International Journal of Database Theory and Application
Vol.9, No.7 (2016), pp.233-242
http://dx.doi.org/10.14257/ijdta.2016.9.7.20

Research on the Impact of Advanced Data Mining Algorithm on Physical Education Quality

Nan Wang

(sports department of Zhengzhou University,China,450001)
henanzzlinan@sina.com

Abstract

Concerning the condition that there is a glittering array of disadvantages such as frequent candidate collection of Apriori algorithm, this paper comes up with cost-sensitive filtering matrix Apriori algorithm based on weighting. What's more, with the help of FP-tree algorithm, we can carry out cost-sensitive learning through relevant data of its constructed decision tree to set different weighting for data and confidence level;

Keywords: *Apriori; Decision Trees; Strong Association Rule*

1. Introduction

Association rule is a key part of data mining, which mine relevant information by studying the storage information in transaction database. Literature [1] proposed to mine Boolean association rules and Aprior algorithm based on frequent sets. With the proposals of cloud computing, big data and other concepts, researches on data mining based on association rule become more significant. Literature [2] constructed a web data mining model that supported the parallel association rule that integrated with the idea converging local computing of storage nodes. Experimental results show that the model can improve the efficiency of web data mining, and the implementation rate increases as the increases in data volume. Literature [3] proposed a preference sensing algorithm based on the least regret, and experiments show the effectiveness of the algorithm. Literature [4-6] proposed to carry out different improvements against the deficiency of Apriori algorithm and achieved certain results. Based on Apriori algorithm, this paper puts forward non-frequency cost-sensitive filtering matrix Apriori algorithm based on weighting. Simulation experiment shows that the proposed algorithm has certain advantages and obtained better results by applying to researches on data mining of higher vocational teaching quality.

2. Apriori Algorithm Based on Association Rules

2.1. Apriori Algorithm

Apriori is composed according to priori knowledge of frequent item set characteristics, which mainly contains two meanings: one is that if the item set I does not meet the minimum support threshold, the item set I is not frequent item set; the other is that if an item set A is added to item set I, the frequency for new IUA in database will be less than or equal to the frequency of original item set I. Therefore, when a set of data cannot be verified, the subset added cannot be verified as well.

For the association rule in the given item set of K $\phi \to \psi$, U is the finite non-empty set, and the setting probability is P, the association rule contains 3 factors: the rule intensity factor, the confidence factor and the coverage factor.

ISSN: 2005-4270 IJDTA

(1) The rule intensity factor: Assume $\sup p_k(\phi,\psi) = card(\| \phi \wedge \psi \|_k)$, the rule intensity is $\sigma_k(\phi,\psi) = \frac{\sup p_k(\phi,\psi)}{card(U)} = \frac{card(\| \phi \wedge \psi \|_k)}{card(U)}$, and the rule intensity factor reproposeds the rule proportion in the entire decision making process.

(2) The confidence factor: If $x \in U$, $P_u(x) = \frac{1}{card(U)}$, so the probability of any equation ϕ in K is defined as: $\pi_k(\varphi \mid \psi) = P_u(\| \psi \|_k \| \phi \|_k) = \frac{card(\| \phi \wedge \psi \|_k)}{card(\| \phi \|_k)}$. Therefore, there is conditional probability in the association rule $\phi \to \psi$: $\pi_k(\varphi \mid \psi) = P_u(\| \psi \|_k \| \phi \|_k) = \frac{card(\| \phi \wedge \psi \|_k)}{card(\| \phi \|_k)}$. Where, $\| \phi \|_k \neq \varnothing$. $\pi_k(\varphi \mid \psi)$ is the confidence factor of decision rule $\phi \to \psi$, which is expressed as $incred_k(\varphi,\psi)$. Therefore, $incred_k(\varphi,\psi)$ reflects the confidence level that the association rule $\phi \to \psi$ is true.

