An Improved Distance Metric Clustering Algorithm for Association Rules

LIANG Futai¹,²*, LI Hongquan¹, ZHANG Weifeng³ and ZHANG Chenhao¹

¹ Air Force Early Warning Academy, Wuhan, China
² 31121 PLA Troops, Nanjing, China
³Corresponding author:417516911@qq.com

Abstract. By mining association rules in large data, we can reveal useful information contained in the data and find out the relationship between things or the law of motion. However, because of the huge transaction data, the association rules obtained by mining are complex and massive. It is difficult to find useful association relations, especially when the requirement is uncertain. To solve this problem, this paper first uses Apriori algorithm to mine association rules from a data set, then defines similarity measure between association rules, and applies DBSCAN clustering algorithm to association rules analysis. The analysis results show that this method is effective in association rules analysis.

1. Introduction

Association rule mining, also known as association analysis, is to find frequent patterns, associations, or causal structures that exist between project collections or object collections in relational data, transaction data, or other information carriers [1]. Association analysis is one of the methods of big data mining [2].

In the process of mining and analyzing Association rules, the analysis and interpretation of association rules is a key link. Because the association rule set extracted from a large amount of data is very large, it is very difficult to analyze and interpret. In order to solve this problem, one method is to cluster transaction data sets, such as such as shown in literature [3, 4]. This method clusters the items with high correlation according to a given threshold, divides transaction data sets into several parts, and then mining association rules in each transaction data set. This method reduces the size of transaction data sets, but from in the perspective of large data, it affects the integrity of the data, which is not conducive to the fullness of useful information. Another method is to analysis association rule sets by clustering algorithm, such as shown in literature [5, 6], literature [6] defined the distance between rules, and cluster association rule sets through K-means method. Literature [5] defined the similarity measure between different association rules. Density-based DBSCAN clustering algorithm is used to cluster Association rules, which greatly reduces the number of output rules. However, the support and confidence factors are not taken into account in the definition of distance and similarity measure, which affects the clustering effect.

In view of the above problems, an improved distance metric clustering algorithm for association rules is proposed in this paper to improve the clustering effect.
2. Basic Concept

Association analysis can mine and analyze association rules such as “the occurrence of some events causes other events to occur” from a large number of historical data [7]. For example, by investigating what customers buy in the mall, it is found that 30% of the customers will buy sheets and pillowcases at the same time, and 80% of the people who buy the sheets buy pillowcases, which hides a connection: \{sheets\} \Rightarrow \{pillowcases\}, that is to say, a large number of customers will buy pillowcases while purchasing sheets, so for the mall, sheets and pillowcases can be placed in the same shopping area of the same mall, which will convenient customers to shop.

Based on the experimental data listed in Table 1, the related concepts of association analysis are explained as follows:

Table 1. Experimental data of association analysis.

| Tid | Sheets | pillowcase | broom | bread | meat | noodles | ... | cabbage |
|-----|--------|------------|-------|-------|------|---------|-----|---------|
| 1   | 1      | 1          | 1     | 0     | 0    | 0       | 0   | 0       |
| 2   | 0      | 0          | 0     | 1     | 0    | 1       | 0   | 0       |
| 3   | 1      | 1          | 0     | 0     | 1    | 0       | 1   | 0       |
| 4   | 0      | 0          | 1     | 1     | 0    | 1       | 0   | 0       |
| ... | ...    | ...        | ...   | ...   | ...  | ...     | ... | ...     |
| m   | 0      | 0          | 0     | 1     | 1    | 0       | 1   | 1       |

(1) Item: A commodity in a transaction in the experimental data is one item. For example, sheets, pillowcases, hamburgers, etc.

(2) Transaction: A transaction represents the sum of all commodities involved in an exchange. For example, the first transaction record includes sheets, pillowcases and brooms. A transaction T is an itemset, and each transaction has a unique identifier Tid. All transactions together constitute the transaction data set of association rule mining.

(3) Itemsets: A set containing one or more items is called an itemset, such as \{sheet, pillowcase, broom\}.

(4) K-Itemsets: An itemset containing K items is called a K-Itemset. For example, \{pillowcase, bread\} is called a 2-itemset.

(5) Support count: also known as the frequency and count of item set, which is the number of transactions containing the itemset. For example, \{spaghetti, bread\} appears in transactions with Tid 2 and Tid 4, so its support count is 2.

(6) Support: The support of association rules \(A \Rightarrow B\) refers to the probability that events A and B occur simultaneously, which is equal to the ratio of support count to total transactions [8]. For example, the total number of transactions listed in Table 1 is 5, and the support count of itemset \{noodles, bread\} is 2, so its support is 40%, indicating that noodles and bread appear in 40% of transactions.

