Computers, Materials & Continua CMC, vol.65, no.1, pp.337-354, 2020

An Improved Algorithm for Mining Correlation Item Pairs

Tao Li[1], Yongzhen Ren[1,*], Yongjun Ren[2] and Jinyue Xia[3]

Abstract: Apriori algorithm is often used in traditional association rules mining, searching for the mode of higher frequency. Then the correlation rules are obtained by detected the correlation of the item sets, but this tends to ignore low-support high-correlation of association rules. In view of the above problems, some scholars put forward the positive correlation coefficient based on Phi correlation to avoid the embarrassment caused by Apriori algorithm. It can dig item sets with low-support but high-correlation. Although the algorithm has pruned the search space, it is not obvious that the performance of the running time based on the big data set is reduced, and the correlation pairs can be meaningless. This paper presents an improved mining algorithm with new association rules based on interestingness for correlation pairs, using an upper bound on interestingness of the supersets to prune the search space. It greatly reduces the running time, and filters the meaningless correlation pairs according to the constraints of the redundancy. Compared with the algorithm based on the Phi correlation coefficient, the new algorithm has been significantly improved in reducing the running time, the result has pruned the redundant correlation pairs. So it improves the mining efficiency and accuracy.

Keywords: Interestingness, item pairs, positive correlation, association rules, redundancy.

1 Introduction

In practice, the association rules $(A \Rightarrow B)$ have the following four combinations between support and correlation: Low-Low, High-High, Low-High, High-Low. The Apriori-based association rules mining algorithm tends to ignore the Low-High rules, and it is also very research-intensive in practice. The purchase records that are often rare are more interesting than those that occur frequently, such as the purchase analysis of luxury goods in shopping malls. Therefore, mining Low-High rules is sometimes more valuable than mining high-support rules.

The traditional association rules mining algorithm, such as Apriori, usually find frequent item sets in Xu et al. [Xu and Dong (2013); Rameshkumar, Sambath and Ravi (2013); Poundekar, Manekar, Baghel et al. (2014); Yuan, Li and Chen (2016); Tandon, Haque

[1] College of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing, 210044, China.

[2] College of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, 210044, China.

[3] International Business Machines Corporation (IBM), New York, USA.

[*] Corresponding Author: Yongzhen Ren. Email: renyongzhen2017@163.com.

Received: 26 February 2019; Accepted: 13 June 2019.

and Mande (2016)] generate the association rules based on item sets. The time and space complexity of the first step is much higher than the second step. Tang et al. [Tang, Xu and Duan (2018)] reduce the complexity of the space. Said et al. [Said, Guillet, Richard et al. (2013)] propose a new association rules of the correlations between item pairs. Feng et al. [Feng, Zhu and Zhang (2016)], MH-Apriori optimizes Apriori algorithm that can improve the efficiency of Apriori algorithm for mining frequent item sets. Pandagale et al. [Pandagale and Surve (2016)], in order to find association rules, the Apriori MapReduce algorithm can be used to better achieve space and time complexity. Xue et al. [Xue, Song, Qin et al. (2015)] propose a mutual-information-based quantitative association rule-mining algorithm (MIQarma) to address traditional approaches to spatio-temporal analysis challenges. Poundekar et al. [Poundekar, Manekar, Baghel et al. (2014)] propose association rule mining and it can reduce the scanning time of database. The classic model of association rules mining is based on support and confidence metrics, Thangarasu et al. [Thangarasu and Sasikala (2014)] use Tree-based Association Rules. Li et al. [Liu and Wang (2013)] propose an association rule mining algorithm and it based on the formal concept analysis that can improve the efficiency of algorithms. Tempaiboolkul [Tempaiboolkul (2013)] propose an algorithm for discovering rare association rules in distributed environment and it can achieve an optimized function. Jiang et al. [Jiang, Luan and Dong (2012)] propose a multi-support (WNAIIMS)-based invariant item set weighted negative association rules mining algorithm. Quan et al. [Quan, Liu, Chen et al. (2012)] propose a new mining frequent item sets algorithm based on matrix and experimental result improves the efficiency. Qian et al. [Qian, Jia, Zhang (2008); Luo and Li (2014)], the improved Apriori algorithms are proposed to improve the efficiency of traditional algorithms. The matrix method is used to scan the database once, and it can optimize the operation and improve the mining efficiency. Although it is relatively simple to extract association rules from frequent item sets, it is easy to produce meaningless misleading rules. Ravi et al. [Ravi and Khare (2014)] propose an Efficient and Optimized Association Rules Mining algorithm EO-ARM. It can increase the efficiency by scanning the data set only once. Yang et al. [Yang, Huang and Jin (2017)] presented an improved algorithm that reduces the time to scan the transaction database while preserving the effect of complete mining, which reduces the running time and improves the efficiency of mining. Davale et al. [Davale and Shende (2015)] use logic to generate the association rules and there is no need to decide value of threshold. In Chen et al. [Chen and Gao (2011)], based on the generation of association rules by frequent item sets, correlation metrics are used to test the rules and to avoid the occurrence of misleading rules. However, the correlation metric introduced in the paper are asymmetrically distributed on both sides with a threshold of 1, and its value does not reflect the correlation strength of the association rules. Su et al. [Su and Guo (2014)] propose an interestingness model based on cosine metric, which makes up for the lack of asymmetry of the probability model of the value 1. Lu [Lu (2012)] avoids weak and misleading association rules. All these algorithms use the Apriori method to mine frequent item sets with high frequencies, and ignore infrequent parts which often contain key value information. Juan et al. [Juan, Li and Feng (2015)] propose the research of deleting redundant association rules and it can get frequent association rules. Xiong et al. [Xiong, Shekhar and Tan (2004); Qian, Feng and Wang (2005)] are non-Apriori class

algorithms. They use the upper bound of the Phi coefficient to reduce the space, and mine all pairs of positive correlations and all pairs of negative correlations. Compared with the traditional Apriori class methods. The algorithms not only mine High-High item pairs, but also mine Low-High item pairs, and at the same time, they improve time performance. However, the running time performance improvement of the algorithms is not obvious on the big data set, and the generated item pairs may be redundant and interestingness for users. To reduce the redundant information and extract the most distinct features, ROI and PCA operations are performed for learned features of convolutional layer or pooling layer. Yue et al. [Yue, Wang and Wang (2014)] reduce the running time and deletes some redundant rules in mining association rules.

