

Study of Positive and Negative Association Rules Based on Multi-confidence and Chi-Squared Test*

Xiangjun Dong¹, Fengrong Sun², Xiqing Han³, and Ruilian Hou¹

¹ School of Information Science and Technology, Shandong Institute of Light Industry,
250100 Jinan, China

{dxj, hrl}@sdili.edu.cn

² School of Information Science and Engineering, Shandong University, 250100 Jinan, China
sunfr@sdu.edu.cn

³ Dept. of business administration, Shandong Institute of Commerce and Technology,
250103 Jinan, China
hanxq@sict.edu.cn

Abstract. Using a single confidence threshold will result in a dilemmatic situation when simultaneously studying *positive and negative association rule* (PNAR), i.e., the forms $A \Rightarrow B$, $A \Rightarrow \neg B$, $\neg A \Rightarrow B$ and $\neg A \Rightarrow \neg B$. A method based on four confidence thresholds for the four forms of PNARs is proposed. The relationships among the four confidences, which show the necessity of using multiple confidence thresholds, are also discussed. In addition, the chi-squared test can avoid generating misleading rules that maybe occur when simultaneously studying the PNARs. The method of how to apply chi-squared test in mining association rules is discussed. An algorithm PNARMC based on the chi-squared test and the four confidence thresholds is proposed. The experimental results demonstrate that the algorithm can not only generate PNARs rightly, but can also control the total number of rules flexibly.

1 Introduction

Traditional *association rule* (AR) is the form $A \Rightarrow B$, which is used to find the relationships of item sets in transaction database. $A \Rightarrow B$ is a valid rule if its support $s(A \Rightarrow B)$ and confidence $c(A \Rightarrow B)$ meet minimum support (*mins*) and minimum confidence (*minc*) [1]. As an important supplement of the form $A \Rightarrow B$, this paper will study the other three forms $A \Rightarrow \neg B$, $\neg A \Rightarrow B$ and $\neg A \Rightarrow \neg B$, which are called *negative association rules* (NAR), and the traditional form $A \Rightarrow B$ *positive association rules* (PAR) [2].

NARs can provide much useful information for decision-making. They play an important role in many areas, especially in competitive analysis and investment analysis. So, the study of NAR is being attached much more attention recently. The negative relationships between two frequent items were first mentioned in 1997 [3].

* This work was supported by the National Nature Science Foundation of China (NNSFC) under the grant 60271015.

The extended association rules and atom association rules are discussed in [4]. In [5] the authors propose strong negative association rules. In [2] and [6] the authors proposed a PR (Probability Ratio) model and proposed a new method to discover both positive and negative association rules in 2002, and in 2004 they proposed an improved efficient method that could be used to large databases with a pruning strategy and an interestingness measure [7]. In [8] the authors propose a genetic algorithm to discover negative association rules.

However, the above-mentioned papers do not consider the problems caused by using a single confidence threshold when mining positive and negative association rules (PNARs) simultaneously. Intuitively, in basket analysis, the support of almost every item set is very low. Suppose we have two item sets A and B , their support $s(A)$ and $s(B)$ are low, the confidence $c(A \Rightarrow B)$ may be low or high, but $c(\neg A \Rightarrow \neg B)$ must be very high. If we use a single confidence threshold, a dilemmatic situation would occur: if we use a lower confidence threshold, the numbers of PNARs would be so many that the user can not easily choose the right rules; if we use a higher confidence threshold, some valued PARs would be missing. A good solution is using four different confidence thresholds for the four forms of rules. But the relationships among the four confidences must be taken careful consideration when setting these confidence thresholds. Although some articles have discussed the question using different supports and confidences to mine association rules [9], but they are limited to the positive association rules, which is different from this paper.

Furthermore, some self-contradictory association rules would occur when we consider the positive and negative association rules simultaneously.

