Searching of interesting itemsets for negative association rules

Hyeok Kong, Dokjun An, Douk Han
Faculty of Mathematics, Kim Il Sung University, Pyongyang, D. P. R. K.
hyeok_kong@yahoo.com
matadj79@yahoo.com
hdu0322@yahoo.com
June 21, 2018

Abstract: In this paper, we propose an algorithm of searching for both positive and negative itemsets of interest which should be given at the first stage for positive and negative association rules mining. Traditional association rule mining algorithms extract positive association rules based on frequent itemsets, for which the frequent itemsets, i.e. only positive itemsets of interest are searched. Further, there are useful itemsets among the frequent itemsets pruned from the traditional algorithms to reduce the search space, for mining of negative association rules. Therefore, the traditional algorithms have not come true to find negative itemsets needed in mining of negative association rules. Our new algorithm to search for both positive and negative itemsets of interest prepares preconditions for mining of all positive and negative association rules.

Keywords: positive and negative itemsets, negative association rules

1 Related works

We gave the definitions of positive and negative itemsets of interest in [1]. Apriori, a traditional association rule mining algorithm [4], has not come true to mine negative association rules because it pruned certain frequent itemsets among the candidate itemsets(Ck), not included in the prior family of frequent itemsets(Lk−1).

Also, the search space problem arised because of a great number of positive and negative itemsets, so it was important to identify only itemsets of interest needed for mining of association rules.

In [2, 3], rare association rule mining such as ¬A→¬C was studied, but it was only a kind of negative association rules, so they did not give clear definition and consider searching problem dealing with all possible cases of negative association rules.

The definitions of itemsets of interest are as follows. [1]

Definition 1. X→Y is called a positive association rule of interest, and X∪Y is called a positive itemset of interest, if they satisfy the following conditions.

1. X∩Y = ϕ
2. sprt(X∪Y) ≥ minsprt,
3. |sprt(X∪Y) − sprt(X)sprt(Y)| ≥ mininterest,
4. sprt(X∪Y)/sprt(X) ≥ minconf

Otherwise, if |sprt(X∪Y)−sprt(X)sprt(Y)| < mininterest, or sprt(X∪Y)/sprt(X) < minconf, then the rule X→Y is not of interest, and X∪Y is called an uninteresting itemset.

Conversely, if Q is a positive itemset of interest, there is at least one expression Q= X∪Y such that X and Y satisfy the above 4 conditions for positive association rules of interest.
**Definition 2.** The rule $A \rightarrow \neg B$ is called a **negative association rule of interest** and $A \cup B$ is called a **negative itemset of interest** if they satisfy the following conditions.

1. $A \cap B = \emptyset$,
2. $\text{sprt}(A) \geq \text{minsprt}$, $\text{sprt}(B) \geq \text{minsprt}$, $\text{sprt}(A \cup \neg B) \geq \text{minsprt}$,
3. $\text{sprt}(A \cup \neg B) - \text{sprt}(A)\text{sprt}(\neg B) \geq \text{mininterest}$,
4. $\text{sprt}(A \cup \neg B)/\text{sprt}(A) \geq \text{minconf}$.

Otherwise, the rule $A \rightarrow \neg B$ is not of interest, and $A \cup B$ is an **uninteresting itemset**. On the other hand, if $Q$ is a negative itemset of interest, there is at least one expression $Q=A \cup B$ such that one of the rules: $A \rightarrow \neg B$, or $\neg A \rightarrow B$, or $\neg A \rightarrow \neg B$, is a valid negative association rule of interest.

Thus, uninteresting itemsets are any itemsets in a database which exclude both positive and negative itemsets of interest. These itemsets need to be pruned to reduce the space searched in mining.

In other words, there are a large number of infrequent itemsets related to uninteresting association rules. If we could extract only positive and negative itemsets of interest among so many itemsets, the space searched would be extremely reduced.

## 2 Design of algorithm

For the purpose to develop an algorithm of mining both positive and negative association rules, we first consider data for the algorithm. Database $D$, minimum support $\text{minsprt}$, minimum confidence $\text{minconf}$ and minimum interest $\text{mininterest}$ are given in the algorithm. The resulting data of the algorithm are $PS$, a family of positive itemsets of interest, and $NS$, a family of negative ones.

