International Journal of Database Theory and Application
Vol.8, No.2 (2015), pp. 245-258
http://dx.doi.org/10.14257/ijdta.2015.8.2.23

The Research on Measure Method of Association Rules Mining

Gao Yongmei and Bao Fuguang

Hangzhou Vocational & Technical College, Hangzhou 310018, China;
Zhejiang Gongs hang University, Hangzhou 310018, China
Gaoyongmeixdx@163com, 2. 506617078@qq.com

Abstract

Data mining receives much attention from artificial intelligence and databases, and the association rule is one of the most important research fields of data mining. In this paper, the advantages and disadvantages of the specific indicators of objective measure, subjective measure, and association rule based on statistical perspective are discussed. Some indicators of statistical perspective are adopted to measure the association rules, which can effectively solve the problems of association rules. Next, a further verification of the advantage and disadvantages of the indicators is made by the combination of the theory and application, a new measure frame is put forward as well. Then, the dynamic association rules are analyzed through making a comparative analysis in the following four aspects: the traditional association analysis without the life cycle, the association rules with the life cycle, the weighted dynamic association rules and the weighted dynamic association rules weighted by the consumption amount, showing the influence of timeliness on association rules analysis, and thus effectively mining some rules with low support in global period but high support in a certain period.

Keywords: *Association mining; Measure; Dynamics; Life circle*

1. Introduction

Currently, the world is in the era of data explosion. With the continuous geometric growth of the data, researchers pay increasing attention to the data mining technology. Data mining receives widespread attention from artificial intelligence and databases. The association rule is an important branch of data mining and it is also an important research orientation in the study of knowledge discovery field in recent years. The research object of the association rule is transaction database and the aim of the research is to discover the relations among transaction items in the transaction database.

The concept of the association rule was first put forward by Agrawal, *etc.* in 1993[1] [2], and it was used to handle transactional database and then was spread to relational database. The main purpose was to research the pattern among the commodities purchased by the customers in the supermarket, and to discover the commodities usually purchased simultaneously by the customers, and then have a reasonable layout, which is convenient for customers to select commodities, and it was called the Shopping Basket Analysis. They put forward the classic Apriori algorithm and then numerous follow-up researchers put forward many improvements for algorithm to increase the mining efficiency and extend the application of the association rule. Association rule algorithm can generate a lot of rules, but due to the limited resources, only a part of the rules may be adopted by policy-makers. In order to avoid illusive association rule, various new thresholds were introduced to strengthen the evaluation for the association rules. Among these, the interestingness is a relative eye-catching point.

The measure research of the interestingness mainly includes the objective measure and the subjective measure [3]. The objective interestingness mainly considers the significant statistical characteristics of the objective data. It includes not only the classic

ISSN: 2005-4270 IJDTA

support, confidence, lift and *etc.*, but also relatively new matching, trusts, improvement, influence, *etc.*. There are still some limitations [3] in both classic theory and new research of the objective interestingness. The subjective interestingness mainly involves the knowledge field, hobbies and other personality characteristics of the main body (users). Compared with the research of objective interestingness, the research of subjective interestingness is relatively rare and immature. Because, the interestingness evaluation of association rule has the significance for the practical application of association rule mining technology, so it is necessary to study and improve it.

2. The Interestingness Measure of Association Rule

A transaction $\boldsymbol{T}$ is usually made up of a transaction mark ($\boldsymbol{TID}$) and item sets, item set $\boldsymbol{X}$ for short. The $\boldsymbol{TID}$ can determine one transaction alone. Let$\boldsymbol{I} = \{\boldsymbol{I_1}, \boldsymbol{I_2}, \dots\dots \boldsymbol{I_k}\}$, $\boldsymbol{I}$ include all of the $\boldsymbol{k}$ items, and the transaction $\boldsymbol{T} \subseteq \boldsymbol{I}$, itemset$\boldsymbol{X} \subseteq \boldsymbol{I}$.

For the convenience of explanation, the formalized description for the association rule is firstly assumed as: $\boldsymbol{A} \rightarrow \boldsymbol{B}$. Among this, $\boldsymbol{A} = \{\boldsymbol{A_1}, \boldsymbol{A_2}, \dots\dots, \boldsymbol{A_j}\} \subset \boldsymbol{I}$, $\boldsymbol{B} = \{\boldsymbol{B_1}, \boldsymbol{B_2}, \dots\dots, \boldsymbol{B_k}\} \subset \boldsymbol{I}$, and $\boldsymbol{A} \cap \boldsymbol{B} = \emptyset$. The rules must satisfy a certain support threshold $\boldsymbol{s}$ and confidence threshold $\boldsymbol{c}$.

Generally, there are two evaluation standards to evaluate whether an association rule is interesting or not: the objective measure and the subjective measure. The method of the objective measure can obtain a quantitative value by the algorithm, and it is relatively visual and easy to operate. However, the rule after the evaluation of the objective measure may be not the mode users interested in, therefore, the subjective measure is required. In order to ensure that the final mining rule can arouse the interests of the users or the experts in the field, they should be involved in the process and make use of their knowledge to pruning the rule.