Example 1. Suppose we are interested in analyzing the transactions of apples (denoted by A) and bananas (denoted by B) in a supermarket. In 10000 transactions, 6000 transactions contain apples, 6500 transactions contain bananas and 3600 transactions contain both apples and bananas. Then,

$$s(A \cup B) = 3600/10000 = 0.36, \text{ and}$$

$$c(A \Rightarrow B) = 3600/6000 = 0.60.$$

Suppose $mins = 0.25$ and $minc = 0.55$, the rule $A \Rightarrow B$ is a valid association rule, which namely shows that increasing the transactions of apples can correspondingly increase the transactions of bananas. But is it true? Let's see another rule $\neg A \Rightarrow B$.

$$s(\neg A \cup B) = s(B) - s(A \cup B) = 0.65 - 0.36 = 0.29 > mins, \text{ and}$$

$$c(\neg A \Rightarrow B) = s(\neg A \cup B) / (1 - s(A)) = 0.725 > minc.$$

Obviously, $\neg A \Rightarrow B$ is also a valid association rule, which shows that reducing the transactions of apples can correspondingly increase the transactions of bananas. Obviously, the two valid rules are self-contradictory. In fact, the transaction of apples and bananas is negatively correlated. Buying one will reduce the possibility of buying another. Here the rule $A \Rightarrow B$ is a misleading rule; another case that can generate misleading rules is when two item sets are independent of each other, buying one doesn't affect buying another.

In order to eliminate these misleading rules, many kinds of measure are proposed, such as interestingness, chi-squared test, correlation coefficient, Laplace, Gini-index, Piatetsky-Shapiro, Conviction and so on [10,11]. In these measures, the chi-squared

test is one of the most interesting measures because of its mature theoretical basis. The authors in [3,12] propose a method to test the correlation and the independence of association rules using chi-squared test. In [13] the authors use chi-squared test to prune non-actionable rules or rules at lower significance level. In [14] the authors also use chi-squared independent test in rules classification. In [15] the authors propose a method to express the chi-squared value with support, confidence and lift.

However, the purpose of using the chi-squared test in existing works is to find those item sets whose chi-squared value shows negative correlation and then in turn avoid these item sets generating rules, while our works are not only to find out PARs in the item sets with positive correlation, to avoid generating rules in independent item sets, but also to find out NARs in the item sets with negative correlation.

The contribution of this paper lies in three aspects: 1) study the relationships of the four confidences; 2) look for a method of how to apply chi-squared test in mining PNARs; and 3) propose an algorithm to generate PNARs based on chi-squared test and four confidence thresholds.

The rest of this paper is organized as follows. Section 2 discusses the relationships among the four confidences. Section 3 discusses the method of how to use the chi-squared test in mining PNARs. Section 4 is about algorithm design. Section 5 is about experiment and comparisons. Section 6 is conclusion.

2 The Study of the Relationships Among the Four Confidences

Now we discuss the relationships among the four confidences in the following four cases (A and B are two item sets).

(1) $s(A)$ and $s(B)$ are very low (the typical case in basket analysis); (2) $s(A)$ and $s(B)$ are very high;

(3) $s(A)$ is very high, $s(B)$ is very low; (4) $s(A)$ is very low, $s(B)$ is very high.

However, the concept of high or low here is not explicit. In order to discuss the relationships easily, we give an explicit value that the high is not less than 0.9 and the low not more than 0.1.

Based on the study in [16], we can express the confidences $c(A \Rightarrow \neg B)$, $c(\neg A \Rightarrow B)$ and $c(\neg A \Rightarrow \neg B)$ as the functions of $s(A)$, $s(B)$ and $c(A \Rightarrow B)$ as follows:

$$c(A \Rightarrow \neg B) = \frac{s(A) - s(A \cup B)}{s(A)} = 1 - c(A \Rightarrow B); \quad (1)$$

$$c(\neg A \Rightarrow B) = \frac{s(B) - s(A \cup B)}{1 - s(A)} = \frac{s(B) - s(A) * c(A \Rightarrow B)}{1 - s(A)}; \quad (2)$$

$$c(\neg A \Rightarrow \neg B) = \frac{1 - s(A) - s(B) + s(A \cup B)}{1 - s(A)} = \frac{1 - s(A) - s(B) + s(A) * c(A \Rightarrow B)}{1 - s(A)} \quad (3)$$

Obviously, we can get: (1) $c(A \Rightarrow B) + c(A \Rightarrow \neg B) = 1$ and (2) $c(\neg A \Rightarrow B) + c(\neg A \Rightarrow \neg B) = 1$. The range of the value $c(A \Rightarrow B)$ can be shown in Theorem 1.

Theorem 1. $\text{MAX}(0, \frac{s(A)+s(B)-1}{s(A)}) \leq c(A \Rightarrow B) \leq \text{MIN}(1, \frac{s(B)}{s(A)})$.