(3) The coverage factor: Express $\pi_k(\varphi \mid \psi) = P_u(\| \psi \|_k \| \phi \|_k) = \frac{card(\| \phi \wedge \psi \|_k)}{card(\| \phi \|_k)}$ as $\mathrm{cov}_S(\phi,\psi)$. The coverage factor reflects the confidence level that the inverse rule $\psi \to \phi$ of $\phi \to \psi$ is true, which determines the confidence level of this decision reasons through given conditions.

2.2. Analysis of Algorithm Deficiency

Although the Apriori algorithm can mine data very well to a certain extent, the algorithm has great temporal and spatial complexity because there are a lot of candidate frequent set when scanning database and subsets of those frequent items also have features of super sets, reducing the mining performance of Apriori algorithm. These deficiencies may be altered by introducing FP-Growth algorithm. The main idea is to compress the database data into a frequent pattern tree and break them down into databases with different conditions for mining. Its features are as follows:

(1) The root node is the control node. Each child node excluding the root node contains 2 parts. One is the set of item prefix subtree, and the other is the header table composed by frequent item;

The node for each item prefix subtree contains 3 parts, which are respectively used to identify the item name of the prefix, the number of node supports contained in each prefix and the pointer field pointing to father node. The record that the item node records this node is the transaction quantity;

Each frequent item table contains two fields: namely, item name and node pointer. Where, the item name refers to the name of frequent item, and the node pointer points to the father node with the same item name in the frequent pattern tree.

Suppose a database is shown in Table 1, the minimum support number is set as 3, and FP-tree after a scan is constructed as Figure 1.

Table 1. Database Example

Id	Item Set	Frequent Items
1	a,b,c,d,e,f,g,i,p	b,c,a,m,p
2	a,b,c,d,m,o,f	c,b,c,m,f
3	a,b,j,o	a,j
4	a,k,s,p	b,p,c
5	a,c,e,d,l,m,n	l,c,b,n,e

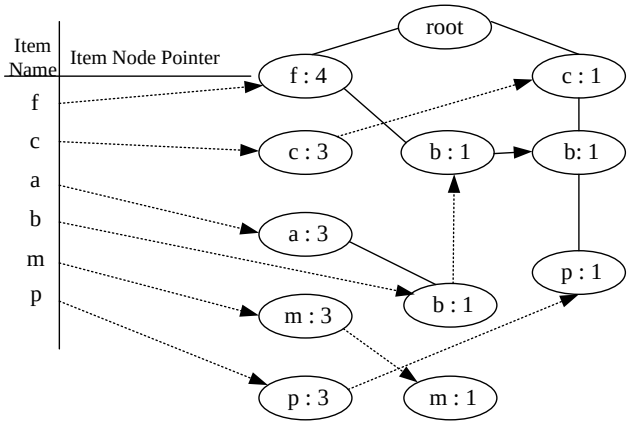

Figure 1. FP-Tree Structure

Compared with Apriori algorithm, FP-tree algorithm has lower frequency of scanning database, the required time is less than three, there will be no frequent item set, and the data retrieval process is simplified. Therefore, FP-tree algorithm improves the deficiencies of Apriori algorithm, yet the improvement effect only reduces the spatial complexity and there is little change in the overall performance of mining algorithm.

3. Advanced Apriori Mining Algorithm

To solve the deficiency of foregoing Apriori algorithm, this paper puts forward non-frequency cost-sensitive filtering matrix Apriori algorithm based on weighting. This algorithm mainly has 6 steps. (1) Carry out cost-sensitive learning over the data corresponded to decision tree built on the basis of FP-tree algorithm and obtain the cost effectiveness of data attribute; (2) Set the weighting for different data and set the confidence level for weighting; (3) Look for K frequency set through non-frequency filter matrix set, and generate strong association rules; (4) Set the initial matrix corresponded to non-frequency filter Apriori algorithm; (5) Build the non-frequency cost-sensitive filtering matrix Apriori algorithm, and increase the scanning times of database, namely, expand the range of finding records, (6) Determine whether the non-frequency filtering matrix satisfies the conditions of K frequency set; if the conditions are met, remove the candidate set with minimum supports; otherwise, continue the iteration.