In this paper, Algorithm is proposed to mine the pairs of non-redundant positive correlations, and to prune the search space by the upper bound of the interestingness of the superset of the item or item pairs, the algorithm improves the time performance significantly compared with the one based on Phi correlation coefficient. At the same time, the pair of items that are interestingness and redundant for the user are pruned.

The arrangement of this paper is as follows. Section 2 introduces conceptual knowledge related to the new algorithm, including support, rules, strong rules, positive association rules, pairs of positive correlations, and pairs of non-redundant items. Section 3 gives knowledge of interest measure, such as interestingness definition, superset interestingness upper bound, interestingness and relevance measure relationship. Section 4 gives the main ideas and algorithm implementation of the new algorithm. Section 5 gives the experimental simulation results and related performance analysis. Finally, Section 6 summarizes the paper and briefly describes the follow-up study work arrangements.

2 Related work

Definition 2.1: Support: Suppose $I = \{i_1, ..., i_m\}$ is a set of m different items, given a transaction database $D = \{T_1, T_2, ..., T_n\}$, with n data. Each data transaction in the database has $T \subseteq I$. Now assume that there is an item set A, $A \subseteq I$, whose support means that in data set D, the transaction data T contains the percentage of item set A.

$$\sup(A) = \frac{|\text{cov(A,D)}|}{n} \tag{1}$$

where, $\text{cov}(A, D) = \{i : 1 \le i \le n \cap A \subseteq T\}$

Definition 2.2: Rules: Given a transaction database D and a set I, the expression like $(A \Rightarrow B)$ is the association rule. The support of the association rule $(A \Rightarrow B)$ is the proportion of the number of records in data set D, which contain both the item set A and B, $\sup(A \Rightarrow B) = \sup(A \cup B)$. The confidence of the rule $(A \Rightarrow B)$ is the ratio of the number of transactions containing both item sets A and B to the number of transactions containing item set A.

$$confidence(A \Rightarrow B) = \frac{\sup(A \cup B)}{\sup(A)} \tag{2}$$

Definition 2.3: Strong rules: Set the minimum support threshold t_1 and the minimum confidence threshold t_2. When rule $(A \Rightarrow B)$ meets the two minimum thresholds, the association rule $(A \Rightarrow B)$ is the strong rule.

Definition 2.4: Positive association rules: Example 1: This is a relationship between people who love coffee and who love tea. Suppose A means that people who buy tea, B means that people who buy coffee. Assume that the minimum support of the rule is 0.1, and the minimum confidence is 0.6. The following conclusions can be drawn by analyzing the Tab. 1.

Table 1: Purchases of coffee and tea

	Buy coffee	Not buying	
Buy tea	150	50	200
Not buying	650	150	800
	800	200	1000

From Tab. 1, We can get $\sup(A \Rightarrow B) = 0.15$, $confidence(A \Rightarrow B) = 0.75$. So $(buy\_tea \Rightarrow buy\_coffee)$ is a strong rule. But on the other hand, the rule $(do\_not\_buy\_tea \Rightarrow buy\_coffee)$ has greater confidence and accuracy, more than 80%, that is, it is more likely to buy coffee without buying tea. Therefore, the strong rule excavated according to the traditional algorithm is wrong at this time and it is a negative association rule. It is impossible to mine the $(do\_not\_buy\_tea \Rightarrow buy\_coffee)$. To avoid the insufficient of traditional algorithms, the correlation metric can be used here.

$$correlation_{A,B} = \frac{\sup(AB)}{\sup(A)\sup(B)} \tag{3}$$

$correlation_{A,B} \in [0,+\infty)$. When $correlation_{A,B} > 1$, it means that there is a positive correlation between A and B; when $correlation_{A,B}=1$, it means that A and B are independent of each other; When $0 < correlation_{A,B} < 1$, A and B are negatively correlated. So the interesting rules can be mined by using this correlation measure. However, this kind of measure has the following disadvantages. First, the range of $correlation_{A,B}$ value is $[0,+\infty)$, which is asymmetrical on both sides of the value 1, and cannot directly reflect the correlation between items. Second, since zero transactions are ubiquitous in real life, the values of $correlation_{A,B}$ are affected by the number of zero transactions and do not have zero invariance. Therefore, it is necessary to make a further improvement for the measurement model of interestingness.

Definition 2.5: Positive correlation pairs: Set the minimum interestingness threshold $min-correlation$. If the interestingness value is bigger than threshold, i.e.,

$Interestingness > min-correlation$, then the items A and B belong to a positive correlation, and then $\{A,B\}$ is a positive correlation pairs.

Definition 2.6: Item pairs without redundancy: A pair of items that satisfy a positive correlation is not necessarily a meaningful pair of items, and what really matters is the pair of items that the user is interested in. If the result of the mining is expected by the user, the item is meaningless to the user. Therefore, the following constraints are met to be a pair of non-redundant pairs of interest to the user. Suppose an item set x has an item i, and if item i contains $x \setminus i$, which is $\text{cov}(x \setminus i) \subseteq (i)$, Then there is an inclusion relationship between the item and the item in the item set x, and the item set x is a redundant item set. For example, there is a pair of items that are positive for $x\{female, pregnancy\}$, but not necessarily meaningful pairs, because pair of items in $x\{female, pregnancy\}$, item $i\{female\}$ already contains $x \setminus i\{pregnancy\}$ this relationship, then the pair of items x is redundant and meaningless.