Proof: Because $s(A \cup B) \leq \text{MIN}(s(A), s(B))$, thus, $c(A \Rightarrow B) \leq \text{MIN}(1, \frac{s(B)}{s(A)})$;

The minimal value of $s(A \cup B)$ obviously is 0 ,but when $s(A)+s(B)>1$,the minimal value of $s(A \cup B)$ is $s(A)+s(B)-1$, so $c(A \Rightarrow B) \geq \text{MAX}(0, (s(A)+s(B)-1)/s(A))$.

We can easily calculate the value ranges of $c(A \Rightarrow B)$, $c(\neg A \Rightarrow B)$ and $c(\neg A \Rightarrow \neg B)$ according to Theorem 1 and Equation (1) to (3). Table 1 shows the value ranges of the four confidences with the different value of $s(A)$ and $s(B)$.

From Table 1 we can see that the value of confidence has an important impact on the quantity of association rules. Take the first case(the typical basket analysis) for example: if we set $\text{minc}=0.6$, there would be a large number of rules of the form $\neg A \Rightarrow \neg B$ in the result sets, but not any rule of the form $\neg A \Rightarrow B$. It is obviously unreasonable. In order to get the rules of all the four forms, multi-confidences for different forms of rules are necessary in real application.

This paper uses minc_{11} , minc_{10} , minc_{01} and minc_{00} to express the four minimum confidence sof the four forms of rules $A \Rightarrow B$, $c(A \Rightarrow B)$, $c(\neg A \Rightarrow B)$ and $c(\neg A \Rightarrow \neg B)$ respectively.

Table 1. The value ranges of the four confidences with different $s(A)$ and $s(B)$

	$c(A \Rightarrow B)$	$c(A \Rightarrow \neg B)$	$c(\neg A \Rightarrow B)$	$c(\neg A \Rightarrow \neg B)$
$s(A)=0.1, s(B)=0.1$	[0,1]	[0,1]	[0,0.11]	[0.89,1]
$s(A)=0.9, s(B)=0.9$	[0.89,1]	[0,0.11]	[0,1]	[0,1]
$s(A)=0.1, s(B)=0.9$	[0,1]	[0,1]	[0.89,1]	[0,0.11]
$s(A)=0.9, s(B)=0.1$	[0,0.11]	[0.89,1]	[0,1]	[0,1]

3 The Study of Applying Chi-Squared Test to Mining PNARs

In statistics, chi-squared test is a widely used method to test the independence or correlation among variables. The variables are first assumed independent, and then their chi-squared value is calculated. If the value is more than a given value (the value is 3.84 when the significance level $\alpha=0.05$), then the independent assumption would be

refused. The chi-squared value can be calculated by $\chi^2 = \sum \frac{[f_o - f_e]^2}{f_e}$, where f_o

denotes observing frequency, f_e expected frequency. Specifically, for two random variables (X, Y) , their n samples of $(X_1, Y_1) (X_2, Y_2) \dots, (X_n, Y_n)$ are placed in a contingency

table. Suppose the contingency table has R rows and C columns, the frequency of a sample falling into certain a cell denotes as N_{ij} . Then we can get: $N_{i\cdot} = \sum_{j=1}^C N_{ij}$,

$$N_{\cdot j} = \sum_{i=1}^R N_{ij}, \text{ and}$$

$$\chi^2 = \sum \frac{[f_o - f_e]^2}{f_e} = \sum_{i=1}^R \sum_{j=1}^C \frac{(N_{ij} - \frac{1}{n} N_{i\cdot} N_{\cdot j})^2}{\frac{1}{n} N_{i\cdot} N_{\cdot j}}. \quad (4)$$

It obeys χ^2 -distribution of $(R-1)(C-1)$ degrees of freedom, where $f_o = N_{ij}$, $f_e = \frac{1}{n} N_{i\cdot} * N_{\cdot j}$.

Chi-squared testⁿ could be used in multi-dimensional variables, but in association rule we only need two-dimensional variables, i.e., two-dimensional

Table 2. Contingency table of items A and B

	A	$\neg A$	Σ
B	$s(A \cup B) * n$	$s(\neg A \cup B) * n$	$s(B) * n$
$\neg B$	$s(A \cup \neg B) * n$	$s(\neg A \cup \neg B) * n$	$s(\neg B) * n$
Σ	$s(A) * n$	$s(\neg A) * n$	n

contingency table, its degree of freedom is 1. The contingency table of item sets A and B is shown in table 2, where n denotes the total number of transactions in database. The purpose of the data expressed in the form of support in Table 2 is to associate the chi-squared test with the concepts of association rules easily.

