Date Mining Model Based on Association Rules Algorithm

Yuelei Li¹, Zhen Xu²

¹School of Electrical Engineering & Mechano-Electric Engineering, Xuchang University, Xuchang 461000, Henan, China
²Institute of Information Technology, Zhejiang Shuren University, Hangzhou 310014, Zhejiang, China

Abstract
The basic Apriori algorithm has the problems of long run time, occupying large memory space and low accuracy. This paper proposes a date mining model based on a fuzzy set and two-dimensional Hash Apriori optimization algorithm. First, we find out the minimum support of no less than the presupposition set, through the fuzzy association rules, interest rate will be final output with the biggest rule, in order to improve the accuracy of the original algorithm of rule mining, then by using two-dimensional Hash, remove those unneeded candidate item sets, which reduces the algorithm's time and space overhead, then in the generation of frequent item sets, it will adaptive assignment which decreases the original algorithm’s load, in the end, we apply the improved algorithm proposed in this paper to the information data mining. Simulation results show that Apriori optimized algorithm based fuzzy set and the two-dimensional Hash proposed in this paper has less memory consumption and faster running speed than the standard Apriori algorithm, the model is applied to the information in data mining which has greatly increased the accuracy of the rules mining.

Key words: Fuzzy Set Optimization, Two-dimensional Hash, Candidate Item Set, Adaptive Assignment, Date Mining.

1. INTRODUCTION

In today's university education system, students are consumers inside of school, they are the key parts to solve the above-mentioned pressure, while enjoying the service that school this distributors bring to them, they also hope that it can offer them diversified methods, let them acquire knowledge freely and easily (Liu, 2014). Therefore, the increasing working effort and various kinds of pressures of educator urge universities changing the management pattern, and they begin to learn credit management system from foreign universities, we propose the construction target of digitized campus, information-based and networked of student information management (Yan, 2014). With the development of digitized campus, how to use student data stored in the digital campus to improve management efficiency, quality and school’s management has become a very practice significance work (Li, 2014).

The domestic and abroad scholars have made many achievements in date mining issues. At present the abroad universities have applied data mining technology to management, which includes predicting failing student (Ye, 2014), university enrollment (Yu, 2014), optimizing the curriculum (Huang, 2014), student graduation (Ma, 2014), assessment of learning outcomes (Li, 2014), investigation of university life learning experience (Zhang, 2014) etc. These applications provide a strong guarantee to improve the quality of students (Xiao, 2013), level (Liu, 2013) and the quality of talent cultivation (Huang, 2014) for universities in foreign countries. The United States California University using "admissions integrated review" system analyze the admissions data of "race, language, application profession, family condition, high school ranking and scores", and do targeted propaganda to students which the enrollment probability above 40%, and made the California University enrolment reach to 22%, over twice admissions of congeneric schools (Fang, 2013). Compared with foreign countries, China’s research on data mining is still at an early stage, there is no deeper application in the management of institutions of higher education and no mature pattern of educational management system with data mining function for educational administration staff. But China is continuously carrying on the research on data mining in educational administration management. Dong Ping adopted the improved association rule Apriori algorithm to analyze student each class’s result, she discovered curriculum relation as well as the foundation curriculum’s influence to professional course, thus provide reference to school educational administration department to reasonably arrange the curriculum (Qi, 2014). Yang and Luo used decision tree algorithm to carry out analysis of the network college student’s achievement, found that the evaluation of learning process namely the usually results are more critical than the final results, according to the results of this study, the methods of letting teacher to strengthen the guidance through learning, focusing on usual learning situation and feedback learning information timely, improve the quality of the network school (Zhou, 2013).
Aiming at the defects of association rule Apriori algorithm, put forward the data mining model based on improved Apriori algorithm, and adopt the fuzzy set and two-dimensional Hash to optimize the accuracy of candidate item sets of Apriori algorithm.