During the running of the algorithm, $Freq_{k}$, a family of frequent itemsets must be generated at the $k$th pass of the algorithm, and based on them, the sets of positive and negative itemsets, namely $P_{k}$ and $N_{k}$, must be also generated respectively, where $P_{k}$ is the same as $L_{k}$, the family of frequent itemsets in the traditional algorithms of mining association rules.

Now, let $\text{Temp}_{k} = P_{k} \cup N_{k}$. Then, $C_{k}$, the candidate set in the traditional algorithms of mining association rules, is a subset of $\text{Temp}_{k}$, and every member of it must include at least one subset which is a member of $P_{k-1}$. Therefore, $\text{Temp}_{k}$ is a family of $k$-itemsets and each itemset of $\text{Temp}_{k}$ is an union of any two frequent itemsets in $Freq_{i}(1 \leq i \leq k-1)$. That is, for itemsets $A$ in $Freq_{i_{0}}$ and $B$ in $Freq_{i_{1}}(1 \leq i_{0}, i_{1} \leq k-1)$, if $A \cup B$ is a $k$-itemset, $A \cup B$ is appended into $\text{Temp}_{k}$. And each itemset in $\text{Temp}_{k}$ must be counted in the database $D$.

Next, if an itemset $Q = X \cup Y$ in $P_{k}$ satisfies $|\text{sprt}(X \cup Y) - \text{sprt}(X)\text{sprt}(Y)|<\text{mininterest}$ for any $X$ and $Y$, then $Q$ is the uninteresting frequent itemset, so it must be pruned from $P_{k}$. $P_{k}$ with all uninteresting itemsets pruned from it is appended to $PS$. Similarly, if an itemset $Q = X \cup Y$ in $N_{k}$ satisfies $|\text{sprt}(X \cup Y) - \text{sprt}(X)\text{sprt}(Y)|<\text{mininterest}$ for any $X$ and $Y$, then $Q$ is the uninteresting frequent itemset, so it must be pruned from $N_{k}$. $N_{k}$ with all uninteresting itemsets pruned from it is appended to $NS$.

The end conditions of the above loop are $P_{k} \neq \emptyset$ and $N_{k} \neq \emptyset$. $PS$ and $NS$ are output as results of the algorithm.

The algorithm is as follows.

**[Algorithm]** **Searching of Interesting Itemsets**

**Input data:** $D$(database), $\text{minsprt}$(minimum support), $\text{minconf}$(minimum confidence), $\text{mininterest}$(minimum interest)

**Output data:** $PS$(family of positive itemsets of interest), $NS$(family of negative itemsets of interest)

1. $PS=\emptyset$; $NS=\emptyset$;
2. $Freq_{1}$=family of 1-frequent itemsets; // First pass of $D$
   
   $k=1$;
   
   // Generation of all positive and negative $k$-itemsets of interest
3. do 
   
   
   
   
   k++;
   
   // Generation of possible k-itemsets
   Temp_k={A∪B| A∈Freq_i0, B∈Freq_i1, (1 ≤ i0, i1 ≤ k − 1), |A∪B|=k};
   
   // Counting of k-itemsets in transactions in D
   for ∀ t ∈ D do 
       
       Temt={k-itemset| k-itemset⊆ t, k-itemset∈Temp_k};
       
       for ∀ itemset ∈ Temt do
           itemset.count= itemset.count+1;
       
   // Generation of positive and negative k-itemsets based on k-candidate itemsets and k-frequent itemsets
   C_k={k-itemset| k-itemset∈Temp_k, ∃ Sub⊂k-itemset: Sub∈P_k−1};
   Freq_k={c| c∈C_k, sprt(c)=c.count/|D| >= minsprt};
   P_k = Freq_k; // k-positive itemsets
   NN_k = Temp_k - P_k
   // Pruning of uninteresting k-positive itemsets
   for ∀ itemset ∈ P_k do
       if itemset: uninteresting itemset then P_k=P_k - {itemset};
   
   PS=PS∪P_k;
   // Pruning of uninteresting k-negative itemsets
   N_k = {itemset ∈ NN_k, itemset: negative itemset}; // k-negative itemset
   for ∀ itemset ∈ N_k do
       if itemset: uninteresting itemset then N_k=N_k - {itemset};
   
   NS=NS∪N_k;
   } while (P_k−1 ≠ Φ, N_k−1 ≠ Φ);
   // Result output
   
   4. output PS, NS;

When uninteresting itemsets are pruned from the space searched of exponential size using the above algorithm, the space searched is extremely reduced.