3.1. Cost-Sensitive Learning

Cost-sensitive learning is a learning system balancing the correct classification costs and test costs. Due to attributes between different records or different difficulty levels in actual database, this method strikes a balance between information classification cost and test cost, and choose highly cost-effective attribute as the basis of classification. The result of cost-sensitive learning is the ratio of correct classification cost value and the test cost value, and the cost effectiveness of attribute i is expressed by $Cost(i)$:

$$Cost(i) = \frac{M - FP * \sum_{i=0}^{r} n_i - FN * \sum_{i=r+1}^{n} p_i}{TestCost(i)} \quad (1)$$

In the equation, $TestCost(i)$ stands for the test cost of attribute i, the numerator reproposeds the decrements of correct classification process corresponded to the chosen attribute i, and M refers to the cost that i is not chosen as correct classification.

3.2. Setting the Confidence Level of Weighting

The weighting refers to the sum of transaction item set value included in the transaction database, expressed as Equation (2)

$$WS(x) = \frac{F_x \prod_{k=1}^{|X|} (\forall [i[W] \in X]) t_i[i_k[w]]}{N \prod_{k=1}^{|X|} (\forall [i[W] \in X]) t_i[i_k[w]]} \quad (2)$$

In the equation, F_x is the frequency of record x, and N is the record number in database.

The confidence level of weighting mainly refers to the ratio of weighted supports meeting $X \bigcup Y$ in the transaction database and weighted supports including the weighting of X, and the equation is expressed as (3).

$$\begin{aligned} confid(X \to Y) &= \frac{WS(X \bigcup Y)}{WS(X)} \\ &= \frac{N_{(x \bigcup y)}}{N_x} \cdot \frac{\prod_{k=1}^{|X \bigcup Y|} (\forall [i[W] \in X]) t_i[i_k[w]]}{\prod_{k=1}^{|X|} (\forall [i[W] \in X]) t_i[i_k[w]]} \\ &\cdot \frac{\prod_{k=1}^{|X|} (\forall [i[W] \in X]) t_i[i_k[w]]}{\prod_{k=1}^{|X \bigcup Y|} (\forall [i[W] \in X]) t_i[i_k[w]]} \end{aligned} \quad (3)$$

3.3. Finding K-Frequency Set by Using Non-frequency Filter Matrix Set

By constructing the non-frequency filtering matrix, this paper removes irrelevant records when finding K-Frequency sets. Non-frequency filtering matrix is a Boolean matrix, the element $X_{i,j}$ reproposeds whether data records in the i-th line appear in the j-th item database, and the value is 0 or 1.

Principle 1: According to the pigeonhole principle, if the record number of database in K frequency set is less than K, the retrieval value of database must not be within the candidate K frequency set.

Principle 2: When the minimum support number required by database transaction is n, and when the retrieval number of database including an item is less than or equal to n, the retrieval record of the item database must not be within the candidate K frequency set.

The filter matrix is constructed as per below method:

Step 1: Check each entry of data information successively. When one record is scanned, a row vector for intermediate matrix is constructed. When the same i-th item data record appears, the value for the i-th element in the row vector is labeled 1; otherwise it is labeled 0.

Step 1: When finishing checking all database records, the last row vector and last column vector are constructed for intermediate matrix. Fill the sum calculated for each column into the last row vector, and similarly, fill the sum of each row into the last column vector. The last element for this matrix is set to 0, indicating no meaning.

Step 3: "Cut" the intermediate matrix. Process is as follows:

(1) According to Principle 1, delete from the row vector that its value of last element in the matrix is less than K, and re-modify data on the last row of matrix;

(2) Set the minimum support number. According to Principle 2, delete the column vector that its value of least element in the matrix is less than $\min$, and re-modify data in the last column of matrix;

(3) Constantly iterate the above procedures until the matrix cannot be "cut".