2. ASSOCIATION RULE MINING ALGORITHM

Apriori algorithm using the iterative method, starting from the 1-item set, according to the given support threshold \( \text{min\_sup} \), pruning the frequent item set 1-, and finding frequent 1-item set \( L_1 \). According to prior principle: if there has a certain frequent item set, then its all subsets should be frequent too. Therefore when producing candidate 2-item, it will record as \( C_2 \), we might use the frequent 1-item set of \( L_1 \) to produce directly. After the generation of candidate 2-item set, according to the given \( \text{min\_sup} \) pruning to the candidate 2-item set \( C_2 \), then produces the frequent 2-item set \( L_2 \). Analogizes in turn, according to \( L_2 \) produce \( C_3 \), pruning \( C_3 \) produce \( L_3 \) ... Until producing the most item of frequent item set \( L_k \). As mentioned above, the process of Apriori algorithm mining rule also may divide into two steps to realize:

1. Finding all the frequent item set \( L \) in data set.
2. Extracting strong rule from frequent item set \( L \).

Step (1) is the key of Apriori algorithm, it decides this algorithm performance’s evaluation, the realization of step(2) is relative quite simple. At present, the improved methods of Apriori algorithm aim at step (1) in majority. The realization of step (1) may again divide into two operations. The first operation produces the candidate item set \( C \), the second operation is pruning the produced the candidate item set \( C \) on the basis of \( \text{min\_sup} \), and then we can find the frequent item set \( L \).

The generation of candidate item sets also have a lot of implementation methods, mainly have: a brute force method, \( F_{k-1} \times F_{k} \) method and \( F_{k-1} \times F_{k-1} \) method etc..

1. Brute force method

If we need to produce the candidate item set \( k- \), the brute force method will arrange and combine all 1-item set, listing all possible candidate item sets. If there has \( n \) 1-item set, then it can produce \( C_n^k \) candidate item set, then trimming off the nonessential candidate item set. Thus it can be seen, although the generation of candidate item set of this method is extremely simple, the operations are extremely complex, because the candidate number of pruning is too big.

2. \( F_{k-1} \times F_{k} \) method

This method uses the combination of \( L_{k-1} \) and \( L_k \) to produce the candidate \( k \) -item set \( C_{k-1} \). Figure (1) demonstrates the process of how to use this method of frequent 2-item set and frequent 1-item set’s combination to produce the candidate 3-item set.

![Figure 1. Generate candidate 3-itemsets process](image)

But because \( L_2 \) in this method is produced by \( L_{2-1} \) and \( L_1 \)’s combination, so it will inevitable produce the repetition of candidate item set.
(3) $F_{k-1} \times F_{k-1}$ method

In this method the candidate $k$–item set is produced by merging a pair of frequent $(k-1)$–item set, and this pair of frequent $(k-1)$ item set must satisfy $k-2$’s two former items are the same. The basic idea can be shown as Equation(1) and (2):

$$A = \{a_1, a_2, ..., a_{k-1}\}$$

(1)

$$B = \{b_1, b_2, ..., b_{k-1}\}$$

(2)

When they satisfy the following conditions, combining $A$ and $B$. The basic idea can be shown as Equation(3):

$$a_i = b_j (i=1,2,3,...,k-2), a_{k-1} \neq b_{k-1}$$

(3)

Figure 2 demonstrates the process of how to use this method to produce a pair of frequent $2$–item set combining into the candidate $3$–item set.

Because this method by combining a pair of frequent $(k-1)$–item set to get candidate $k$–item set, so before combination, we need to add one more step to ensure that frequent $(k-1)$–item set’s $(k-2)$ is the same.

The generation of Apriori algorithm’s frequent item sets has two important characteristics: first, it is a layer-by-layer iterative process (level-wise), which is from frequent 1–item set to the maximum frequent item set, each time the generation of a new frequent item set should traverse the transaction sets; secondly, it use generated pruning rule to produce frequent item sets. In each iteration, the new generated candidate items should use the frequent item set which the last time finding, and then calculate the count of support of each candidate item set, then compare to given support threshold value, remove candidate item sets which the support value is less than the support threshold value.