(7) Frequent itemsets: The itemset whose support is greater than or equal to a threshold are called frequent item set. For example, when the threshold is set to 30%, the support of the itemset \{noodles, bread\} is 40%, so it is a frequent itemset.

(8) Right (RHS) and Left (LHS): For rules \{bread\} \Rightarrow \{noodles\}, \{bread\} is called left (also called the former), and \{noodles\} is called the right part (also called the later).

(9) Confidence: Confidence of association rules \(A \Rightarrow B\) is the probability of occurrence of event B on the basis of occurrence of event A. The formula is as follows:

\[
\text{confidence} = \frac{P(B|A)}{P(A)} = \frac{P(AB)}{P(A)}
\]  

For the rule \{noodles\} \Rightarrow \{bread\}, the support count of itemset \{noodles, bread\} divided by the support count of \{noodles\} is the confidence of the rule, which is 100%, indicating that the person who bought the noodles 100% will buy bread.
(10) Strong Association rules: is an association rules that is greater than or equal to the minimum support and the minimum confidence threshold.

The goal of association analysis is to find strong association rules.

3. Association Rule Mining

Apriori algorithm and FP-growth algorithm are commonly used in association rules mining. Apriori algorithm is a classical data mining algorithm for frequent item set and association rules [9]. The purpose of association rules is to find out the relationship between items in a data set, also known as shopping basket analysis, because "shopping basket analysis" expresses a scenario suitable for the algorithm. The core of the algorithm is the iteration based on the idea of two-stage frequent itemsets:

The first step is to find all frequent itemsets.

The steps of the process are as follows:

Scan all transactions to generate candidate 1-itemset C1. Then, according to the minimum support threshold, the item satisfying is selected from C1 (as shown in Table 2), that is, the frequent itemset L1.

| 1-itemset | Support count | Support(%) | frequent itemset L1 | Support(%) |
|-----------|---------------|------------|---------------------|------------|
| {bread}   | 3             | 60         | {bread}             | 60         |
| {sheets}  | 2             | 40         | {sheets}            | 40         |
| {broom}   | 2             | 40         | {broom}             | 40         |
| {meat}    | 2             | 40         | {meat}              | 40         |
| {noodles} | 2             | 40         | {noodles}           | 40         |
| {pillowcase} | 2       | 40         | {pillowcase}        | 40         |
| {cabbage} | 2             | 40         | {cabbage}           | 40         |

By pruning the set generated by L1's own connection, candidate 2-itemset C2 are generated. By scanning all transactions, and satisfying items are selected from C2 according to the minimum support threshold(as shown in Table 3), that is, frequent sets L2.

| 2-itemset | Support count | Support(%) | frequent itemset L2 | Support(%) |
|-----------|---------------|------------|---------------------|------------|
| {bread, broom} | 1             | 20         | {bread, broom}      | 20         |
| {bread, cabbage} | 1             | 20         | {bread, cabbage}    | 20         |
| {bread, meat} | 1             | 20         | {bread, meat}       | 20         |
| {bread, noodles} | 2             | 40         | {bread, noodles}    | 40         |
| {bread, pillowcase} | 0           | 0          |                     |            |
| {broom, cabbage} | 0             | 0          |                     |            |
| {broom, meat} | 0             | 0          |                     |            |
| {cabbage, meat} | 2             | 40         | {cabbage, meat}     | 40         |
| {cabbage, noodles} | 0           | 0          |                     |            |
| {cabbage, pillowcase} | 1          | 20         | {cabbage, pillowcase} | 20         |
| {meat, noodles} | 0             | 0          |                     |            |
| {meat, pillowcase} | 1             | 20         | {meat, pillowcase}  | 20         |
| {noodles, pillowcase} | 0            | 0          |                     |            |
By pruning the set generated by L2's own connection, candidate 3-itemset C3 are generated. By scanning all transactions, and the satisfied items are selected from C3 according to the minimum support threshold (as shown in Table 4), which means frequent 3-itemset L3.

