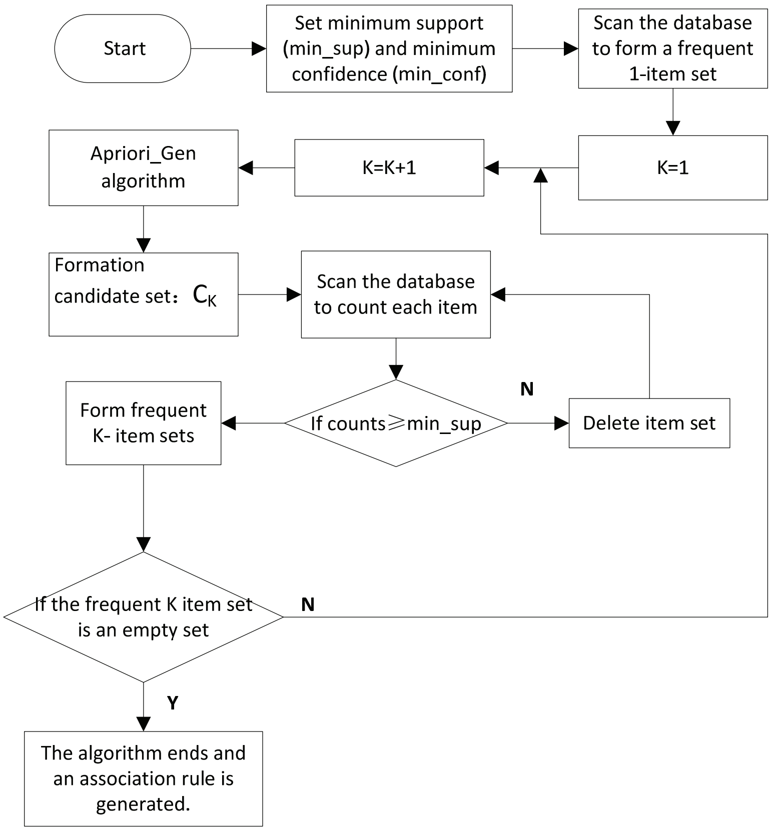
Start
Set minimum support (min_sup) and minimum confidence (min_conf)
Scan the database to form a frequent 1-item set
K=1
K=K+1
Apriori_Gen algorithm
Formation candidate set：C_K
Scan the database to count each item
If counts⩾min_sup
N
Delete item set
Form frequent K- item sets
If the frequent K item set is an empty set
N
Y
The algorithm ends and an association rule is generated.

Figure 1. Apriori algorithm flow chart.

3 Mining Test Scores Based on Apriori Algorithm

With the development of the computer network and information technologies, more and more information is collected and stored by the audited entities or individuals. However, auditors can only judge the effect of the test participants and the learning situation by scoring. Some hidden values or situations have not been fully explored. Therefore, we analysed the data of the "One learn and one test" test data of public officials in a certain district of Chengdu, and use the improved Apriori algorithm for analysis.

Taking into account the relevant circumstances of the audit, this paper will improve the Apriori algorithm accordingly.

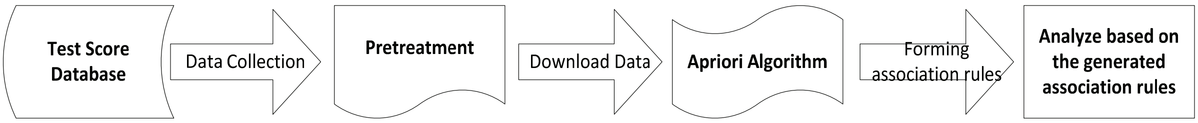

Figure 2. Data mining process.

Firstly, it is found that the number of questions answered each month is not the same as the number of questions answered based on the data obtained. Therefore, the monthly frequent item sets are calculated separately and the support and confidence are calculated to form a strong association rule every month.

Secondly, the month attribute is also added to the data, and all the data is integrated and analysed by Apriori algorithm to calculate the confidence and form a strong association rule.

Finally, the first two strong association rules are synthesized, and finally a global strong association rule is formed. The overall data mining process is shown in Figure 2.

4 Experimental Data Collection and Pretreatment

4.1 Collecting Data

The experimental data used in this paper is for the public officials in a certain district of Chengdu to participate in the “One learn and one test” test data from August to December 2018. The data information includes test submission time (ST), answer time (AT), score (S), test questions, etc.

4.2 Data Preprocessing

The main job of data pre-processing is data cleaning and conversion.

The main job of cleaning is to format the useful data, convert it into a unified format, delete some obviously problematic data, and at the same time remove the extraneous features and reduce the amount of algorithm calculation. This can avoid the use of useless information into the algorithm and improve the accuracy of the actual model. The characteristics selected in this experiment are divided into the following three items: Submission Time (ST), Answer Time (AT), and score (S).

According to the standard of working time and non-working time, the submission time is divided into five time periods, which are before the morning: 00:00-09:00, morning: 09:00-12:00, Noon: 12:00-14:00, afternoon: 14:00-17:00, evening: 17:00-00:00.

For the time attribute of the answer, the normal answer time (unit: second) is 180-600sin terms of the communication with the auditor and the actual experience, so the answer time is divided into extremely short: 0-180s, normal: 180-3600s Very long: >3600s.

For the score attribute, it is divided into three levels according to the answer question: Excellent: 80-100, Good: 60-80, Fail: 0-60.

The processed data is shown in Table 1.

Table 1. Discredited data table. The overall data distribution used in the experiment.