The Apriori algorithm based on frequent item set using the iterative method to search step by step. Algorithm is simple and clear, no complicated theoretical derivation, also it is easy to implement. However, it has some deficiencies:

(1) In the process of generating frequent item sets, it needs scan database over and over again, which causing enormous load to input and output. Especially in the case of long set, algorithm needs to individually scan each subset of $k-1$ candidate set and then match them, check whether these subsets are belong to $L_{k-1}$, if there is one does not belong to $L_{k-1}$, it will be reduced, this operation will cause a huge amount of computation, and causing enormous load to input and output, and also the running time is long, the efficiency is very low.

(2) In the process of connecting $L_{k-1}$ to produce $C_k$, it will generate very huge candidate item sets. The Apriori algorithm to generate $k$–candidate item set should use for self connection of $k-1$–frequent item set.
The operation should judge whether the \( k - 2 \) item is the same as \( k - 1 \) or not, this operation will occupy a significant running time, the efficiency of the algorithm is relatively low.

(3) The frequent item set which lower than the minimum support threshold value is difficult to discover, in this entire process, as long as the candidate items are smaller than the minimum support value should be delete. However, this deleted information may also valuable. If decrease the minimum support threshold value, it will be able to produce a huger candidate item set, the efficiency also will be a big problem.

3. IMPROVEMENT OF APRIORI ALGORITHM

3.1. Optimization of Accuracy Based on Fuzzy Set

The fuzzy concept applies the data mining algorithm is the process of promoting Apriori algorithm to the fuzzy project, it will discover the minimum support of item set which its support is not smaller than supposed, then through the fuzzy connection rule, output the interest rate in biggest rule.

FARMA () enter \( n \) transaction database \( D \) consisting of transaction, each transaction data requires \( m \) projects; \( j(j = 1, 2, \ldots, n) \) transaction in \( p(p = 1, 2, \ldots, m) \) project can be described as \( k \) 's \( u_j^p(R_p^s) \) \( (s = 1, 2, \ldots, k) \); the default minimum support value and \( \min confidence \).

Output one pair fuzzy association rule.

Step1. Each project \( t_j^p(p = 1, 2, \ldots, m) \) of transaction data \( T_j(j = 1, 2, \ldots, n) \) in \( D \), \( k \) fuzzy concept corresponding to \( k \) fuzzy set \( R_p^s (s = 1, 2, \ldots, k) \), taking the given membership function as fuzzy sets. Set \( j \) in \( T_j \) that \( P \) project the formation of fuzzy set is \( f_j^p \), then \( t_j^p \) described by Zadeh as fuzzy set program. The basic idea can be shown as Equation(4):

\[
f_j^p = \mu_j^p(R_p^s) / R_p^s + \mu_j^p(R_p^s) / R_p^s + \ldots + \mu_j^p(R_p^s) / R_p^s\]

Step2. Calculate the membership value’s weight of each item \( t_j^p(p = 1, 2, \ldots, m) \) in \( n \) transaction data community \( T_j(j = 1, 2, \ldots, n) \) the corresponding fuzzy set \( R_p^s (s = 1, 2, \ldots, k) \). The basic idea can be shown as Equation(5):

\[
weight_j^p = \sum_{j=1}^{n} \mu_j^p(R_p^s)\]

Step 3. Calculate the maximum fuzzy set \( R_p^{max} \) of the weight of each item, the basic idea can be shown as Equation(6):

\[
weight_{p}^{max} = \max_{j=1}^{k}(weight_j^p)\]

Then its mutual fuzzy set \( R_p^{max} \), in following model we use \( R_p^{max} \) to describe the project \( t_j^p \).

Step4. Inspects each fuzzy set \( R_p^{max} (p = 1, 2, \ldots, m) \) correspondence weight value \( weight_p^{max} \), to determine whether it is smaller than the \( \min sup port \) or not. If the weight value \( weight_p^{max} \) is not smaller than the \( \min sup port \), then we will put the fuzzy set \( R_p^{max} \) in the first major term \( L_t \), the basic idea can be shown as Equation(7):

\[
L_t = \{ R_p^{max} | weight_p^{max} \geq \min sup \} \quad (7)\]

Step5. Set \( r = 1 \), here \( r \) expresses the quantity of current retention in major term centralism project.