According to Equation (4), we can get:

$$\begin{aligned} \chi^2 &= \frac{[s(A \cup B) * n - \frac{1}{n} s(A) * n * s(B) * n]^2}{\frac{1}{n} s(A) * n * s(B) * n} + \frac{[s(\neg A \cup B) * n - \frac{1}{n} s(\neg A) * n * s(B) * n]^2}{\frac{1}{n} s(\neg A) * n * s(B) * n} \\ &+ \frac{[s(A \cup \neg B) * n - \frac{1}{n} s(A) * n * s(\neg B) * n]^2}{\frac{1}{n} s(A) * n * s(\neg B) * n} + \frac{[s(\neg A \cup \neg B) * n - \frac{1}{n} s(\neg A) * n * s(\neg B) * n]^2}{\frac{1}{n} s(\neg A) * n * s(\neg B) * n} \\ &= \frac{n * [s(A \cup B) - s(A)s(B)]^2}{s(A)s(B)(1-s(A))(1-s(B))} \end{aligned} \quad (5)$$

Here is an instance. Table 3 shows the contingency table of data in Example 1. Its chi-squared value is:

$$\frac{\left(3600 - \frac{6000 * 6500}{10000}\right)^2}{\frac{6000 * 6500}{10000}} + \frac{\left(2900 - \frac{4000 * 6500}{10000}\right)^2}{\frac{4000 * 6500}{10000}} + \frac{\left(2400 - \frac{6000 * 3500}{10000}\right)^2}{\frac{6000 * 3500}{10000}} + \frac{\left(1100 - \frac{4000 * 3500}{10000}\right)^2}{\frac{4000 * 3500}{10000}} = 16$$

4.8 > 3.84.

Since 16.48 are more than 3.84, we reject the independence assumption at the 5% significance level. Thus, the transactions of apples and bananas are correlated.

But how to judge the correlation between item sets are positive or negative?

Definition 1. The correlation of item sets A and B can be defined as $\text{corr}(A, B) = f_o / f_E$, where f_o and f_E is the observing frequency and expected frequency respectively of the cell that A and B cross. Then we can get: (1) if $\text{corr}(A, B) > 1$, then $A \Rightarrow B$ is positively correlated, and (2) if $\text{corr}(A, B) < 1$, then $A \Rightarrow B$ is negatively correlated.

According to Definition 1 and Table 2 we can get:

- (1) $\text{corr}(A, B) = s(A \cup B) / (s(A)s(B))$;
- (2) $\text{corr}(A, \neg B) = s(A \cup \neg B) / (s(A)s(\neg B))$;
- (3) $\text{corr}(\neg A, B) = s(\neg A \cup B) / ((s(\neg A)s(B))$; and
- (4) $\text{corr}(\neg A, \neg B) = s(\neg A \cup \neg B) / (s(\neg A)s(\neg B))$.

Now we discuss the relationships of the above four correlations.

Theorem 2. If $\text{corr}(A, B) > 1$, then (1) $\text{corr}(A, \neg B) < 1$; (2) $\text{corr}(\neg A, B) < 1$; (3) $\text{corr}(\neg A, \neg B) > 1$ and vice versa.

Table 3. Contingency table of the data in Example 1

Proof: Here we only prove (3), (1) and (2) can be proved in the same way.

(3) Since $\text{corr}(A, B) > 1$, thus $s(A \cup B) > s(A)s(B)$,

so $1 - s(A) - s(B) + s(A \cup B) > 1 - s(A) - s(B) + s(A)s(B)$.

Because $\text{corr}(\neg A, \neg B) > 0$, namely $1 - s(A) - s(B) + s(A)s(B) > 0$, therefore,

$$\frac{1 - s(A) - s(B) + s(A \cup B)}{1 - s(A) - s(B) + s(A)s(B)} = \frac{s(\neg A \cup \neg B)}{s(\neg A)s(\neg B)} > 1, \text{ i.e., } \text{corr}(\neg A, \neg B) > 1.$$

These relationships of the four correlations tell us self-contradictory rules would not occur if we judge the correlation between item sets before generating rules. In detail, we only mine rules of the forms $\neg A \Rightarrow B$ and $A \Rightarrow \neg B$ if $\text{corr}(A, B) < 1$ and mine rules of the forms $A \Rightarrow B$ and $\neg A \Rightarrow \neg B$ if $\text{corr}(A, B) > 1$. In Example 1, we only mine rules of the forms $\neg A \Rightarrow B$ and $A \Rightarrow \neg B$ because of $\text{corr}(A, B) = 0.92 < 1$.