2.1 The Indicators of Objective Measure

2.1.1. Support and Confidence: Support and Confidence are two common indicators of the objective measure to evaluate the association rule; the former measures the usefulness of the rules while the latter reflects the effectiveness of the rules. Support [4] refers to the frequency that the concurrence of data domain $\boldsymbol{A}$ and $\boldsymbol{B}$ involved by the association rule occupies in all of item sets, during the researching data item sets. The accuracy will be higher only when the researching association rule frequently appears in item sets. Only when the support of the concurrence of $\boldsymbol{A}$ and $\boldsymbol{B}$ is greater than or equal to the designated minimum support threshold, $\boldsymbol{A}$ and $\boldsymbol{B}$ will be confirmed to be the frequent item sets. The Support can be expressed as:

$$\boldsymbol{s(A \rightarrow B) = P(AB) = N(AB)/|D|}$$

Among this, $\boldsymbol{N(AB)}$ stands for the number of records of the concurrence of $\boldsymbol{A}$ and $\boldsymbol{B}$, and $|\boldsymbol{D}|$ refers to all the number of records of the data sets.

Confidence [4] is the statistical probability of the occurrence of the consequent after the occurrence of the antecedent among the transactional data sets. Confidence is used to measure the reliability of the rules. The formula is as follows:

$$\boldsymbol{c(A \rightarrow B) = P(B|A) = P(AB)/P(A)}$$

The mining of the association rules can be divided into two steps: (1) find out the entire maximal frequent item sets meeting the condition; (2) generate the association rules by the frequent item sets. During the generation of the association rules, the rules that cannot meet the support-confidence threshold are filtered out, and the strong association rules are finally generated. This is the generating process of the association rules based on the Support-Confidence evaluation system.

2.1.2. Lift: Due to the deficiency of support-confidence frame, some scholars carried out the correlation analysis for the mined association rules, namely Lift. Lift [5] is also called Correlation or Interestingness in some references. Lift refers to the ratio of the rule's confidence to the consequent's occurrence probability of the rule, reflecting the positive and negative correlation between the antecedent and consequent. The research of the correlation can partly remove some rules that have little correlation among the rules mined based on the support-confidence frame. The correlation reflects the probability ratio of ***B*** occurrence under the condition of ***A*** to the ***B*** occurrence without the condition of ***A***, and reflects the relation between ***A*** and ***B***. Lift does not possess the downward closure or the problem of rare itemset.

$$\boldsymbol{Lift(A \rightarrow B) = c(A \rightarrow B)/P(B) = P(AB)/P(A)P(B)}$$

However, the lift still has some deficiency: (1) the value direction of lift can reflect the influential direction of ***A*** to ***B***, but the quantity of the value cannot effectively indicate the influential degree of ***A*** to ***B***. (2) there is no standard for the value of the lift; whatever the value is, the probability of ***B*** occurrence under the condition of ***A*** is obviously greater than that without the condition of ***A***; (3) it is easy to flitter out the rules which is made up of high frequent item sets. In this case, some scholars made some amendments for the computation of the lift, and their idea is to introduce a negative item in order to strengthen the expression means of knowledge and to improve the evaluation of the existing association rules.

$$\boldsymbol{Lift(A|B) = \frac{c(A|B) - P(B)}{\max\{c(A|B) - P(B)\}}}$$

The value range of the lift is $[-1,+1]$, a closer value to 1 indicates that the rule has a greater value.

2.1.3. Improvement: Because there is some deficiency for the method of the traditional interestingness measure, so a new measure method is put forward, it is called improvement for the moment [3, 12]. The improvement refers to the difference of probability between ***B*** occurrence under the condition of ***A*** and the ***B*** occurrence without the condition of ***A***.

$$\boldsymbol{Improve(A \rightarrow B) = [P(B|A) - P(B)]}$$

2.1.4. Validity: The new measure method of the association rule is called validity [6]. The validity is defined as the probability of the co-occurrence of ***A*** and ***B*** subtracts the probability of ***B*** occurrence without the condition of ***A*** in the database ***D.*** Because the value range between ***P(AB)*** and $\boldsymbol{P(\bar{A}B)}$ is [0,1], it is obvious that the range of the validity is [1,1].

$$\boldsymbol{Validity(A \rightarrow B) = P(AB) - P(\bar{A}B)}$$

2.1.5. Influence: The interestingness measure standard based on ***T*** verification is put forward, namely influence [3]. The statistics ***T*** verification method is used to analyze the difference between the association confidence ***P(B|A)*** and the expectation confidence ***P(B)***. If the difference is large, it indicates that the occurrence of ***A*** has a relatively large influence on the occurrence of ***B***. The rule (A → B) is interesting, and the formula is shown as:

$$\boldsymbol{Influence(A \rightarrow B) = [P(B|A) - P(B)]/\sigma}$$

$$\boldsymbol{\sigma} = \sqrt{\frac{\boldsymbol{P(B)(1-P(B))}}{\boldsymbol{n}}}$$

2.2. The Indicators of Subjective Measure

The subjective evaluation indicator mainly embodies the subjective factors, such as user participation and the integration in the field of knowledge, *etc*. This evaluation of this level is from the perspective of rules, regardless of the data in the database.