3 Interestingness measure

3.1 Interestingness definition

Usually, if the rule $(A \Rightarrow B)$ is true, the condition must be met: $p(B \mid A) > p(B)$. It means that the probability of B appearing in the presence of A is higher than the probability of B appearing directly, so that the appearance of B can promoted by A. So a new measure of interestingness will be defined here from the perspective of correlation.

$$Interestingness(A \Rightarrow B) = \frac{p(B \mid A) - p(B)}{p(B \mid A) + p(B)} \tag{4}$$

The interestingness measure has the following properties. First, the measure has lower and upper bounds $[-1,1]$, which can effectively control the input setting of the parameter. Second, if the $Interestingness(A \Rightarrow B) = 0$, it means that the item sets A and B are independent of each other in the pattern; If the $Interestingness(A \Rightarrow B) > 0$, it means that the item set A is positively related to B in the pattern; If the $Interestingness(A \Rightarrow B) < 0$ means that the item set A is negatively correlated with B in the pattern.

Property 3.1: Given an item pair set X, $Interestingness(A \Rightarrow B)$ has lower and upper bounds $[-1,1]$.

Proof:

$$\begin{aligned} Interestingness(A \Rightarrow B) &= \frac{p(B \mid A) - p(B)}{p(B \mid A) + p(B)} \\ &= \frac{p(AB) - p(A)p(B)}{p(AB) + p(A)p(B)} \\ &\le \frac{p(AB)}{p(AB) + p(A)p(B)} \\ &\le 1 \end{aligned} \tag{5}$$

$$\begin{aligned} Interestingness(A \Rightarrow B) &= \frac{p(B \mid A) - p(B)}{p(B \mid A) + p(B)} \\ &= \frac{p(AB) - p(A)p(B)}{p(AB) + p(A)p(B)} \\ &\ge \frac{-p(A)p(B)}{p(AB) + p(A)p(B)} \\ &\ge -1 \end{aligned} \tag{6}$$

Property 3.2: Given an item pair set $X = \{A, B\}$ and $Interestingness(A \Rightarrow B)$.

If $Interestingness = 0$, A and B are independent of each other;

If $Interestingness > 0$, it means that A is positively correlated with B;

If $Interestingness < 0$, it means that A is negatively correlated with B;

Proof:

If A and B are independent of each other in the item pairs set $\{A, B\}$, then $p(AB) = p(A)p(B)$, then $Interestingness(A \Rightarrow B) = 0$.

If $Interestingness > 0$, $\frac{p(AB) - p(A)p(B)}{p(AB) + p(A)p(B)} > 0$, $p(AB) > p(A)p(B)$, it means that A is positively correlated with B.

If $Interestingness < 0$, $\frac{p(AB) - p(A)p(B)}{p(AB) + p(A)p(B)} < 0$, $p(AB) < p(A)p(B)$, it means that A is negatively correlated with B.

3.2 Superset interestingness upper bound

Property 3.3: There is an item set x, and the correlation measure:

$$\begin{aligned} & M(\sup(x), \sup(y), \sup(z)) \\ &= M(\sup(x), \sup(y), \sup(x \setminus y)) \\ &= correlation_{A,B} = \frac{\sup(AB)}{\sup(A)\sup(B)} \\ &= \sup(x) / [\sup(y) * \sup(x \setminus y)] \end{aligned} \tag{7}$$

where, $z = x \setminus y$.

The theorem about the correlation measure $M(\sup(x), \sup(y), \sup(z))$ is as following:

Theorem 1: When $\sup(y)$, $\sup(z)$ and the number of transaction data set n are constant, the correlation measure $M(\sup(x), \sup(y), \sup(z))$ and $\sup(x)$ are proportional to each other.

Theorem 2: When $\sup(x)$ and $\sup(z)$ (or $\sup(y)$) remain constant, the correlation measure $M(\sup(x), \sup(y), \sup(z))$ and $\sup(y)$ (or $\sup(z)$) are inverse.

For the above correlation measure theorems, the upper limit value of the correlation measure can be obtained. There is an item set x, and the superset x', $x \subseteq x'$, the upper limit value of the correlation measure of the superset x' of item set x is:

$$M(\sup(x), \sup(y), \max(\sup(\{i\}))) = \frac{1}{\max}(\sup(\{i\})) \tag{8}$$

Theory is as follows:

When $\sup(x') \le \sup(x)$, compared to other parameters, if the first parameter $\sup(x')$ is the largest in the measurement, the correlation measure M is the largest.

When $\sup(y)$ and $\sup(z)$ are the smallest, M is the largest, and $\sup(y)$ and $\sup(z)$ are both not less than $\sup(x')$, and $\sup(x') = \sup(x)$ at this time, then M is the largest when $\sup(y) = \sup(x)$.

When $\sup(x) = \sup(y)$, then $z = x' \setminus y = x' \setminus x$, where $\{j\} = x' \setminus x$. The items is sorted in ascending order of support, then $\sup(\{j\}) \ge \max(\sup(\{i\}))$, where $i \in x$. Therefore, when the third parameter of the correlation measure M is the smallest, that is $\max(\sup(\{i\}))$, M is the largest, so the maximum value is the upper bound of the item set correlation measure.

Property 3.4: The upper bound of the measure of the correlation of all item pairs to the superset of x is $\frac{1}{\max(\sup(\{i\}))}$. Further, the upper bound of the superset of each term is $\frac{1}{\sup(\{i\})}$.