Now we give the definition of positive and negative association rules based on chi-squared test.

Definition 2. Let I be a set of items, D a database, $A, B \subseteq I$ and $A \cap B = \emptyset$. $s(A)$, $s(\neg A)$, $s(B)$ and $s(\neg B) > 0$, mins and minc are given by users. Let χ^2 express the Chi-squared value of A, B and χ_α^2 the critical value at the significance level α . If $\chi^2 \leq \chi_\alpha^2$, A and B are independent, we don't generate any rules from them. Otherwise, A and B are correlated, and

(1) If $\text{corr}(A, B) > 1$, $s(A \cup B) \geq \text{mins}$ and $c(A \Rightarrow B) \geq \text{minc}$ then $A \Rightarrow B$ is a PAR;

(2) If $\text{corr}(A, \neg B)$ ($\text{corr}(\neg A, B)$, $\text{corr}(\neg A, \neg B)$) > 1 , $s(A \cup B) \geq \text{mins}$ and $c(A \Rightarrow \neg B)$ ($c(\neg A \Rightarrow B)$, $c(\neg A \Rightarrow \neg B)$) $\geq \text{minc}$, then $A \Rightarrow \neg B$ ($\neg A \Rightarrow B$, $\neg A \Rightarrow \neg B$) is a NAR;

It is obvious that chi-squared test is a very effective method. But chi-squared test also has some limitations [3]. The chi-squared test rests on the normal approximation to

the binomial distribution. This approximation breaks down when the expected values are small. As a rule of thumb, statistics texts recommend the use of chi-squared test only if (1) all cells in the contingency table have expected value greater than 1; and (2) at least 80% of the cells in the contingency table have expected value greater than 5.

4 Algorithm Design

According to Definition 2, we propose an algorithm PNARMC (Positive and Negative Association Rules based on Multi-confidence and Chi-squared test). We suppose the frequent item sets (those item sets whose support is great than the minimum support threshold $mins$) are saved in set L .

Algorithm. PNARMC

Input: L : frequent item sets; $minc_11, minc_10, minc_01, minc_00$: minimum confidence; χ_a^2 : the critical value at the significance level α ;

Output: PAR : set of positive associate rule; NAR : set of negative associate rule;

(1) $PAR = \emptyset$; $NAR = \emptyset$;

(2) //mining PNARs in L .

```

for any item set  $X$  in  $L$  do {
  for any item set  $A \cup B = X$  and  $A \cap B = \emptyset$  do {
    calculate  $\chi^2$  with formula 5;
     $corr = s(A \cup B) / (s(A) s(B)) \square$ 
    if  $\chi^2 > \chi_a^2$  then {
      if  $corr > 1$  then {
        (2.1) //generate rules of the forms  $A \Rightarrow B$  and  $\neg A \Rightarrow \neg B$ .
        if  $c(A \Rightarrow B) \geq minc\_11$  then
           $PAR = PAR \cup \{A \Rightarrow B\}$ ;
        if  $c(\neg A \Rightarrow \neg B) \geq minc\_00$  then
           $NAR = NAR \cup \{\neg A \Rightarrow \neg B\}$ ; }
      if  $corr < 1$  then {
        (2.2) //generate rules of the forms  $A \Rightarrow \neg B$  and  $\neg A \Rightarrow B$ .
        if  $c(A \Rightarrow \neg B) \geq minc\_10$  then
           $NAR = NAR \cup \{A \Rightarrow \neg B\}$ ;
        if  $c(\neg A \Rightarrow B) \geq minc\_01$  then
           $NAR = NAR \cup \{\neg A \Rightarrow B\}$ ; }
    } } }

```

(3) **return** PAR and NAR ;

Step (1) initializes PAR and NAR with empty set. Step (2) calculates chi-squared value χ^2 and correlation flag $corr$ and generates rules. Step (2.1) generates rules of the forms $A \Rightarrow B$ and $\neg A \Rightarrow \neg B$ and step (2.2) generates rules of the forms $A \Rightarrow \neg B$ and $\neg A \Rightarrow B$. Particularly, it is four confidence thresholds that are used to their corresponding four forms of rules, but not a single confidence threshold. Step (3) returns the result and finishes the whole algorithm.