2.2.1. Novelty: Novelty is a relative concept to the primal knowledge, and its extent reflects on the difference in each item between the discovered rules and the rules based on the knowledge base and the difference is respectively reflected on the difference extent in each item of the antecedent and consequent [7]. Assuming the set made up of the discovered rules is $\boldsymbol{E}$, and the rules set in the basic knowledge base is $\boldsymbol{K}$. The number of rules in $\boldsymbol{E}$ is $|\boldsymbol{E}|$, and the number of rules in $\boldsymbol{K}$ is $|\boldsymbol{K}|$. Assuming $\boldsymbol{W_i}$ is the novelty of the rule $\boldsymbol{E_i}$ in $\boldsymbol{E}$ relative to $\boldsymbol{K}$. $\boldsymbol{W_{(i,j)}}$ is the novelty between the rule $\boldsymbol{E_i}$ and the rule $\boldsymbol{K_j}$ in the basic knowledge base, namely the difference degree. $\boldsymbol{W_{(i,j)}}$ includes two parts: the novelty $\boldsymbol{L_{(i,j)}}$ of the antecedent and the novelty $\boldsymbol{Z_{(i,j)}}$ of the consequent. Assuming $\boldsymbol{J}$ is the set of all the antecedents among the rule $\boldsymbol{K_j}$ in the basic knowledge base, and the $\boldsymbol{I}$ is the set of all the antecedents among the rule $\boldsymbol{E_i}$ in the $\boldsymbol{E}$. As for any item $\boldsymbol{I_k}$ in the $\boldsymbol{I}$, $\boldsymbol{V_{(i,j)k}}$ is the difference degree between this item and the rule$\boldsymbol{K_j}$, we conclude:

$$\boldsymbol{V_{(i,j)k}} = \begin{cases} \boldsymbol{2,\ I_k \notin J} \\ \boldsymbol{1 + neg_k, I_k \in J} \end{cases}$$

$\boldsymbol{neg_k}$is the difference degree of the values between the $\boldsymbol{K}$th item in $\boldsymbol{I}$ and the same item in $\boldsymbol{J}$. The novelty of the antecedent is equal to the accumulation for the difference degree on each item of the antecedent, namely:

$$\boldsymbol{L(i,j)} = \sum_{\boldsymbol{k=1}}^{|\boldsymbol{I}|} \boldsymbol{V_{(i,j)k}}$$

After the simplification of the rule, the items number of the consequent of all the rules in the basic knowledge base is 1, so is the items number of the consequent of the rules obtained through data mining algorithm. Therefore, the newly discovered rule $\boldsymbol{E_i}$ has only two possible relations with any rule $\boldsymbol{K_j}$ in the knowledge base: (1) the consequent of the two rules belongs to the same attribute. At this time, the difference degree (*neg*) of the corresponding value between the two rules should be calculated first, namely $\boldsymbol{Z(i,j) = 1 + neg}$. (2) the consequent of the two rules does not belong to the same linguistic variable. At this time, make $\boldsymbol{Z_{(i,j)} = 2}$ and then calculate the sum:

$$\boldsymbol{W_i} = \frac{\sum_{\boldsymbol{j=1}}^{|\boldsymbol{k}|} \boldsymbol{w(i,j)}}{|\boldsymbol{k}|}$$

Finally, the novelty of the new rules can be determined through the calculated novelty$\boldsymbol{W_i}$. The rules with higher novelty should be kept while the rules with lower novelty should be deleted, and put the leaving rules into the knowledge base.

2.2.2. Availability: The aim that clients analyze data by the data mining tools is to utilize the result of the data mining to support the decision. If the clients can improve the workflow and enhance efficiency according to a certain mode of data mining, then the mode is interesting. If the clients can utilize the obtained knowledge to take some actions, and thus improve the work efficiency or bringing some economic profit, it is thought to be practical. Availability is the function of the costs needed by the researched system transformed into the associated space (interested state) $\boldsymbol{M_i}$ from the current space (primal state) $\boldsymbol{M}_0$. Namely:

$$\boldsymbol{S = \frac{1}{f(M_0 - M_i)}}$$

The more it costs to transform, the smaller the availability of the mode has. When the system cannot be transformed into the associated space from the current space, it is $\boldsymbol{s \to 0}$.

In addition, many references also mention the simplicity, trust, comprehensive evaluation, *etc.*. And the comprehensive evaluation indicator is the measure indicator including all kinds of indicators, which is obtained through the set weighted average of the objective and subjective measure indicators.

2.3. The Indicators of Association Rule Based on Statistical Perspective [8]

2.3.1. Contingency Table: Contingency table is a frequency table listed by the classification of the observation data according to two or more attributes (qualitative variable). The basic analysis problem of the contingency table is to ascertain whether there are associations among all the observed attributes, namely whether independent. The statistics used by the contingency table to examine whether there is an association among all the attributes is chi-squared statistic:

$$\boldsymbol{x^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij} - e_{ij})^2}{e_{ij}}}$$

The analysis steps as follows: firstly, put forward an hypotheses, and the primal hypotheses is that the line variation and row variation are independent, and the selected hypotheses is that the line variation and row variation are not independent; secondly, calculate chi-squared statistic; lastly, make decisions, and find out the critical value χ_α according to the significance level α and free degree (r-1)(c-1). If $\chi 2 \geq \chi_\alpha$, refuse the primal hypothesesH_0, and if $\chi 2 < \chi\alpha 2$, not refuse.