3.3 Measurement of interestingness and relevance

Interestingness is defined as follows:

$$
\begin{aligned}
Interestingness\left(A \Rightarrow B\right) &= \frac{P\left(B \mid A\right) - P\left(B\right)}{P\left(B \mid A\right) + P\left(B\right)} \\
&= \frac{\dfrac{P\left(AB\right)}{P\left(A\right)P\left(B\right)} - 1}{\dfrac{P\left(AB\right)}{P\left(A\right)P\left(B\right)} + 1} \\
&= \frac{corr_{A,B} - 1}{corr_{A,B} + 1} \\
&= \frac{corr_{A,B} + 1 - 2}{corr_{A,B} + 1} \\
&= 1 - \frac{2}{corr_{A,B} + 1}
\end{aligned} \tag{9}
$$

It is obvious that the interestingness and the correlation are directly proportional to each other in the Eq. (9). Assumed that an item set x could be divided into two parts of y and z, then the upper bound of the measure of interestingness of the superset x' of the item set x is evaluated when the correlation is equal to maximum value, namely:

$$
Interestingness_{\max} = 1 - \frac{2}{M+1} = 1 - \frac{2}{\dfrac{1}{\max(\sup(\{i\}))} + 1} \tag{10}
$$

3.4 Usage of item pairs to supersets interestingness upper bound

The upper bound of interestingness can be used to prune the search space in the algorithm. When the upper bound of the interestingness of the item $\{i\}$ or the superset of the item x' is smaller than the threshold t, the search space of item $\{i\}$ or item set x may be pruned. The complexity of the algorithm will be reduced.

4 Item pairs mining algorithm based on interestingness

4.1 Main idea of the algorithm

Based on the redundancy condition limit for the upper bound of interestingness, the algorithm automatically generates the item pair pattern search traversal space and finds the non-redundant positive correlation terms. First, calculate the upper bound of the interestingness level of the superset of each item I_n and arrange from the largest to the smallest according to the upper bound value, and prune the items of the superset whose upper bound of interestingness is less than the threshold t, and reduce the space. Then the item set combination extension is performed, and item I_1 and item I_2 are combined into an item pair. First check whether the pair is redundant. If it is redundant and traverses the next pair. Otherwise, calculate the upper bound of the interestingness of the superset of the pair, observe whether it is greater than the threshold t, and if it is not greater, find the next pair of items. If satisfied, continue to calculate whether the two

pairs are positive correlation pairs. Then get the item pair combinations $\{I_1,I_3\}.....\{I_1,I_n\}$; $\{I_2,I_3\}.....\{I_2,I_n\}$; $.....\{I_{(n-1)},I_n\}$; Finally, all pairs of non-redundant positive correlations are mined.

4.2 Implementation of the algorithm

// i is all the items in the transaction data, $M(\sup(i),\sup(i),\sup(i))$ is the upper bound of interestingness of the superset corresponding to each item

for i in descending order on $M(\sup(i),\sup(i),\sup(i))$

if $M(\sup(i),\sup(i),\sup(i)) > t$ // t is the minimum threshold of interestingness, and the upper bound of interestingness of the item superset is greater than t

for k from $i+1$ to n

$x = \exp and(\{i\},\{k\})$; //extension pair

if $M(\sup(x),\sup(x),\max(\sup_{j\in x}(\{j\}))) > t$ //Whether the upper bound of interestingness of the superset of the item set is greater than t

check Redundancy(x); // Check the redundancy of item sets

if !redundancy // If non-redundant

Partition(x); // Calculate whether item set x is a positive correlation pair

if positive // Positive correlation

output // Output pair x

end if

else

break;

end if

end if

end for

else

break;

end if

end for

5 Experiment analysis

5.1 Time performance analysis

In order to analyze the performance of the algorithm, the algorithm is implemented in matlab and is runned the real data sets. The experimental environment includes: Intel(R) Core(TM) 2 Quad CPU, 2.00 GB memory, and Window 7.

The running time results of the algorithm are compared with the algorithm of Xiong et al.

[Xiong, Shekhar and Tan (2004)] on experimental data sets. The experimental data sets include T10I4D100K, T40I10D100K, and Kosarak, which are collected from the UCI website and are preprocessed. The characteristics of the data sets are shown in Tab. 2.

Table 2: Experimental data sets

Data sets	Record	Number of items
T10I4D100K	21452	451
T40I10D100K	40454	511
Accidents	13040	418
Kosarak	12137	190
Pumsb	22044	304

The running time of the algorithm is compared with the running time of Xiong et al. [Xiong, Shekhar and Tan (2004)] on the five data sets, T10I4D100K, Pumsb, Accidents, Kosarak, T40I10D100K. As shown in Figs. 1-5, the running time of the new algorithm and Xiong et al. [Xiong, Shekhar and Tan (2004)] algorithm decreases as the minimum interestingness threshold increases. However, the running time taken by the new algorithm relative to the previous algorithm is greatly reduced, and the time performance is significantly improved.