Table 4. Third iteration result.

| 3-itemset C3                      | Support count | Support(%) | frequent itemset L3            | Support (%) |
|-----------------------------------|---------------|------------|-------------------------------|-------------|
| {sheets, broom, cabbage}         | 0             | 0          | {sheets, broom, pillowcase}   | 20          |
| {sheets, broom, meat}            | 0             | 0          | {sheets, cabbage, meat}      | 20          |
| {sheets, cabbage, pillowcase}    | 1             | 20         | {sheets, cabbage, pillowcase} | 20          |
| {sheets, meat, pillowcase}       | 1             | 20         | {sheets, meat, pillowcase}    | 20          |
| {bread, broom, cabbage}          | 0             | 0          | {bread, broom, noodles}      | 20          |
| {bread, broom, meat}             | 0             | 0          | {bread, cabbage, meat}       | 20          |
| {bread, cabbage, noodles}        | 1             | 20         | {bread, meat, noodles}       | 20          |
| {bread, cabbage, meat}           | 1             | 20         | {bread, meat, pillowcase}    | 20          |
| {broom, noodles, pillowcase}     | 0             | 0          | {cabbage, meat, pillowcase}  | 20          |
| {cabbage, meat, pillowcase}      | 1             | 20         | {cabbage, meat, pillowcase}  | 20          |

Through an iterative operation, frequent k-itemsets Lk are obtained from L(k-1).

Now find frequent itemsets from the transactions listed in Table 1. Set the minimum support threshold to 0.2, and finally get the frequent 4-itemset is: {cabbage, meat, pillowcase, Sheets}, and its support is 0.2.

The second step is to generate strong association rules from frequent itemsets.

When frequent itemsets are found, strong association rules are generated directly from them:

All non-empty subsets (frequent itemsets) are obtained from each frequent itemset Ln. For each non-empty subset P of Ln, if

$$\frac{\text{support}(L_n)}{\text{support}(P)} \geq \text{min\_conf}$$

(2)

Then strong association rules are obtained, where min\_conf is the minimum confidence threshold.

For the frequent itemsets obtained by the first iteration {cabbage, meat, pillowcase, Sheets}, made the minimum confidence threshold is 0.8, and strong association rules whose RHS is limited to a single item can be obtained, as shown in Table 5:

Table 5. Strong association rules.

| strong association rules       | Support (%) | Confidence (%) |
|--------------------------------|-------------|----------------|
| {cabbage, meat, pillowcase} ⇒ {Sheets} | 20          | 100            |
| {cabbage, meat, Sheets} ⇒ {pillowcase}   | 20          | 100            |
| {cabbage, pillowcase, Sheets} ⇒ {meat}   | 20          | 100            |
| {meat, pillowcase, Sheets} ⇒ {cabbage}   | 20          | 100            |
4. Association Rule Analysis

The association rules which satisfy the minimum support and confidence are mined to form the association rules set [10].

For Association rules \( \text{LHS} \Rightarrow \text{RHS} \), LHS and RHS can be itemsets containing multiple items. Generally speaking, association rule mining is demand-driven. Once the user’s needs is determined, RHS is usually limited to a single item, while the number of items contained in LHS is uncertain. For example, in order to determine the storage location of meat cabinet, it is necessary to find out the rules and related factors of customers buying meat. At this time, RHS is \{meat\}, and items of LHS can be one or many. Considering that the association rules with high confidence are often concerned by users, the rules with high confidence are chosen to facilitate users to interpret association rules and analyse the rules of activities.

In most cases, mining and analysis needs are not clear or general. In this case, clustering algorithm can be used to mine and analyse association rules set [11]. Clustering similar rules in the same cluster helps to reduce the size of rules and facilitate users to analyse and understand the meaning of association rules [6].

In this paper, the similarity measure of association rules is defined according to the relationship between the former and the latter. Combining with density-based DBSCAN clustering algorithm, the association rules are clustered, and the similarity rules in clustering cluster are selected to output. Based on the density distribution of the sample, the DBSCAN clustering algorithm continuously checks the connectivity between samples, and finally obtains the clustering result. It can cluster clusters of arbitrary shapes, but the user must select the appropriate distance metric [12].

For a given rule \( R_i: X_i \Rightarrow Y_i \) and \( R_j: X_j \Rightarrow Y_j \), the distance measure between the two can be defined as:
\[
dist_{\alpha\beta}(R_i, R_j) = \alpha \cdot \text{dist}_s(X_i \cup Y_i, X_j \cup Y_j) + \beta \cdot \text{dist}_s(X_i, X_j) + \text{dist}_s(Y_i, Y_j)
\]
where, \( \alpha = |\text{support}_{R_i} - \text{support}_{R_j}| + 0.1 \)
\( \beta = |\text{confidence}_{R_i} - \text{confidence}_{R_j}| + 0.1 \)

The Jaccard Distance measures the distinction between two sets by the proportion of different elements in the two sets to all elements:
\[
J_s(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}
\]