3.4. Generation of Strong Association Rules

The purpose of association rule is to find frequent K predicate set and then find all frequent predicate items meeting the minimum supports. After obtaining frequent K predicate set, seek the association rule meeting the minimum confidence level, and then generate the strong association rules. The process is shown in Figure 2.

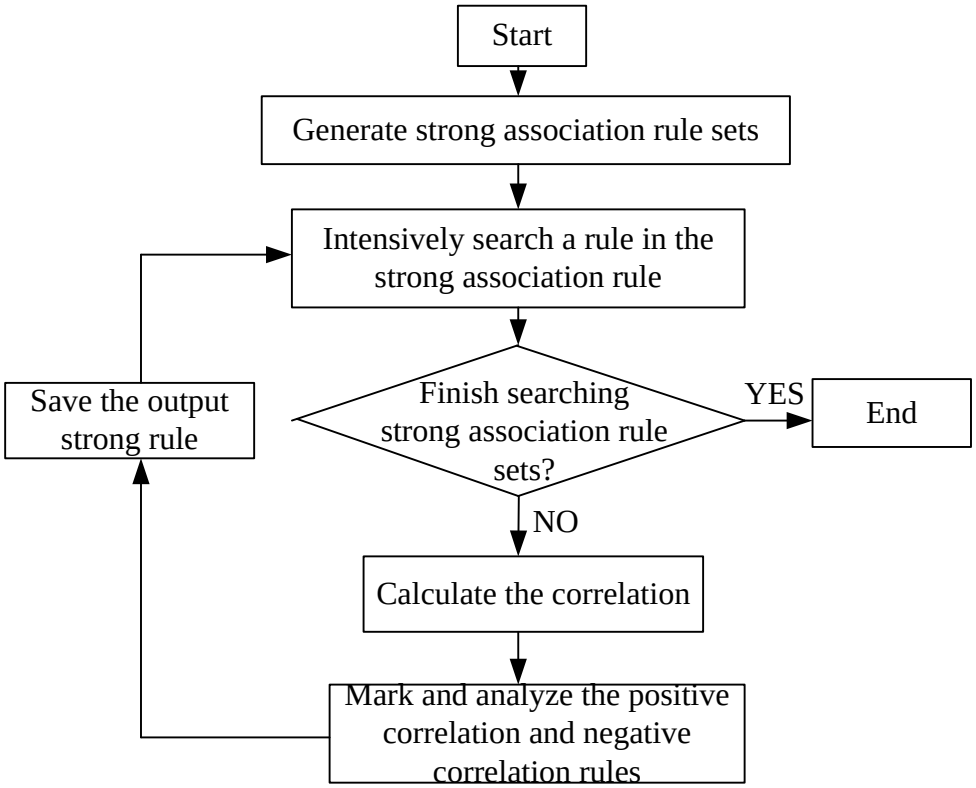

Figure 2. Flow Chart for Generation of Strong Association Rules

Description for Steps of Figure 2 Flow Chart for Generation of Strong Association Rules

Step 1: Generate K frequency set according to the non-frequency filter matrix Apriori algorithm;

Step 2: Choose one set from K frequency set for study;

Step 3: Choose one set from K frequency set as the result of association rule, and take the related attribute corresponded to database in the item as the result of association rules;

Step 4: Divide K frequency set into two frequency item sets of cause and effect, obtain the comparison results according to the confidence level of weighting $conf(x \to y) = \frac{\sup(x \bigcup y)}{\sup(x)}$, and calculate the ratio of K frequency set and effect in the matrix. If it is less than the minimum confidence level, find an association rule.

3.5 Initial Matrix Required by Non-frequency Filter Matrix Apriori Algorithm

In construction of rule matrix system, the initial matrix column required by non-frequency filter matrix Apriori algorithm will be generated through table look-up, and then the database is required to be scanned and copied item by item, in order to generate the initial matrix, generated as shown in Figure 3.