3 Application of algorithm

The following example shows an application of the algorithm. A transaction database of 10 transactions is shown in Table 1. There are 6 items A, B, C, D, E and F. And, minsprt=0.3 and mininterest=0.07 are supposed.

Table 1. Example database
| Transaction ID | Items           |
|----------------|----------------|
| T1             | A, B, D        |
| T2             | A, B, C, D     |
| T3             | B, D           |
| T4             | B, C, D, E     |
| T5             | B, D           |
| T6             | B, D, F        |
| T7             | A, E, F        |
| T8             | C, F           |
| T9             | B, C, F        |
| T10            | A, B, C, D, F  |

Through the first scan of database, the support of each item is counted as follows.

A: 5/10 = 0.5
B: 7/10 = 0.7
C: 6/10 = 0.6
D: 6/10 = 0.6
E: 3/10 = 0.3
F: 5/10 = 0.5

From minsprt = 0.3, all items are 1-frequent itemsets. So $Freq_1 = \{A, B, C, D, E, F\}$. $Temp_2 = \{AB, AC, AD, AE, AF, BC, BD, BE, BF, CD, CE, CF, DE, DF, EF\}$ is generated from $Freq_1$, and $Freq_2 = P_2 = \{AB, AC, AD, BC, BD, BF, CD, CE, CF, DE, DF, EF\}$ is generated through Temt and $C_2$, based on minsprt = 0.3. So $NN_2 = Temp_2 - P_2 = \{AE, AF, BE, CE, DE, DF, EF\}$.

For pruning of uninteresting itemsets from $P_2$ on the basis of mininterest = 0.07, the interest measure of each itemset is calculated as follows.

$|sprt(A \cup B) - sprt(A)sprt(B)| = 0.05 < $mininterest,
$|sprt(A \cup C) - sprt(A)sprt(C)| = 0 < $mininterest,
$|sprt(A \cup D) - sprt(A)sprt(D)| = 0.11 < $mininterest,
$|sprt(B \cup C) - sprt(B)sprt(C)| = 0.02 < $mininterest,
$|sprt(B \cup D) - sprt(B)sprt(D)| = 0.02 < $mininterest,
$|sprt(B \cup E) - sprt(B)sprt(E)| = 0.11 < $mininterest,
$|sprt(B \cup F) - sprt(B)sprt(F)| = 0.05 < $mininterest,
$|sprt(C \cup D) - sprt(C)sprt(D)| = 0.02 < $mininterest,
$|sprt(C \cup F) - sprt(C)sprt(F)| = 0.05 < $mininterest,
$|sprt(D \cup E) - sprt(D)sprt(E)| = 0.02 < $mininterest,
$|sprt(D \cup F) - sprt(D)sprt(F)| = 0.11 < $mininterest,
$|sprt(E \cup F) - sprt(E)sprt(F)| = 0.05 < $mininterest.

As shown in the above calculations, only $BD$ is an itemset of interest and the rest of $P_2$ are all uninteresting itemsets. From this, pruning all uninteresting itemsets from $P_2$ before appending them to $PS$ gives $P_2 = \{BD\}$.

Similarly, the interest measure of each itemset in $NN_2$ is calculated as follows.