\( \text{dist}_s(X_i, X_j) \) represents the Jaccard distance of the former of two association rules, \( \text{dist}_s(Y_i, Y_j) \) represents the Jaccard Distance of the latter of two association rules, support_{R_i} and support_{R_j} are the supports of the two rules, confidence_{R_i} and confidence_{R_j} are the confidences of the two rules. In order to reflect the difference degree of support between rules, the distance of frequent itemsets corresponding to rules is multiplied by coefficient \( \alpha \). In order to reflect the difference degree of confidence between rules, the distance between the former and the latter of rules is multiplied by coefficient \( \beta \). Adding 0.1 to \( \alpha \) and \( \beta \) is to avoid their values becoming zero. Thus, the distance between the two rules is compared at the same level of support and confidence. By clustering experimental data, the central rules of three kinds of clusters in association rules are obtained (as shown in Table 6).

| Category | The former | The later | support | Confidences |
|----------|------------|-----------|---------|-------------|
| 1        | {sheets, broom, pillowcase} | {meat}    | 0.3     | 0.8         |
| 2        | {cabbage, meat}            | {sheets, pillowcase} | 0.2     | 0.9         |
| 3        | {bread, cabbage, meat}    | {noodles}  | 0.3     | 1.0         |
By observing the three association rules in cluster category 2 (as shown in Table 7), we can find that there are strong similarities among them.

Table 7. Association rules with high confidence.

| Tid | The former                  | The later | support | confidences |
|-----|-----------------------------|-----------|---------|-------------|
| R1  | {sheets, broom, pillowcase} | {meat}    | 0.3     | 0.8         |
| R2  | {sheets, cabbage, pillowcase} | {meat}    | 0.2     | 0.9         |
| R3  | {sheets, bread, pillowcase} | {meat}    | 0.2     | 0.7         |

5. Conclusion

In this paper, Apriori algorithm is used to mine association rules from a shopping mall transaction data set. Aiming at the problem that the association rules sets are so big that it’s difficult to analyse and interpret, a definition system of rule distance is established. A large number of association rules are clustered by density-based DBSCAN clustering algorithm, and different similar rule clusters are formed to make the classification of association rules. Analytical comprehension becomes easy and effective. The results show the effectiveness of this method.

References:

[1] Cui Y, Bao Z Q 2016 Survey of Association Rule Mining *Application of Research Computers* 33 pp 330-334.
[2] Xu Y, Hu W M and Yang Z F 2016 A Multivariate Partial Least Squares Approach to Joint Association Analysis for Multiple Correlated Traits *The Crop Journal* 4 pp 21-29.
[3] Mao Y X, Chen T B and Shi Bole 2011 Efficient Method for Mining Multiple-Level and Generalized Association Rules *Journal of Software* 22 pp 2965-2980.
[4] He Q 2017 Research and Improvement of Clustering-based Multi-level Association Rules Mining Algorithms *Shanghai Normal University Master of Science*.
[5] Zhang L L, Zhou Q L and Tang G G 2015 Research on Algorithm of Post-processing Association Rules Based on Clustering and Domain Knowledge *Chinese Journal of Management Science* 23 pp 154-161.
[6] Wang Zh, Xun Y L, Zhang J F 2016 An Association Rule Clustering Algorithm Based on K-means *Journal of Taiyuan University of Science and Technology* 37 pp 429-437.
[7] Song Ch X 2016 Research of Association Rule Algorithm based on Data Mining[A]. IEEE Beijing Section.Proceedings of 2016 IEEE International Conference on Big Data Analysis (ICBDA) IEEE Beijing Section pp 4.
[8] Zhao Ch Y 2016 Efficient Association Rule Mining Algorithm Based on User Behavior for Cloud Security Auditing IEEE Computer Society.Proceedings of 2016 IEEE International Conference of Online Analysis and Computing Science(ICOACS 2016) IEEE Computer Society pp 5.
[9] Cheng M C 2016 Research on Audit Log Association Rule Mining Based on Improved Apriori Algorithm IEEE Beijing Section.Proceedings of 2016 IEEE International Conference on Big Data Analysis (ICBDA)[C]. IEEE Beijing Section pp 7.
[10] Liu J Sh, Li J W and Liu L 2019 Ships’ Association Rule Minging Based on Improved Interestingness *Ship Electronic Engineering* 39 pp 78-82.
[11] Yan M Q,Guo Zh M and Ren Zh H 2017 Spatio-temporal analysis of bus pickpocketing using association rules based on clustering *Journal of East China Normal University (Natural Science)* pp 145-152.
[12] Yang Y L 2017 The PGST-DBSCAN Algorithm for Mining Clusters from Massive Spatial-temporal Data Wuhan Zhicheng Times Cultural Development Co., Ltd.Proceedings of the 2nd International Conference on Artificial Intelligence and Engineering Applications(AIEA 2017)[C].Wuhan Zhicheng Times Cultural Development Co., Ltd pp 12.