Step6. With the algorithm which is similar to Apriori algorithm to produce the candidate major term set \( C_{r+1} \) from \( L_t \). The algorithm inspects \( L_t \) firstly, preset \( L_t \) has \( r - 1 \) same in two item sets, keep it in \( C_{r+1} \) in the project set, and this set includes all \( r \) project subsets exist in \( L_t \).

Step7. Set the candidate major term \( C_{r+1} \)'s \( r + 1 \) project is \( t = (t_1, t_2, \ldots, t_{r+1}) \), the process for the set produced recently in each candidate project set:

(1) Set \( \mu_j^p(R_p^{max}) \) is membership degree in fuzzy set \( R_p^{max} \) of \( P \) project in \( T_j \), the membership in project calculation of each event’s candidate item \( t \in (r=1, 2, \ldots, r+1) \) is \( \mu_j^p = \mu_j^p(R_p^{max}) \cap \mu_j^p(R_p^{max}) \ldots \mu_j^p(R_p^{max}) : \) but there must exists some intersection, then the basic idea can be shown as Equation(8):
\[ \mu_i^r = \min_{\mu_i^j (R_i^m)} (8) \]

(2) weight calculation of candidate item: \( \text{weight}_i = \sum_{j=1}^{n} \mu_i^j \).

(3) If \( \text{weight}_i \) is not smaller than \( \minsup \), then puts \((t_1,t_2,\ldots,t_{r+1})\) in the project \( L_{r+1} \).

Step8 If \( L_{r+1} \) is empty, then continue to process step 9; otherwise make \( r = r+1 \) and do step 6 and 8 again.

Step9 Make association rules for big \( i \) \( q(q \geq 2) \) tem set \( t \) which contains \((t_1,t_2,\ldots,t_q)\), the steps are as follows:

1. Calculate all possible situations: \( t_1 \land \ldots \land t_{r-1} \land t_{r+1} \land \ldots \land t_q = t_k \), \( k = 1,2,\ldots,q \).
2. Calculate all confidence factors, the basic idea can be shown as Equation(9):

\[ \sum_{j=1}^{n} (\mu_i^j \land \mu_o^j, \mu_i^j, \mu_o^j) = \sum_{j=1}^{n} \mu_i^j \]

Step10. Output confidence factor (confidence factors) is not less than the minimum confidence degree.

Among them, some symbols and variables in the algorithm, \( \text{weight}_i^r \) is the sum for membership; weight value \( \text{weight}_r^\text{max} \) is the maximum value in \( \text{weight}_r^j \); \( L \), the large item set for the project (attribute) \( r \), referred to as large \( r \) item set; \( C \), the candidate large for the project (attribute) \( r \), referred to as the candidate large \( r \) item set.

Association rules output by the mining algorithm with fuzzy attribute, it can be used as potential meta knowledge of the default transaction database of \( D \). Fuzzy rules are most widely, the most popular in expert system, it is based on reasoning and acceptable accuracy, and has good protection.

3.2. Optimization of Candidate Item Set Based on Two-dimensional Hash

The Apriori algorithm shows that before item set \( k \) produced in each step, it must be first based on frequent item set \( k-1 \) generate the corresponding candidate set \( C_k \). Then search database to calculate each candidate item set’s support degree, then generate frequent item sets, it needs great cost of time and space, so to generate a smaller candidate item set play a key role on improving efficiency of discovering frequent item set, but in the Apriori algorithm, candidate set \( C_k \) is generated through \( L_{k-1} \), then the potential of \( C_k \) is very large. By using Hash technology, remove unnecessary candidate item set to reduce the potential of \( C_k \), therefore to reduce the algorithm’s cost in time and space, so as to improve the efficiency of the algorithm.