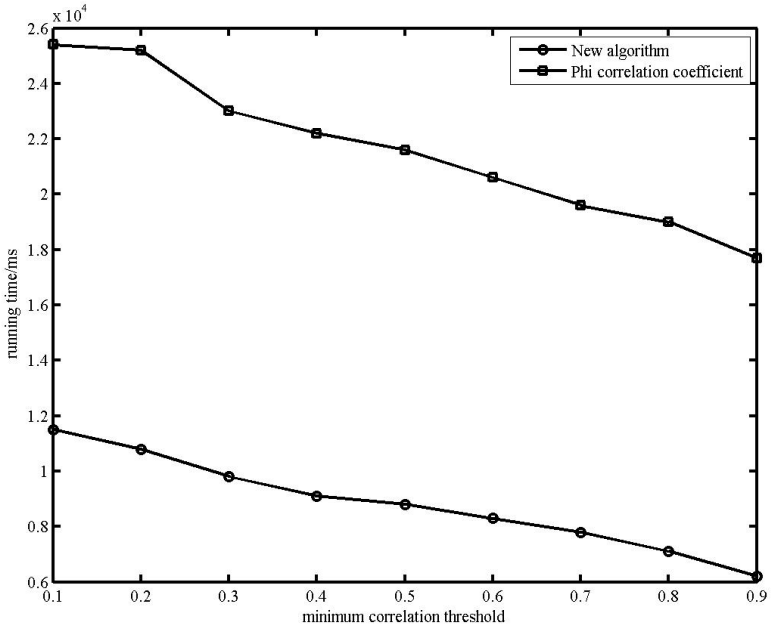

Figure 1: Running time on Accidents

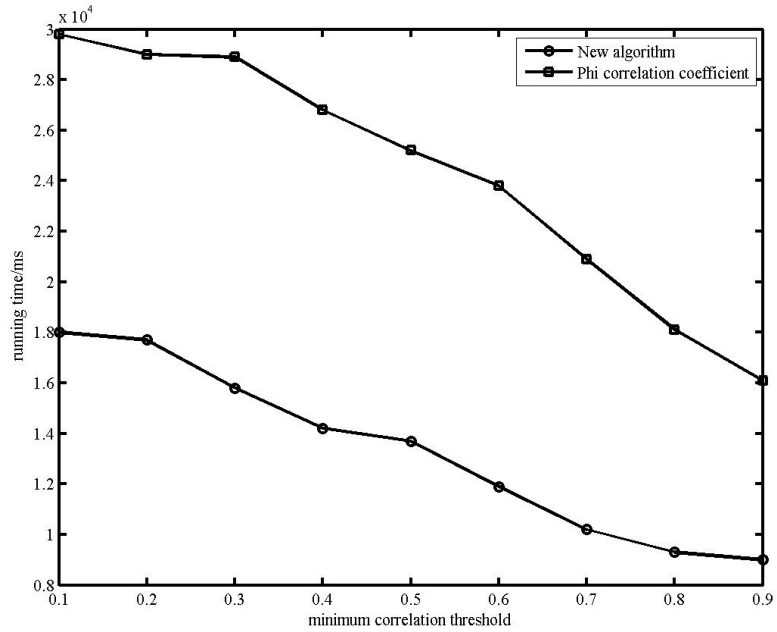

Figure 2: Running time on Pumsb

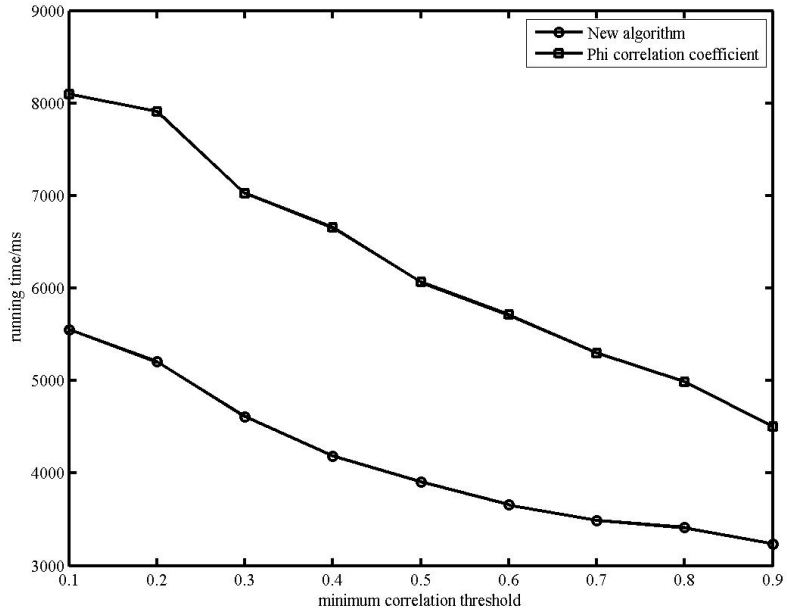

Figure 3: Running time on Kosarak

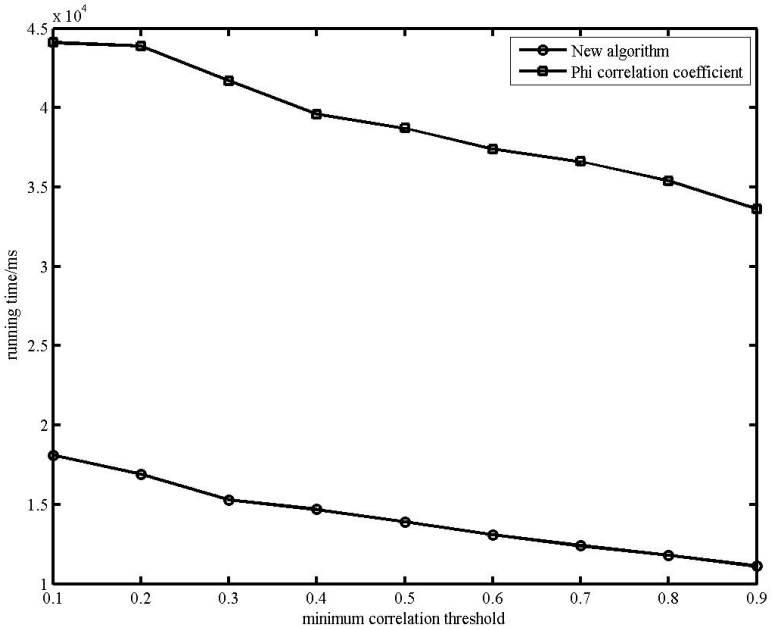

Figure 4: Running time on T10I4D100K

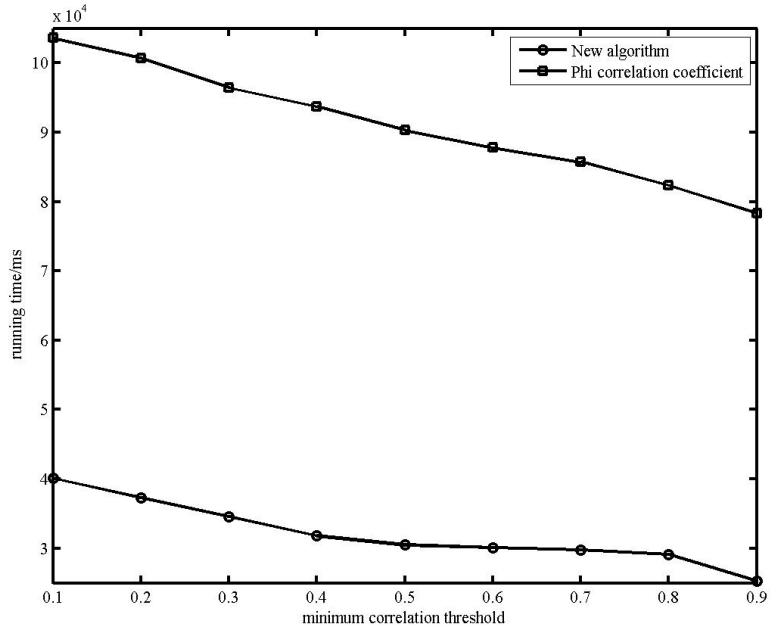

Figure 5: Running time on T40I10D100K

5.2 Pruning rate

Assume that n is the number of items, and $Pairs$ is the number of the item pairs after pruning in the algorithm, then the algorithm pruning rate can be expressed as follows,

$$prunerate = 1 - \frac{Pairs}{C_n^2} = 1 - \frac{Pairs}{n(n-1)/2} \tag{11}$$

As shown in Fig. 6, the pruning rate increase with the increasing of the minimum interestingness threshold on the different experiments.

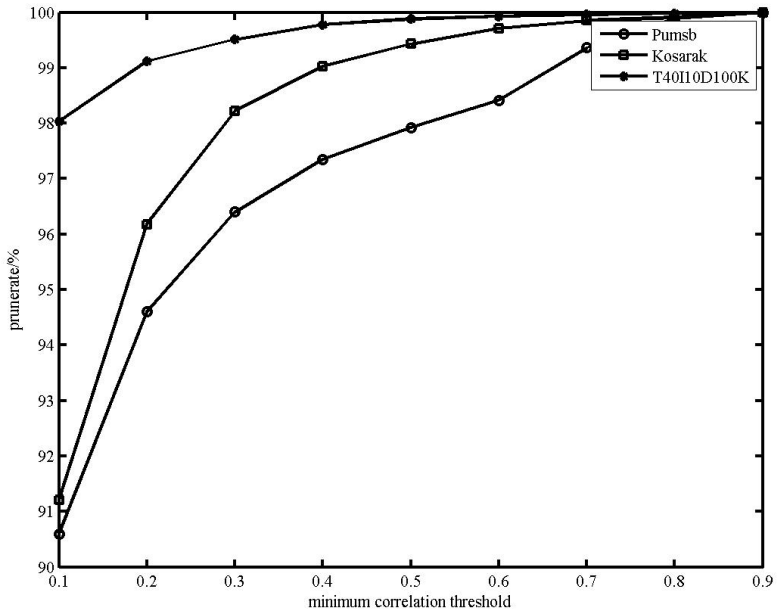

Figure 6: Pruning rate data set

5.3 Number of correlated item pairs

The algorithm proposed in this paper can not only prune the search space, but shorten greatly the running time compared with the Phi correlation coefficient algorithm in Xiong et al. [Xiong, Shekhar and Tan (2004)], the pruning efficiency increases with the increasing of the interestingness threshold. At the same time, the number of correlated item pairs, decreases with the increasing of threshold interestingness. And compared with proposed in the algorithm [Xiong, Shekhar and Tan (2004)], algorithm can retain most of the correlated item pairs, the redundant meaningless ones, it improves efficiency and accuracy of mining method. As shown in Figs. 7 and 8, there are differences between the two algorithms results in the Kosarak and T40I10D100K data sets. It is found that the result item pairs of our algorithm are less than the ones in Xiong et al. [Xiong, Shekhar and Tan (2004)] under the same interestingness threshold constraint. Because it can filter out redundant item pairs, and get the meaningful item pairs results with positive correlations.