In fact, the algorithm PNARMC can also be used to generate NARs from infrequent item sets if only we delete the sentence “if $c(A \Rightarrow B) \geq \text{minc_11}$ then $PAR = PAR \cup \{A \Rightarrow B\}$ ” and let L save the infrequent item sets.

5 Experimental Results and Comparison

5.1 Experimental Results

The experimental dataset records areas of www.microsoft.com that each user visited in a one-week timeframe in February 1998. Summary statistical information of the dataset is: 32711 training instances, 5000 testing instances, 294 attributes and the mean area visits per case is 3.0 (http://www.cse.ohio-state.edu/~yanghu/CIS788_dm_proj.htm#datasets). The experimental results are shown in Table 4. The number of PARs generated by traditional algorithm (without considering correlation) is also given for comparison.

Table 4. The number of rules generated by different algorithms and in different confidences ($\text{mins}=0.014$, $\alpha=0.05$)

#	traditional algorithm	PNARMC ($\alpha=0.05$)					total
	$A \Rightarrow B$	$A \Rightarrow B$	$A \Rightarrow \neg B$	$\neg A \Rightarrow B$	$\neg A \Rightarrow \neg B$	##	
$\text{minc_}* = 0.3$	167	158	30	7	339	4	538
$\text{minc_}* = 0.6$	43	43	30	0	339	0	412
$\text{minc_}* = 0.85$	7	7	14	0	263	0	284
$\text{minc_}* = 0.985$	1	1	0	0	32	0	33
$\text{minc_11} = 0.6$							
$\text{minc_10} = 0.85$	—	43	14	7	32	4	100
$\text{minc_01} = 0.3$							
$\text{minc_00} = 0.985$							

Note: #: $\text{minc_}*$ denotes minc_11 , minc_10 , minc_01 , minc_00

##: This column expresses the number of rules generated by independent item sets.

From Table 4 we can draw two conclusions as follows:

1. The Chi-squared test is effective.

From Table 4, when $\text{minc_}* = 0.3$, there are 158 PARs generated by PNARMC, while 167 PARs generated by traditional algorithm. This shows that there are 9 misleading PARs detected and eliminated by PNARMC. In these 9 rules, 4 rules are generated by independent item sets, the other 5 rules are maybe generated by negative correlated item sets. Moreover, a lot of NARs are generated by PNARMC. These data adequately show that chi-squared test is effective.

2. Multi-confidence is effective.

First, let's look at the changes of the rule number when using a single minimum confidence $\text{minc_}*$. When $\text{minc_}*$ changes from 0.3 to 0.6, the rule number of the forms $A \Rightarrow B$ and $\neg A \Rightarrow B$ decreases sharply while the rule number of the forms $A \Rightarrow \neg B$

and $\neg A \Rightarrow \neg B$ doesn't change, the total number of the rules is still great. The only method to decrease the total number of the rules is to increase the minimum confidence $minc_*$. But the higher the minimum confidence $minc_*$ is, the higher the percentage of the rules of the form $\neg A \Rightarrow \neg B$ in the total is, the lower the percentage of the rules of the other three forms in the total is. These data adequately show the shortcomings of using a single minimum confidence.

Second, let's look at the number of the rules when using multiple minimum confidences. when $minc_{11}=0.6$, $minc_{10}=0.85$, $minc_{01}=0.3$ and $minc_{00}=0.985$, the total number of rules is 100, which contain 43 rules of the form $A \Rightarrow B$, 14 rules of the form $A \Rightarrow \neg B$, 7 rules of the form $\neg A \Rightarrow B$, 32 rules of the form $\neg A \Rightarrow \neg B$ and 4 rules generated by independent item sets. It is obvious that the algorithm PNARMC can flexibly control the total number of rules by setting the four confidence thresholds according to the user's actual need. These data adequately show that using multi-confidence is necessary and the algorithm PNARMC is very effective.