Apriori algorithm shows that when project is very large in event \( D \) and \( I \), there may be a large item in \( L_k \), we set it as \( m \), then the candidate \( 2 \)-item set \( C_2 \) number generated by self connection of \( L_k \) is \( \frac{m(m-1)}{2} \), then it will be a great amount of computation of calculating support degree for \( C_2 \).

Using the Hash technology to solve “the conflict” issue, this paper gives a two-dimensional hash function, which avoided “the conflict” perfectly.

Item set \( I = \{I_1, I_2, \ldots, I_k, \ldots, I_m\} \) assigns a sequence value to each project \( I_k (k = 1,2,\ldots,m) \), which is \( I_1, I_2, \ldots, I_k, \ldots, I_m \), then sequence value is \( 1,2,\ldots,k,\ldots,m \) respectively. \( \text{order}(x) \) and \( \text{order}(y) \) express the sequence value of two projects \( x \) and \( y \) of generated \( 2 \)-item set. Given a new two-dimensional Hash function:

\[ h_1(x,y) = (|L| \times \text{order}(x) + \text{order}(y) - \frac{x(x-1)}{2}) \mod p_1 \]

\[ h_2(x,y) = (|L| \times \text{order}(x) + \text{order}(y) - \frac{x(x-1)}{2}) \mod p_2 \]

Among them

\[ p_1 \times p_2 \geq (1 - \minsup) \min \]

\[ p_1 \neq p_2 \text{ and } p_1 \text{ and } p_2 \text{ coprime.} \]
The value of \( p_1 \) and \( p_2 \) is too large to occupy too much space, and too small will cause the conflict. The value of \( p_1 \) and \( p_2 \) according to the number of item and the minimum support degree that user sets to change accordingly.

\[
H(x, y) = H(h_1(x, y), h_2(x, y))
\]  

(7)

The value of \( h_1(x, y) \) and \( h_2(x, y) \) represent subscript of \( H(x, y) \), when project one unit to Hash count, then add 1 to this unit count.

\[ |L| \text{ is item number.} \]

Function \( |L| \times order(x) + order(y) - \frac{x(x - 1)}{2} \) has uniqueness, using two-dimensional Hash table reduces “the conflict” of algorithm.

While scanning databases to generate \( L_i \), counting two-dimensional Hash to each item’s \( 2-t \) combination, when scanning accomplished, not only receive \( L_i \) but get a two-dimensional Hash table, each value in the table is an accumulation of a count. Compare count value to \( \text{min}_\text{sup} \), if the count value is bigger than or equal to \( \text{min}_\text{sup} \), then this \( 2-t \) combination belongs to \( L_k \), otherwise it is not frequent \( 2-t \) item set.

### 3.3. Resource Optimization Based on Adaptive Assignment

The standard Apriori algorithm cannot solve the issue of occupation of resource, in order to solve it, this paper has done the following treatment in the process of item sets mining: when improving the Apriori algorithm in the frequent \( k-1 \) item set generates frequent \( (k+1)- \) item set, it will assign a weight dynamically. When the frequent \( k \) item set generates frequent \( (k+1) \) item set, frequent \( k \) item set according to the number of descending in transaction database. The \( \text{len}(\text{freq}_i) \) represents the total number of frequent \( k \) item set, the \( i \) frequent \( k \) item set’s weight is defined as:

\[
\text{weight}(I_i) = \text{len}(\text{freq}_i) - i - 1
\]

(8)

All frequent \( k \) item sets’ total weight are:

\[
\text{TotalWeight} = \sum_{i=0}^{\text{len}(\text{freq}_i) - 1} \text{weight}(I_i)
\]

(9)