2.3.2. Canonical Correlation Analysis: Canonical correlation analysis is a multivariate statistical method to study the association between two groups of variables, and it can reveal the internal association between two groups of variables. The aim of canonical correlation analysis is to identify and quantity the association between two groups of variables and to make the analysis of the correlation between two groups of variables transform into the analysis of the correlation between one group of linear combination and the other group of linear combination.

Currently, canonical correlation analysis has been applied into psychology, marketing and other fields, such as the association study between the personal character and vocational interest, as well as the association between sales promotion activities and consumer responses, *etc.*.

The thought of canonical correlation analysis: firstly, find out the first pair of linear combination from each group of variables and make it have the maximum correlation.

$$\boldsymbol{\begin{cases} u_1 = a_{11}x_1 + a_{21}x_2 + \cdots + a_{p1}x_p \\ v_1 = b_{11}y_1 + b_{21}y_2 + \cdots + b_{q1}y_q \end{cases}}$$

And then find out the second pair of linear combination from each group of variables and make it have no correlation with the first pair of linear combination, and the second one has the second largest correlation.

$$\boldsymbol{\begin{cases} u_2 = a_{12}x_1 + a_{22}x_2 + \cdots + a_{p2}x_p \\ v_2 = b_{12}y_1 + b_{22}y_2 + \cdots + b_{q2}y_q \end{cases}}$$

Here, (***u₂*** and ***v₂***) and (***u₁*** and ***v₁***) are mutual independence, but u_2 and v_2 are correlative. Continue like this till the ***r***th step and the correlation between two groups of variables is completely extracted. When $r \leq min(p,q)$, ***r*** groups of variables will be obtained.

2.3.3. Obtaining the Comprehensive Indicator through the Principal Component Analysis: The first principal component is obtained through the principal component analysis, namely it is regarded as a new comprehensive indicator. The computational formula of the principal component is shown as follows:

$$F_1 = a_{11}x_1 + a_{21}x_2 + \cdots + a_{p1}x_p$$

Through the principal component analysis, we can regard the indicator associated with the association rule as the variable X, and not consider the redundancy among the variables.

2.3.4. Obtaining the Comprehensive Indicator through the Geometric Method: $RI = s^{w_1} * c^{w_2} * lift^{w_3} * wi^{w_4} * s1^{w_5}$, in this formula, ***s*** is support, ***c*** is confidence, ***lift*** is lift , ***wi*** is novelty and ***s1*** is availability. There are several advantages using the comprehensive indicator ***RI***: firstly, support, confidence and lift respectively represent practicality, credibility, correlation, which can comprehensively reflect all aspects of association with little redundancy; secondly, this comprehensive indicator can not only reflect the objectiveness, but also users' objective interestingness; thirdly, through the weight, this comprehensive indicator can embody the importance of each indicator and it is not affected by the inconsistent dimension.

2.4. The Comparison for the Objective Measure Indicator

This paper concisely compares the differences of each objective measure indicator, and the sample data is the customer transaction data, which is made into a fact table and variables' names are the specific items, including ten items and ten transactions, and the variable value is 1 or 0, and 1 stands for purchase and 0 stands for no purchase, as shown in Table 1.

The association rule may be effective only when a certain support and confidence are satisfied. The support threshold is 40% and the confidence threshold is 50% in this paper. Use ***Apriori*** algorithm to obtain 8 rules in Table 2 and respectively calculate the indicator value of the objective measure.

Table 1. The Transaction Data of the Customers

TID	A	B	C	D	E	F	G	H	I	J
1	0	1	1	1	0	0	0	0	0	0
2	0	1	0	0	1	0	0	1	0	1
3	0	1	0	1	0	1	1	0	0	1
4	0	1	1	1	0	1	0	0	1	0
5	1	1	0	0	1	1	0	0	1	0
6	1	1	0	0	0	0	0	0	0	0
7	1	0	1	1	0	1	1	1	0	1
8	1	1	0	0	0	1	0	1	0	1
9	0	0	1	1	1	1	1	0	0	0
10	0	1	0	0	1	1	1	0	0	0

Table 2 shows that we cannot tell which the interesting rule is from the support indicator alone. Although the confidence indicator of the rule ***C→D*** and ***G→F*** are 100%, and are highest, we cannot conclude that they are really interesting rules. Similarly, we cannot select out which the interesting rules are through the comparison of the other indicator alone. Analyzing these indicators simultaneously can determine confidently which will be the really interesting rules.

Table 2. The Association Rule and its Objective Measure indicator

Rules	Support	Confidence	Lift	Improvement	Validity	Influence
B→F	0.5	0.63	0.89	-0.08	0.29	-0.54
F→B	0.5	0.71	0.89	-0.09	0.20	-0.69
C→D	0.4	1.00	2.00	0.50	0.30	3.16
D→C	0.4	0.80	2.00	0.40	0.40	2.58
D→F	0.4	0.80	1.14	0.10	0.10	0.68
F→D	0.4	0.57	1.14	0.07	0.30	0.44
F→G	0.4	0.57	1.43	0.17	0.40	1.11
G→F	0.4	1.00	1.43	0.30	0.10	2.07

According to the Support-Confidence evaluation system, the minimum confidence indicator is set as 70%, then five strong association rules can be obtained, respectively ***F→B***,***C→D***,***D→C***,***D→F*** and ***G→F***. On this basis, because the improvement indicator of the rule ***F→B*** is negative, it illustrates that the occurrence of the antecedent ***F*** lowers the possibility of the occurrence of the consequent. Therefore this rule can be excluded. According to the lift indicator, the values of the rule ***C→D*** and ***D→C*** are equal; moreover the values are obviously higher than that of others. Although the validity indicator of the rule ***D→C*** is slightly higher than that of the rule ***C→D***, the improvement and influence indicator of ***C→D*** are higher than those of the ***D→C.*** Therefore, it is thought that ***C→D*** is more interesting than ***D→C***.