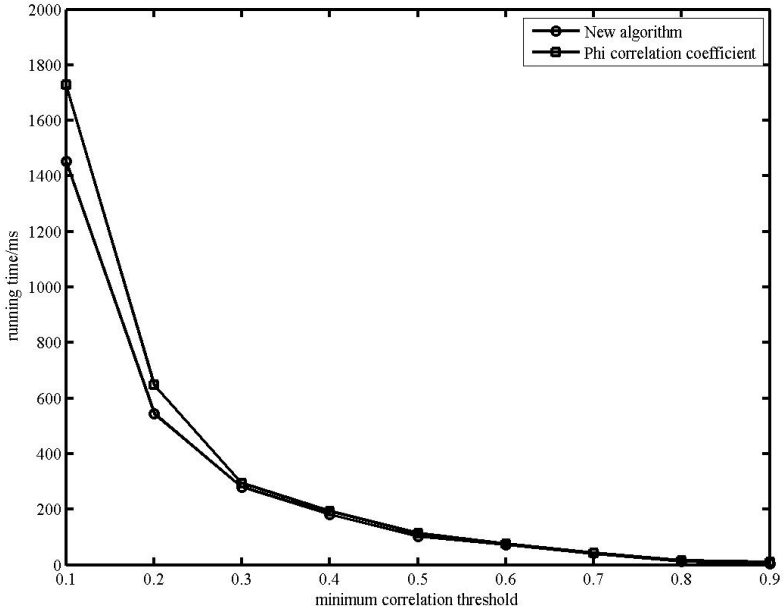

Figure 7: The number of item pairs on the Kosarak

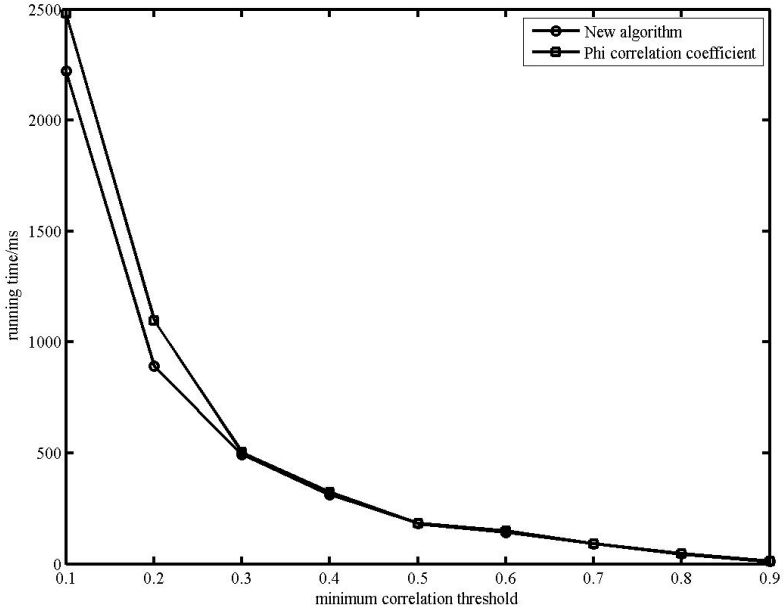

Figure 8: The number of item pairs on the T40I10D100K

5.4 Verification analysis

In the above experiments, the number of item pairs mined by the new algorithm is less than the ones of the Phi algorithm. But it is not completely certain that the pruned pair are meaningless. The correctness of the algorithm is verified by comparing the results on the real data sets of Kosarak and T40I10D100K respectively.

5.4.1 Verification based on Kosarak

There are 190 items in data set, and the items' number set is $\{1\},\{2\},\{3\}...\{190\}$. The experiments used 0.6, 0.7, 0.8, and 0.9 as the minimum thresholds, and the number of item pairs mined by two algorithms are shown in Tab. 3.

Table 3: Kosarak data set item pairs

Interestingness	New algorithm	Phi algorithm
0.6	35	37
0.7	17	18
0.8	5	5
0.9	2	2

On the data sets of Kosarak, when the interestingness threshold is 0.6, the algorithm mined 35 pairs of items, such as $\{64,32\},\{52,55\},\{101,97\},\{31,93\}$..... While the Phi algorithm finally dug out not only the above 35 pairs of items, but also two more pairs of items, i.e., $\{99,98\},\{96,91\}$. After analysis the original data sets, it is found that item 98 is subset of item 99 and item 96 is subset of item 91. So they are meaningless item pairs, and there are ignored by new algorithm. Then set the threshold to 0.7, the new algorithm mines $\{52,87\},\{99,55\}$..... and so on. The Phi algorithm has mined 18 item pairs, including the above 17 item pairs, and redundant item pair $\{96,91\}$. Finally, the minimum threshold of interestingness is 0.8, 0.9, the two algorithms mine the same result item pairs. When the threshold is 0.8, both sets of algorithms mine 5 item pairs. The five result item pairs are $\{52,55\},\{52,95\},\{52,87\},\{47,55\},\{99,96\}$. When the threshold is 0.9, the two algorithms also dig out two sets of result item pairs, namely $\{52,55\},\{47,55\}$.

5.4.2 Verification based on T40I10D100K

There are 511 items in the T40I10D100K data set, assuming the named T40I10D100K data set items are named $\{1\},\{2\},\{3\},...,\{511\}$. The interestingness threshold is 0.6, 0.7, 0.8, 0.9, and the number of item pairs excavated by the two algorithms at different minimum thresholds is shown in Tab. 4.

Table 4: T40I10D100K data set item pairs

Interestingness	New algorithm	Phi algorithm
0.6	81	88
0.7	35	38
0.8	12	12
0.9	3	3

When the threshold is 0.6, the new algorithm mines the item pairs $\{6,180\},\{193,212\},\{65,186\},...,$ which the total is 81. The Phi algorithm has a total of 88 item pairs, including 81 pairs of $\{6,180\},\{65,186\}$ and more than 7 item pairs such as $\{181,210\}$ and $\{97,183\}$. Bringing into the original data set, it is found that these seven item pairs have an ownership relationship with each other, and are meaningless item pairs, so they need to be pruned. When the threshold is 0.7, the new algorithm mines 35 item pairs such as $\{60,68\},\{181,43\},.....$ The Phi algorithm excavates the above 35 pairs of item, and there are more than three pairs of redundant item pairs of $\{183,182\},\{10,23\},\{184,136\}$. When the threshold is 0.8, 0.9, the new algorithm and the Phi algorithm mine the same result item pairs.

After verifying and analyzing the experimental results in detail, it is shown that the new algorithm can prune the redundant pairs of items more efficiently than the Phi algorithm. Therefore, the new algorithm can not only greatly shorten the running time, but also filter out the meaningless item pairs and improve the efficiency.

6 Conclusions

There are several shortcomings in the Phi correlation coefficient mining algorithm, such as, the time performance is not efficiency enough, the item pairs that are mined may be redundant and interestingness for users. A new interestingness model is proposed in this paper, which can use the superset of interestingness upper bound to prune the search space. Compared with the Phi correlation coefficient algorithm, the time performance is improved, and the meaningless item pairs are filtered according to the redundant constraints. Through experimental verification on the real data set, the mining efficiency and accuracy are indeed improved, i.e., the algorithm is feasible. The follow-up work will extend the algorithm to the mining of the entire frequent item sets.