5.2 Comparisons

The superiorities of PR model have been illustrated in [6] when the PR model is compared with some other models including interest model, exception mining model and strong negative association model [5]. So, here we only compare the PANRMC model with PR model and positive and negative association rules on correlation (PANRC) model in [16]. The PR model uses a single confidence threshold to generate PARs from frequent item sets and to generate NARs only from infrequent item sets, however, it does not consider the NARs in frequent item sets and does not consider the problems using a single confidence threshold we mentioned above. The PANRC model can generate PARs and NARs from frequent item sets, but it also uses a single confidence threshold. The PANRMC model, as we have seen, has overcome the shortcomings in these models.

6 Conclusions

How to set up the confidence threshold of association rules becomes very important when we simultaneously study the PNARs. It has an important influence on the total number of rules, and in turn on the efficiency for user to choose the valued rules. A single confidence threshold can't meet the need of actual application, while four confidence thresholds are required for the four forms of rule. The relationships of the four confidences are discussed. The corresponding conclusions are very important in how to set up the four minimum confidences. In addition, chi-squared test can avoid generating misleading rules that may occur when simultaneously studying the PNARs. The method of how to apply chi-squared test to mining association rules is discussed. An algorithm PNARMC is proposed to discover both positive and negative association rules based on chi-squared test and four confidence thresholds. The algorithm can not only generate PARs in the item sets with positive correlation, generate NARs in the item sets with negative correlation, but can also detect and eliminate rules generated in independent item sets. It can flexibly control the number of rules, too. The experiment and comparisons demonstrate that the algorithm is effective.

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large database. Proceeding of the 1993 ACM SIGMOD International Conference on Management of Data. New York: ACM Press(1993) 207-216
2. Wu, X., Zhang, C., Zhang, S.: Mining both positive and negative association rules. Proceedings of the 19th International Conference on Machine Learning(ICML-2002). San Francisco: Morgan Kaufmann Publishers (2002) 658-665
3. Brin, S., Motwani, R., Silverstein, C.: Beyond Market: Generalizing Association Rules to Correlations. In Processing of the ACM SIGMOD Conference (1997) 265-276
4. Li, X., Liu, Y., Peng, J.: The extended association rules and atom association rules. Journal of Computer Research and Application, China(2002.12)1740-1750
5. Savasere, A., Omiecinski, E., Navathe, S.: Mining for Strong Negative Associations in a Large Database of Customer Transaction. In Proceedings of the 1998 International Conference on Data Engineering(1998) 494-502
6. Zhang, C., Zhang, S.: Association Rule Mining, LNAI 2307, Springer-Verlag Berlin Heidelberg (2002)47-84
7. Wu, X., Zhang, C., Zhang, S.: Efficient Mining of Both Positive and Negative Association Rules, ACM Transactions on Information Systems, **22**(2004), 3: 381-405
8. Boulicaut, J-F., Bykowski, A., Jeudy, B.: Towards the Tractable Discovery of Association Rules with Negations. In Proceedings of the Fourth International Conference on Flexible Query Answering Systems FQAS'00, Warsaw (PL)(2000) 425-434
9. Liu, B., Hsu, W., Ma, Y.: Mining Association Rules with Multiple Minimum Supports. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, San Diego, CA, USA. (1999)
10. Tan, P., Kumar, V.: Interestingness measures for association patterns: a perspective. KDD-2000 Workshop on Post-processing in Machine Learning and Data Mining(2000)
11. Tan, P-N., Kumar, V., Srivastava, J.: Selecting the Right Interestingness Measure for Association Patterns . Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton (CA) (2002) 32-41
12. Silverstein, C., Brin, S., Motwani, R.: Beyond market baskets: Generalizing association rules to dependence rules. Data Mining and Knowledge Discovery, 2(1)(1998) 39-68
13. Liu, B., Hsu, W., Ma, Y.: Identifying Non-Actionable Association Rules. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, San Francisco, CA;(2001) 329-334
14. Hilderman, R.J., Hamilton, H.J.: Applying Objective Interestingness Measures in Data Mining Systems. In Proceedings of the 4th European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'00), Lecture Notes in Computer Science, Springer-Verlag, Lyon, France(2000) 432-439
15. Sergio, A. Alvarez.: Chi-squared computation for association rules: preliminary results. Technical Report BC-CS-2003-01 July 2003, Computer Science Dept. Boston College Chestnut Hill, MA 02467 USA (2003)
16. Dong, X., Wang, S., Song, H., Lu, Y.: Study on Negative Association Rules. Transactions of Beijing Institute of Technology, China(2004.11)978-981