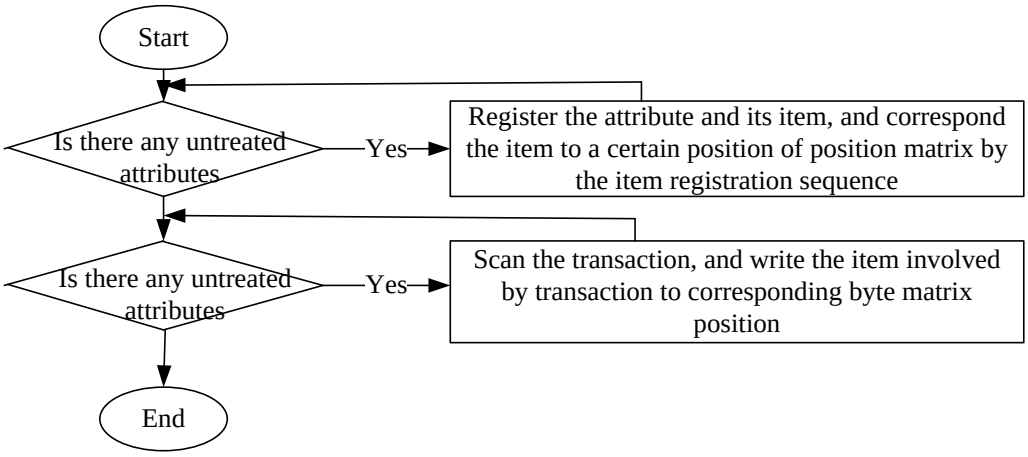

Figure 3. Construction of Initial Matrix

Convert the non-frequency filter matrix into non-frequency filter matrix with sequenced rows and columns, express as shown in Table 2 and Table 3, construct non-frequency filter matrix FF_M(m×n order) based on these two algorithms according to results from Table 4, and seek K frequency set, as shown in Table 3.

Table 2. Construct Intermediate Matrix

```
    MJZ [count,i]=Last_count    //Here calculate the last variable value in each row
    Calculate the last variable value in each column
      count++
    Last_count=0
```

Table 3. Generate Non-frequency Filter Matrix FF_M

```
                FF_M=M_A      //save the changed results to FF_M
            return    FF_M
          }
}
    When FF_M is converted and there is no change, it is the final result
                { delete(M_A[i])// delete rows not meeting the conditions
              }
                FF_M=M_A      //save the changed results to FF_M
            return    FF_M
          }
}
```


Table 4. Seek K Frequency Set Algorithm According to the Filter Matrix FF_M (m×n Order)

```
Procedure Des(lines,m)
 {flag=true
   if (k==0) then
     flag=false
     else{
     M_A=FF_M
   for(i=1;i<n;i++)
      for(j=i+1;j<n;j++)
      {Mid_A(row[i])
        Delete(M_A[i])    ///delete the column after removing i-th item
        while   ( | M_A[k]≠k)   and   flag
                k=k-1
            FS_search(M_A,min,k)
  }
return   Ck
}
```

4. Simulation Experiment

Experimental environment settings are Core i3 CPU, 4GDDR3 memory, 500G hard disk, WindowsXp operating system, related records involving vocational teaching quality for the data objects, and a total of 20,000 data records selected.

4.1. Application of the Present Algorithm

In order to better experience the effects of proposed algorithm on data mining improvements, this paper adopts teaching quality data as the main mining object, and divides the teaching quality items into 4 parts, including teaching contents, teaching methods, curriculum design and student assessment. Then, it sets up the number of input item as 4 and the minimum support min=0.5, mines the association rules for 4 data item contents, generates the frequency item set, and the item set is Item ={{I1},{I2},{I3},{I4}}, and the corresponding support (Count)={4,6,3,2 }. Through the parameter initializing database setting up by users, count the support number and candidate sets for 4 data item sets {I1,I2,I3,I4}, respectively as shown in Table 5 and Table 6; the frequency set lower than the minimum degree of support is shown in Table 7. Therefore, it is very important to adopt the proposed data mining method for teaching contents and curriculum design.