Make a comparative analysis on the above objective measure indicator and combine the relevant references, it is though: (1) Support can filter out most of the non-associated or negative association rules;(2) Support-Confidence evaluation system can generate strong association rules, but it cannot distinguish positive association, negative association or non-associated rules. Moreover, it cannot analyze the rare data that is less than the support threshold; (3) The value direction of the lift can reflect the influential direction of ***A*** to ***B***, but its value cannot effectively indicate the influential degree of ***A*** to ***B***, and the rules which are made up of the high frequency item sets are easily filtered out; (4) The validity can reduce part of the redundant rules, but it cannot eliminate the non-associated rules; (5) generally, the obtained association rule is significant when the improvement is higher than a certain minimum, but there is no standard to set this minimum yet.

3. The Evaluation for the Measure Frame of the Traditional Association Rule

Firstly, we make an analysis on the Support-Confidence of the customers' shopping basket in different time periods. We not only analyze the association rule of the consumers' shopping in different time periods, but also find out the changes for the association rule of the consumers' shopping with the time changing. Therefore, we can speculate the consuming habits and consuming preferences of a certain region over a certain time. Then, this paper makes an analysis for the advantages and disadvantages of the traditional association rules' measure frame based on Support-Confidence. Organize the data into the form required by Clementine12.0 and operate the Clementine12.0, the result is shown in Table 3.

Table 3. The Support and Confidence of Each Rule in Different Time

Time periods	*1*	*2*	*3*	*4*	*5*	*6*
AF Support	50	40				40
AF Confidence	60	50				50
CE Support		40		30	40	40
CE Confidence		75		66.7	50	75
EB Support			40	70	70	50
EB Confidence			75	85.7	57.14	60
BE Support		60	60	70	60	60
BE Confidence		50	50	85.7	66.67	50
CA Support	50					
CA Confidence	60					

Note: the blank in this table does not represent that there is no support and confidence. Actually these support and confidence are less than the minimum support (40 %) and confidence (50%) that we define.

The Figure 1 showed that the support and confidence of ***A→F*** at the first, second and six period are more than or equal to the designated minimum threshold, therefore ***A*** and ***F*** had a strong association, so the occurrence of ***A*** had a big influence on ***F***. However, the support and confidence of ***A→F*** at the third, fourth and fifth period are less than the designated minimum threshold, therefore we don't think that there is a strong association among them, so the occurrence of ***A*** have not a big influence on ***F***. This change may be related to unexpected events. For example, recently, bird flu in Hangzhou made people worry about the food safety for chicken and people gradually reduced the purchase for the chicken. Hypothetically, when the residents purchased the vegetables, they would also purchase chicken at the beginning. Green vegetables and chicken had a strong association. As people reduced the consumption for chicken, they began to purchase the substitute for chicken, thus leading to the result that the primal strong association between chicken and vegetables were no longer associated. ***A*** and ***F*** at the sixth period had a strong association, which shows that ***A*** and ***F*** had a long-term stable association, and this association meets the minimum threshold of the support and confidence, and therefore, ***A*** and ***F*** had a strongly practical value. We firmly believed that, when the bird flu (H7N9) disappeared, the residents would re-adjust the impact for the food safety of chicken and they would re-select the combination of chicken and green vegetables, so that the strong association between chicken and green vegetables would be restored.

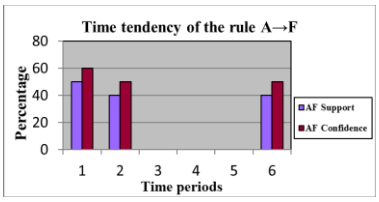

Figure 1. The Support and Confidence of A→F with Time Changing

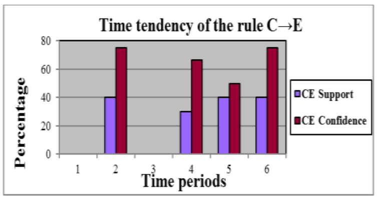

Figure 2. The Support and Confidence of C→E with Time Changing

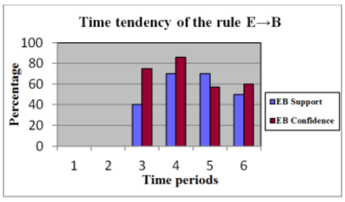

Figure 3. The Support and Confidence of E→B with Time Changing

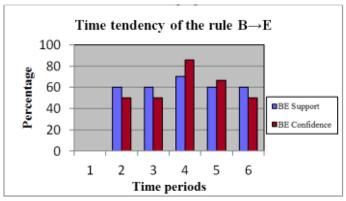

Figure 4. The Support and Confidence of B→E with Time Changing

Figure 2, Figure 3 and Figure 4 show that the support and confidence of ***B→E*** and ***E→B*** are more than and equal to the designated minimum support and confidence threshold with the time changing, and the fluctuation is small. So the rule relation between ***B→E*** and ***E→B*** is relatively stable, this association rule is an important reference for decision-makers. For example, if the owners of fast food restaurants understand the customer's consumption preference, then they can arrange the package series for customers to increase the customer flow.