	ST	AT	S
1	afternoon	normal	Excellent
2	morning	very short	Excellent
3	morning	very short	Excellent
4	morning	normal	Excellent
5	evening	very short	Excellent
……	……	……	……
16103	Before morning	normal	Fail
16104	noon	normal	Excellent
16105	afternoon	normal	Excellent
16106	noon	very short	Good

The total number of data used in this experiment was 16,106.Discretization of data can better mine the intrinsic relationship between data. The pre-processed data is loaded into the Apriori algorithm, and finally a strong association rule is formed.

4.3 Experimental Results

In the Python locale, the Apriori algorithm is written for testing, and the minimum support threshold and the minimum confidence threshold are set to perform association rule mining. The minimum support is set to 0.06 and the minimum confidence threshold is set to 0.6.

According to the set thresholds, the results of the party members' test scores for three consecutive months were respectively obtained, and the following results were obtained, as shown in Table 2 to Table 4.

Table 2. Data mining results for the first month

	Rule	conf
1	frozenset({'long'}) => frozenset({'Fail'})	0.892
2	frozenset({'very short '}) => frozenset({'Excellent'}	0.726

Table 3. Data mining results for the second month

	Rule	conf
1	frozenset({'morning '}) => frozenset({'Excellent '}	0.615
2	frozenset({' normal '}) => frozenset({'Fail'}	0.747
3	frozenset({' very short '}) => frozenset({' Excellent '}	0.698
4	frozenset({'evening'}) => frozenset({'Fail'}	0.661

Table 4. Data mining results for the third month

	Rule	conf
1	frozenset({'long '}) => frozenset({'Fail '}	0.842
2	frozenset({' normal '}) => frozenset({'Fail'}	0.690
3	frozenset({' very short '}) => frozenset({' Excellent '}	0.717
4	frozenset({'evening'}) => frozenset({'Fail'}	0.642

It can be seen from the mining results of the above association rules that in the first month of the data, the good and bad results are only related to the time of the answer. If the answer time become longer, the result will be unsatisfactory. If the answer time is short, the result will be excellent. In the second month of the data, the quality of the results is related to the time of the answer and the time of submission. If the answer time is normal and the submission time is in the night, the test score is unsuccessful. When the answer time is extremely short and the submission time is in the morning, the test score is excellent; In the third month of the data, the rules obtained with this part of the data are the same as in the previous two months.

By combining the above association rules, the following conclusions can be drawn: First, when the time of the participants in the test is very short, their test scores are excellent. This situation is not reasonable. According to this rule, the auditor can further judge whether the participating tester predicts the answer in advance and whether the attitude of the person to be tested is correct. Secondly, the person who answered the question in the regular time will also fail. Auditors can focus on this type of person with this rule, and use this as a basis to monitor their normal learning.

5 Conclusion and Future Work

Under the big data environment, potential useful values can be tapped from party members' test data. The Apriori algorithm is employed to pre-process the party members' test data and mine the association rules, and provide auditors with ideas and work suggestions in this paper. Through the tested results information, mined scores related factors, generated association rules, it can provide reference for audit work and management.

The effectiveness of the proposed model remains to be further validated in future since the number of data features used in this experiment is small, and the associated association rules are not enough. In addition, the application of Apriori algorithm in the auditing industry may be improved by adding more features and enriching the data.

Acknowledgments. This work was supported by grants from the National Natural Science Foundation of China #71774134 and #71373216, in part by the Innovation Scientific Research Program for Graduates in Southwest Minzu University (No.CX2019SP33)

References

1. Wu Xiaodong, Zeng Yuzhu. University Students' Achievement Data Mining Based on Apriori Algorithm[J]. Journal of Langfang Teachers College (Natural Science Edition), 2019, 19(01): 33-38.
2. Ji Lianen [1,2], Gao Fang [1,2], Huang Kaihong [1,2] , et al. Visual exploration and analysis of university curriculum scores for multi-agents[J]. Computer Aided Design & Journal of Graphics, 2018.
3. Hu Zuhui, Shi Wei. Research on College Students' Online Behavior Analysis and Data Mining[J]. China Distance Education (Comprehensive Edition), 2017(2).
4. He Chu [1], Song Jian [1], Zhuo Tong [1]. Curriculum relevance classification model based on frequent pattern spectral clustering and student achievement prediction algorithm[J]. Journal of Computer Applications, 2015, 32(10): 2930-2933.
5. Peng Hailei. Research on the Relationship between Online Education Teacher Behavior and Student Learning Effect[J]. Journal of Northwest Normal University (Social Science Edition), 2018, v.55; No.260(04):111-117.
6. Cao Yuyu, Cao Weiquan, Li Wei, et al. A method for predicting college students' performance against uncertain missing data [J]. Modern Electronic Technology, 2018.
7. Du Qiu, Wu Ying. Application of Association Rules Based on Apriori Algorithm in Enterprise Personnel Management System[J]. Journal of Science and Technology Monthly, 2008(12): 136-138.
8. Sun Ting. Research on the law of medical drugs in insomnia based on Apriori algorithm [D].
9. Jia Kebin, Li Hanzhen, Yuan Ye. Application of Data Mining Based on Apriori Algorithm in Mobile Medical System[J]. Journal of Beijing University of Technology, 2017(3).
10. Zhao Hongli. Research on improved Apriori algorithm in psychology analysis of college students [D]. 2015.
11. Wang Hua, Liu Ping. Application of Improved Association Rules Algorithm in Student Score Early Warning[J]. Computer Engineering and Design, 2015(3): 679-682.