Table 5

Item Set	Support Number
I1	4
I2	6
I3	3
I4	2

Table 6. Candidate Set

Item Set	Support Number
I1	0. 72
I2	0. 71
I3	0. 62
I4	0. 52

Table 7. Frequency Set Lower than the Minimum Degree of Support

项集	支持数
I1	0. 72
I3	0. 62

4.2. Comparison with Other Related Mining Algorithms

According to Figure 4 suppose the setting of the same degree of support, compared with basic Aprior algorithm, the proposed algorithm has most stable and effective maximum frequency set mining capacity, and also has certain advantages in execution time. As the degree of support gradually increases, the algorithm shows stability.

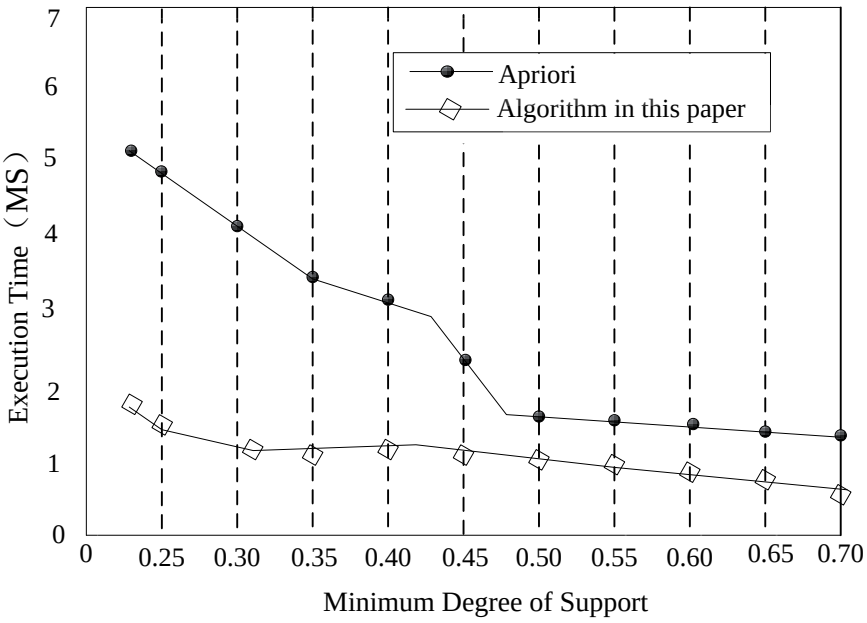

Figure 4. Comparison of Implementation Time

5. Conclusion

This paper puts forward a non-frequency cost-sensitive filtering matrix Apriori algorithm based on weighting. Compared with traditional Aprior algorithms, the proposed algorithm has obvious advantages, and satisfactory results will be achieved by applying the proposed algorithm to researches on higher vocational teaching quality.

References

[1] R. A. Srikan, “Fast algorithms for mining association rules in large databases”, Proceedings of the 20th International Conference on Very Large Data Bases San Francisca MorganKaufn ann Publishers, **(1994)**, pp. 487-499.
[2] L. Xiao and L. Y. Long, “Web data mining of association rules based on an improved iterative algorithm”, Science & Technology Review, vol. 33, no. 3, **(2015)**, pp. 90-93.
[3] S. Jing, “Preference perception algorithm in data mining based on regret minimization”, Computer Applications and Software, vol. 32, no. 5, **(2015)**, pp. 59-63.
[4] D. Li, “The Research about Data Mining of User's Behaviors Based on Apriori Algorithm”, Bulletin of Science and Technology, vol. 29, no. 12, **(2013)**, pp. 214-216.
[5] W. Ling and W. Y. Jiang, “Improved Apriori Algorithm Based on Bigtable and MapReduce”, Computer Science, vol. 42, no. 10, **(2015)**, pp. 208-211.
[6] L. Dan, “Research on Improved Apriori Algorithm Based on Compressed Matrix”, Computer Science, vol. 40, no. 12, **(2015)**, pp. 75-80.

Authors

Nan Wang, (1981.07), male, master, research direction: National Traditional Sports Science.