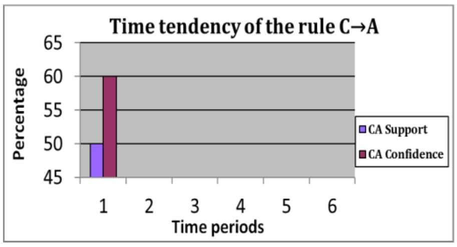

Figure 5. The Support and Confidence of C→A with Time Changing

Figure 5 showed us that ***C→A*** at the first period has a higher support and confidence. However, in the later periods, the support and confidence of ***C→A*** are less than the designated minimum support and confidence. So the rule disappeared in the later periods. It is obvious that this rule has no great practical significance. And it only reflected the association of ***C*** and ***A*** at a certain period. For this case, it is common in real life. For example, the sales of the moon cakes and pomelo rapidly rose, during the Mid-autumn Day, but after August, the sale of the moon cakes is almost zero.

The association rule generated by the evaluation system of the traditional Support - Confidence is the strong association rule, and it can filter out some uninteresting association rules. This evaluation belongs to quantitative evaluation standard, and avoids the influence of subjective factors, therefore the evaluation standard is convincing to some extents. However, the objective evaluation of the association rule is only based on the structure of the data, while the generation of the association rule is completely based on fact data without considering the relations among the rules and the identity degree of the users. In real life, the association rules are significant, only when the users are interested in and the rules are useful. In addition, this evaluation system cannot analyze the rare data item which is less than the minimum support threshold; because it don’t consider the association of the statistics, it can generate a fault which might mislead the users’ judgment.

On this, some scholars put forward a new ternary evaluation system of support-confidence -correlation based on the primal evaluation system of support-confidence. The statistic correlation theory is used as follows:

1. **A and B are positive correlation; ⇔ $\bar{A}$ and B are negative correlation; ⇔ A and $\bar{B}$ are negative correlation ⇔ $\bar{A}$ and $\bar{B}$ are positive correlation.**
2. **A and B are mutual independence ⇔ $\bar{A}$ and B are mutual independence ⇔ A and $\bar{B}$ are mutual independence ⇔ $\bar{A}$ and $\bar{B}$ are mutual independence.**

The rule made up of the two negative correlation events should be utilized, rather than simply deleted.

4. The Analysis of Dynamic Association Rules

The traditional association rule hardly takes the applicability of time into consideration. Currently, the rules are assumed to be valid forever, but not show when becoming valid or invalid. Invalid rules don't illustrate whether it was valid in the past or will be valid in the future. Therefore, the analysis of the dynamic association rule can effectively mine some rules whose support is low in the whole times, but at a certain time the support and confidence is high [7]. For example, the purchases of the Christmas hats at Christmas Day have an important value for users, which would usually ignored in the traditional association analysis.

In the analysis of dynamic association rules, the minimum support threshold is set as 0.5. In reality, because the time of some itemsets existing is too short, the dynamic threshold is established to put forward the useless rules. This paper will set the dynamic threshold as 2(i.e. a item appears in two or more time periods, it may be simply presumed that there is a certain reason for its appearance, which has a great guidance meaning to the real situation). The following part analyzes four evaluation methods of the dynamic association rules. For simplify, this article only analyzes the shopping basket data of the first record at six different time periods. The customer's data at six different time periods is shown as table 4:

Table 4. The First Customer's the Shopping Basket Data at Six Different Time Periods

Time periods	1	2	3	4	5	6
Shopping Basket	BE	CE	AF	BE	CE	BCE
Weight	0.2	0.15	0.15	0.35	0.05	0.1

Note: the weight refers to the proportion of customers' respective consumption amount to the total consumption amount at six time periods.

If there is not specific instruction, the dynamic analysis of this part is all based on the six time periods.

4.1. The Traditional Association Analysis without the Life Circle

Table 5. The Analysis Table of the Traditional Association Rule

C1	*Itemsets*	*A*	*B*	*C*	*E*	*F*
	Support	0.167	0.5	0.5	0.83	0.1667
L1	Itemsets	B	C	E		
	Support	0.5	0.5	0.8		
C2	Itemsets	BC	BE	CE		
	Support	0.167	0.5	0.5		

L2	Itemsets	BE	CE			
	Support	0.5	0.5			
C3	Itemsets					

The table 5 shows that the customer frequently purchases ***B***, ***C*** and ***E*** products, especially for ***E***, and its support can be up to 83%. So, we can conclude that these three products may be daily or customer's preferred commodities. In addition, we also find that when the customer purchases ***B*** or ***C***, they usually purchase ***E*** also. Therefore, when display the commodities on the shelves, ***B***, ***C*** and ***E*** should be displayed together. Moreover, ***B*** and ***E*** or ***C*** and ***E*** can be tied up for promotion, which can bring more profits for the mall. The most important thing is that it can save time for customers and provide happier shopping experience for customers.