Acknowledgement: This research was supported by the National Natural Science Foundation of China under Grant No. 61772280; by the China Special Fund for Meteorological Research in the Public Interest under Grant GYHY201306070; and by the Jiangsu Province Innovation and Entrepreneurship Training Program for College Students under Grant No. 201810300079X. Tao Li and Yongjun Ren are the co-corresponding authors.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

Funding Statement: This research was supported by the National Natural Science Foundation of China under Grant No.61772280; by the China Special Fund for Meteorological Research in the Public Interest under Grant GYHY201306070; and by the Jiangsu Province Innovation and Entrepreneurship Training Program for College Students under Grant No.201810300079X.

References

Chen, N. J.; Gao, Z. N. (2011): An improved positive and negative association rule mining algorithm. *Computer Science*, pp. 191-193.

Davale, A. A.; Shende, S. W. (2015): Implementation of coherent rule mining algorithm for association rule mining. *International Conference on Futuristic Trends on Computational Analysis and Knowledge Management*, pp. 538-541.

Feng, D. Y.; Zhu, L.; Zhang, L. (2016): Research on improved apriori algorithm based on MapReduce and HBase. *IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference*, pp. 887-891.

Jiang, H.; Luan, X.; Dong, X. (2012): Mining weighted negative association rules from infrequent itemsets based on multiple supports. *International Conference on Industrial Control and Electronics Engineering*, pp. 89-92.

Juan, W.; Li, S.; Feng, X. (2015): Extraction of non-redundant association rules from concept lattices based on IsoFCA system. *4th International Conference on Computer Science and Network Technology*, vol. 1, pp. 479-484.

Liu, B.; Wang, C. (2013): Association rule discovery based on formal concept analysis. *International Conference on Intelligent Computing Applications*, pp. 884-887.

Lu, J. L. (2012): Mining association rules based on correlation metrics. *Journal of Zhejiang University (Science Edition)*, pp. 284-288.

Luo, D.; Li, T. S. (2013): An improved apriori algorithm based on compression matrix. *Computer Science*, pp. 75-80.

Pandagale, A. A.; Surve, A. R. (2016): Hadoop-HBase for finding association rules using apriori MapReduce algorithm. *IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology*, pp. 795-798.

Poundekar, M.; Manekar, A. S.; Baghel, M.; Gupta, H. (2014): Mining strong valid association rule from frequent pattern and infrequent pattern based on min-max sinc constraints. *Fourth International Conference on Communication Systems and Network Technologies*, pp. 450-453.

Qian, G. C.; Jia, R. Y.; Zhang, R.; Li, L. S. (2008): One optimized method of apriori algorithm. *Computer Engineering*, vol. 34, no. 23, pp. 196-198.

Qian, T. Y.; Feng, X. N.; Wang, Y. Z. (2005): Over-support-the negative correlation of rules of confidence framework mining. *Computer Science*, pp. 124-127.

Quan, J.; Liu, Z.; Chen, D.; Zhao, H. (2012): High-efficiency algorithm for mining maximal frequent item sets based on matrix. *Fourth International Conference on Computational Intelligence and Communication Networks*, pp. 930-933.

Rameshkumar, K.; Sambath, M.; Ravi, S. (2013): Relevant association rule mining from medical dataset using new irrelevant rule elimination technique. *International Conference on Information Communication and Embedded Systems*, pp. 300-304.

Ravi, C.; Khare, N. (2014): EO-ARM: an efficient and optimized k-map based positive-negative association rule mining technique. *International Conference on Circuits, Power and Computing Technologies*, pp. 1723-1727.

Said, Z. B.; Guillet, F.; Richard, P.; Picarougne, F.; Blanchard, J. (2013): Visualisation of association rules based on a molecular representation. *17th International Conference on Information Visualization*, pp. 577-581.

Su, X. F.; Guo, Y. P. (2014): Research on measurement method of interest degree of negative association rules. *Agricultural Network Information*, pp. 76-79.

Tandon, D.; Haque, M. M.; Mande, S. S. (2016): Inferring intra-community microbial interaction patterns from metagenomic datasets using associative rule mining techniques. *PLoS One*, vol. 11, no. 4, e0154493.

Tang, X. W.; Xu, J.; Duan, B. J. (2018): A memory-efficient simulation method of grover's search algorithm. *Computers, Materials & Continua*, vol. 57, no. 2, pp. 307-319.

Tempaiboolkul, J. (2013): Mining rare association rules in a distributed environment using multiple minimum supports. *IEEE/ACIS 12th International Conference on Computer and Information Science*, pp. 295-299.

Thangarasu, S.; Sasikala, D. (2014): Extracting knowledge from XML document using tree-based association rules. *International Conference on Intelligent Computing Applications*, pp. 134-137.

Xiong, H.; Shekhar, S.; Tan, P. N. (2004): Exploiting a support-based upper bound of phi's correlation coefficient for identifying strongly correlated pairs. *Proceedings of the Tenth ACM SIGKDD Conference.*

Xu, T.; Dong, X. (2013): Mining frequent patterns with multiple minimum supports using basic apriori. *Ninth International Conference on Natural Computation*, pp. 957-961.

Xue, C.; Song, W.; Qin, L.; Dong, Q.; Wen, X. (2015): A mutual-information-based mining method for marine abnormal association rules. *Computers & Geosciences*, vol. 76, pp. 121-129.

Yang, J.; Huang, H.; Jin, X. (2017): Mining web access sequence with improved apriori algorithm. *IEEE International Conference on Computational Science and Engineering and IEEE International Conference on Embedded and Ubiquitous Computing*, vol. 1, pp. 780-784.

Yuan, L.; Li, D.; Chen, Y. (2016): Optimization and realization of parallel frequent item set mining algorithm. *International Conference on Audio, Language and Image Processing*, pp. 546-551.

Yue, Z.; Wang, L.; Wang, N. (2014): Efficient weighted association rule mining using lattice. *The 26th Chinese Control and Decision Conference*, pp. 4913-4917.