Suppose there’s \( P \) available resource, each resource processes subsets of \( k-1 \), these subsets’ weight are close to \( \text{TotalWeight} / p \). For simplicity, put these \( k-1 \) frequent to the first resource, until these frequent sets’ weight greater than \( \text{TotalWeight} / p \). Other item sets assign to the second resources in the same way. Then repeat the above allocation process, until all item sets’ assignment are finished. According to this assignment, the candidate set number of resource is basically the same. Assuming there are \( m \) frequent \( k \) item sets, after descending order in accordance with support of item sets, teach item set is \( I_0, I_1, \ldots, I_{m-1} \) respectively; there are \( P \) resource, they are \( p_0, p_1, \ldots, p_{\text{p}-1} \) respectively. Item sets connect with \( I_0 \) are \( I_0, I_1, \ldots, I_{m-1} \) respectively, the number is \( \text{len}(\text{freq}_i) - 1 \) which equal to weight of \( I_0 \) set. If \( I_0, I_1, \ldots, I_{m-1} \) is the item set assigned to the first resource, so the total number of generated candidate set is \( \sum_{i=0}^{m-1} \text{weight}(I_i) \). The number is equal to the total weight of \( I_0, I_1, \ldots, I_{m-1} \). So the number of each resource generated candidate \( (k+1) \)-item set equal to the weight sum of its all assignment frequent \( k \) -item set. Because the weight sum of each resource processes frequent \( k \) -item sets close to \( \text{TotalWeight} / p \), so the number of generated candidate \( (k+1) \)-item set is basically the same, so each resource load is balanced.

### 4. SIMULATION OF ALGORITHM PERFORMANCE

In order to confirm the effectiveness of this improved algorithm proposed in this paper, we carry on the simulation experiment to it. In order to test the disparity of improved algorithm proposed in this paper and the original algorithm, this paper uses 3 different data quantities’ databases to carry on the performance test. Among them, the NO.1 database’s data quantity is 100, the NO.2 database’s data quantity is 1000, two algorithms’ speed test results are in Figure 3-5.
It can be seen from the above figures, with the increasing of the amount of data, the improved algorithm proposed in this paper is better than the original algorithm; it saves computational overhead, so as to improve the execution efficiency of the algorithm.

Then using the improved Apriori algorithm that proposed in this paper to carry on the association rules mining to the data, the final results are as follows.

**Table 1. Data mining results**

<table>
<thead>
<tr>
<th>Rules prefix</th>
<th>Rules suffix</th>
<th>Support</th>
<th>Confidence</th>
<th>influence</th>
</tr>
</thead>
<tbody>
<tr>
<td>The school year =4</td>
<td>Job = monitor</td>
<td>0.13</td>
<td>0.64</td>
<td>1.26</td>
</tr>
<tr>
<td>Score = good</td>
<td>Subjects through rate&gt;90%</td>
<td>0.21</td>
<td>0.62</td>
<td>1.96</td>
</tr>
<tr>
<td>Attendance&gt;80</td>
<td>Fail rate&gt;40%</td>
<td>0.21</td>
<td>0.62</td>
<td>1.20</td>
</tr>
</tbody>
</table>

It can be seen from the simulation results, the improved Apriori algorithm proposed in this paper can have a very good performance in mining the useful information from data in practical application.

**5. CONCLUSIONS**

At present, the administrators of institutions of higher learning are merely through the simple data statistics as well as sorting basic functions to gain surface information, the further information hidden in these huge databases are not discovered and used. How to go facing these data and to carry on the advanced data processing, to discover the rule of school administrators from these data, therefore to help administrators to better to carry on the decision-making, to promote institutions of higher learning’s management level as well as quality, it becomes the major issue in management of institutions of higher learning. This paper proposed data mining...
model based on fuzzy set and two-dimensional Hash optimized Apriori algorithm, it can be seen from the simulation result, the performance of algorithm is good and worth promoting.

Acknowledgements

This work was supported by the Department of Science and Technology of Zhejiang Province (Grant No. 2014C31065), the Department of Science and Technology of Zhejiang Province (Grant No. 2015C31091), and the Department of Science and Technology of Zhejiang Province (Grant No. 2016C31116).

REFERENCES

Zhan Haiju (2014) “Information Query System Based on National Distance Education Standard and Data Mining”, Computer Simulation, 7(7), pp.313-316.