4.2. The Association Rule with the Life Circle

In real life, we often hear the life circle of products, plants, economic fluctuations, *etc*.. Similarly, the customers' the consumption habit and preference have the life circle also. In table 6, the life circle of item set ***A*** is shown as [3, 3]. Among the six time periods, the customer purchases ***A*** only at the third time period. Therefore, we can conclude that the consumer purchases ***A*** by chance, and finds out that ***A*** is not good, so he no longer purchases ***A***. Another situation is that ***A*** is a household electrical appliance and the consumer will not often buy it. The life circle of ***C*** is shown as [2, 6], the consumer purchases ***C*** during the second and sixth time period. Generally, the analysis result obtained by the association rule with the life cycle is similar to that of the traditional association rule without the life cycle, but the support (0.6) of purchasing ***C*** and ***E*** simultaneously is larger than that (0.5) of purchasing ***B*** and ***E*** simultaneously**,** therefore, the association rule with life circle has a higher recognition.

Table 6. The Analysis Table of the Association Rule Based on Life Circle

C1	***Itemsets***	***A***	***B***	***C***	***E***	***F***
	Support	1	0.5	0.6	0.83	1
	Life circle	[3,3]	[1,6]	[2,6]	[1,6]	[3,3]
L1	Itemsets	B	C	E		
	Support	0.5	0.6	0.8		
	Life circle	[1,6]	[2,6]	[1,6]		
C2	Itemsets	BC	BE	CE		
	Support	0.2	0.5	0.6		
	Life circle	[2,6]	[1,6]	[2,6]		
L2	Itemsets	BE	CE			
	Support	0.5	0.6			
	Life circle	[1,6]	[2,6]			
C3	Itemsets					

4.3. The Weighted Dynamic Association Rule

On the basis of the analysis on the association rule with life circle, we put forward a new dynamic association rule, namely the weighted dynamic association rule [9]. It not only absorbs the advantages of the association rule with life circle, but also provides more useful information (i.e., if the time period is closer, the information provided is more, more accurate, and more efficient, especially for something with stronger timeliness.). Therefore, we set the weight coefficients successively as 0.1, 0.2, 0.3, 0.4, 0.5 and 0.6 according to the different time period from far to near. The calculating result is shown in Table 7.

Table 7. The Analysis Table of the Weighted Dynamic Association Rule

C1	*Itemsets*	*A*	*B*	*C*	*E*	*F*
	Support	1	0.52	0.65	0.9	1
	Life circle	[3,3]	[1,6]	[2,6]	[1,6]	[3,3]
L1	Itemsets	B	C	E		
	Support	0.52	0.65	0.85		
	Life circle	[1,6]	[2,6]	[1,6]		
C2	Itemsets	BC	BE	CE		
	Support	0.3	0.52	0.61		
	Life circle	[2,6]	[1,6]	[2,6]		
L2	Itemsets	BE	CE			
	Support	0.52	0.61			
	Life circle	[1,6]	[2,6]			
C3	Itemsets					

The table 7 shows that the result of the weighted dynamic association rule is roughly similar to that of the association rule with the life cycle. The main difference is that the time period is closer, the greater impact on the result. Therefore, the result based on the weighted dynamic association rule is more sensitive to the time changing, and the calculating result is more accurate than that of the association rule with life cycle. In addition, when we only consider the shopping basket during the former five time periods, the result is shown as Table 8.

Table 8. The Analysis Table of the Weighted Dynamic Association Rule during the Former Five Time Periods

C1	*Itemsets*	*A*	*B*	*C*	*E*	*F*
	Support	1	0.681	0.23	0.8	1
	Life circle	[3,3]	[1,4]	[2,5]	[1,5]	[3,3]
L1	Itemsets	B	E			
	Support	0.6809	0.827			
	Life circle	[1,4]	[1,5]			
C2	Itemsets	BE				
	Support	0.681				
	Life circle	[1,4]				

Compared the above two tables, we can see that, when the sixth time period is added, the support to purchase ***B*** and ***E*** simultaneously is reducing by 0.161. Meanwhile, purchasing ***C*** and ***E*** simultaneously occurs, and the support is up to 0.61. Therefore, the weighted dynamic association rule has a good sensitivity for the consumption situation at different time periods, that is to say, the consumption situation at the closer time period, the greater impact on the judgment of the association rule. Conversely, we can also conclude the recent customers' consumer behaviors according to the weighted dynamic association rule.

4.4. The Weighted Dynamic Association Rule Weighted by the Consumption Amount

Although the weighted dynamic association rules provide more effective and timely information, the weighted dynamic association rules can't truly reflect the customers' consumption features. As for the following situations, the results that the weighted dynamic association rules conclude are the same: (1) the customer purchases a lot of ***C*** and ***E*** during the fifth time period and purchases rather little ***C*** and ***E*** during the sixth time period; (2) the customer purchases rather little ***C*** and ***E*** during the fifth time period and purchases a lot of ***C*** and ***E*** during the sixth time period. It is obvious that the situation 1 is different from the situation 2, but the weighted dynamic association rules can not accurately display this difference. On account of this, we put forward the weighted dynamic association rule weighted by the consumption amount. The analysis result is as shown in Table 9.

Table 9. The Analysis Table of the Weighted Dynamic Association Rule Weighted by the Consumption Amount

C1	***Itemsets***	*A*	*B*	*C*	*E*	*F*
	Support	1	0.688	0.38	0.9	1
	Life circle	[3,3]	[1,6]	[2,6]	[1,6]	[3,3]
L1	Itemsets	B	E			
	Support	0.6875	0.859			
	Life circle	[1,6]	[1,6]			
C2	Itemsets	BE				
	Support	0.688				
	Life circle	[1,6]				

The Table 9 shows that when ***C*** is purchased each time, the customer’s consumption amount is rather small, and due to this, the support to purchase ***C*** for this customer is low, only 0.3833. However, when ***B*** and ***E*** are purchased each time, the consumption weight is relatively high, therefore the support to purchase ***B*** and ***E*** for this customer is relatively high and up to 0.688, which is 0.168 higher than the support (0.52) obtained by the weighted dynamic association rules.

5. Summary

This paper makes an analysis from five parts for the improvement of the evaluation method and the measure frame of the association rule. In part one, the article introduces the necessity of the association technology in the data mining. Because the interestingness evaluation of the association rule has a very important significance for the practical application of the association rule mining technology, this paper discusses the improvement the evaluation method and the measure frame of the association rule. In part two, this paper analyzes the interestingness measure indicators of the association rule from three aspects: the objective measure indicator, the subjective measure indicator and the association rule indicator based on statistical perspective. In part three, this article analyzes the evaluation and measure frame of the traditional association rule based on the support-confidence, and finds out some significant association rules. On the basis of the above analyses, In order to improve the efficiency of the association rules and meet the user's interestingness better, this paper puts forward the dynamic association rule analysis in part four, and make a comparative analyses from the following four aspects: the traditional association analysis without the life cycle, the association rules with the life cycle, the weighted dynamic association rules, and the weighted dynamic association rules weighted by the consumption amount. These four methods are progressive, showing us the impact of the timeliness on the association rules as well as the deep influence of the different proportions of the customers shopping amount at different times on the association rule technology. In the last part, we make a summary.

References

[1] R. Agrawal and R. Srikant, "Mining association rules between sets of items in large databases", Proceedings of the ACM SIGMOD international conference on Management of data, Washington DC, **(1993)**, pp. 207- 216.

[2] R. Agrawal, T. Imielinski and A. Swami, "Database mining: a performance perspective", IEEE Transactions on Knowledge and Data Engineering, 1993, vol. 5, no. 6, **(1993)**, pp. 914-925.

[3] J. Chunguang, B. Fuguang and W. Zongge, "Research on Improvement of Assessment Method and Measure Framework for Association Rules, Journal of the China Society for Scientific and Technical Information, vol. 32, no. 6, **(2013)**, pp. 584-592.

[4] T. Xueqing, L. Lin and Z. Dongru, "The Comparative Study on the Interestingness Measures for Mining Association Rule, Journal of the China Society for Scientific and Technical Information, vol. 26, no. 2, **(2007)**, pp. 266-270.

[5] Y. Weiguo, W. Jinmao and W. Mingyang, "Improvement on Mining Association Rules", Journal of Northeast Normal University (Natural science Edition), vol. 38, no. 2, **(2006)**, pp. 15-18.

[6] L. Ke and W. Jie, "Research on Judgment Criterion of Association Rule, Control and Decision, vol. 18, no. 3, **(2003)**, pp. 277-280.

[7] Q. Yanxie, "Measurement of Novelty: Factor of Evaluation for the Association Rules", Application Research of Computers, **(2004)**, no. 1, pp. 7-19.

[8] G. Chonghui and Z. Zhen, "Interestingness Evaluation of Association Rules Based on Combination Evaluation Method", Journal of the China Society for Scientific and Technical Information, 2011, vo. 30, no 4, **(2011)**, pp. 353-360.

[9] Z. Xinxie and W. Yaoqing, "Correlation–Based Interestingness Association Rules Mining", Computer Engineering & Science, **(2003)**, no. 3, pp. 60-62.

[10] Z. Jianping and L. Yanbo, "Construction of Weighted Temporal Association Rules in Data Mining [J]，Computer Engineering, vol. 34, no. 6, **(2008)**, pp. 51-53.

[11] L. Weidong, N. Zhiwei and L. Xiao, "Mining Association Rues Based on Interest Measure", Computer Technology and Development, vol. 17, no. 6, **(2007)**, pp. 80- 86.

[12] W. Yongliang and C. Lian, "Valid Association rules Based on Lift-calculation", Computer Engineering, no. 3, **(2003)**, pp. 60-62.

[13] D. H. Choi, B. S. Ahn, S. H. Kim, "Prioritization of Association Rules in Data Mining: Multiple criteria decision approach", Expert Systems with Application, vol. 29, no. 4, **(2005)**, pp. 867-878.

[14] M. C. Chen, "Ranking Discovered rules from Data Mining with Multiple Criteria by Data Envelopment Analysis", Expert Systems with Application, vol. 33, no. 4, **(2007)**, pp. 1100-1106.

[15] M. Toloo, B. Sohrabi, S. A. Nalchigar, "New Method for Ranking Discovered Rules from Data Mining by DEA", Expert Systems with Application, vol. 36, no. 4, **(2009)**, pp. 8503-8508.

Authors

Gao Yongmei, Female, 1975 birth, master, associate professor; Research area: data processing, data mining, *etc.*

Bao Fuguang, Male, 1986 birth, Dr.; Research area: Intelligent information processing le**:** data mining, *etc.*