$|sprt(A \cup E) - sprt(A)sprt(E)| = 0.05 < $mininterest,
$|sprt(A \cup F) - sprt(A)sprt(F)| = 0.05 < $mininterest,
$|sprt(B \cup E) - sprt(B)sprt(E)| = 0.11 < $mininterest,
$|sprt(C \cup E) - sprt(C)sprt(E)| = 0.02 < $mininterest,
$|sprt(D \cup E) - sprt(D)sprt(E)| = 0.02 < $mininterest,
$|sprt(D \cup F) - sprt(D)sprt(F)| = 0.11 < $mininterest,
$|sprt(E \cup F) - sprt(E)sprt(F)| = 0.05 < $mininterest.

As shown in the above calculations, only $BE$ and $DF$ are itemsets of interest and the rest of $N_2$ are all uninteresting sets. From this, pruning all uninteresting itemsets from $N_2$ before appending them to $NS$ gives $N_2 = \{BE, DF\}$.

Next, $Temp_3 = \{ABC, ABD, ABE, ABF, ACD, ACE, ACF, ADE, AEF, BCD, BCE, BCF, BDE, BDF, BEF, CDE, CDF, CEF, DEF\}$ is generated from $Freq_1$ and Temt, and $Freq_3 = P_3 = \{ABD, BCD\}$ is generated through $Freq_2$ and $C_3$, based on minsprt = 0.3. So $NN_3 = Temp_3 - P_3 = \{ABC, ABE, ABF, ACD, ACE, ACF, ADE, AEF, BCE, BCF, BDE, BDF, BEF, CDE, CDF, CEF, DEF\}$.

For pruning of uninteresting itemsets from $P_3$ on the basis of mininterest = 0.07, the interest measure of each itemset is calculated as follows (for the sake of convenience, $A \cup B \cup C$ is represented as $ABC$, and $A \cup B$ is also represented as $AB$).
Similarly, the interest measure of each itemset in NN results in N. Next, Temp

From the above calculations, pruning all uninteresting itemsets from N

This shows that ABD and BCD are both itemsets of interest. So P₃= {ABD, BCD}. Similarly, the interest measure of each itemset in NN₃ is calculated as follows.

| sprt(ABD)−sprt(AB)sprt(D)| = 0.12 > mininterest,
| sprt(BCD)−sprt(B)sprt(CD)| = 0.09 > mininterest.

From the above calculations, pruning all uninteresting itemsets from N₃ before appending them to NS results in N₅ = {ABE, ADE, BDE, BDF, BEF, CDF, CEF}.

Next, Temp₄ = {ABCD, ABCF, ABDE, ABDF, BCDE, BCDF} is generated from Freq₁, Freq₃, and Freq₅, and Freq₄ = P₄ = ∅ is generated through Temp and C₄, based on minsprt = 0.3. So NN₄ = Temp₄ - P₄ = {ABCD, ABCF, ABDE, ABDF, BCDE, BCDF}.

For pruning of uninteresting itemsets from NN₄ on the basis of mininterest = 0.07, the interest measure of each itemset is calculated as follows.

| sprt(ABCD)−sprt(AB)sprt(CD)| = 0.11 > mininterest,
| sprt(ABDE)−sprt(AB)sprt(DE)| = 0.07 > mininterest.

From the above calculations, pruning all uninteresting itemsets from N₄ before appending them to NS yields N₅ = {ABCD, ABDE}.

Finally, Temp₅ = {ABCDEFG} is generated and Freq₅ = P₅ = ∅ based on minsprt = 0.3. So NN₅ = Temp₅ - P₅ = {ABCDEFG}.

In this algorithm, it is not considered whether confidences of the rules are greater than minconf or not, identifying the positive and the negative itemsets of interest. It means that each of positive and negative itemsets of interest should be dealt with together with the forth condition in Definitions 1 and 2. Dealing with this condition, the space searched would be more reduced.

Through the application of the algorithm, we can see that if a frequent itemset is pruned straight away when frequent itemsets of interest are searched, it can be found no longer in the frequent itemsets, but conversely it can be found in infrequent itemsets of interest. After then, if the itemset is pruned when infrequent itemsets of interest are searched, it is removed from the rest of the searched space. If an infrequent itemset is pruned when infrequent itemsets of interest are searched, it does not impact on searching for frequent itemsets of interest.

### 4 Conclusion

For mining both positive and negative association rules, not only frequent itemsets but also infrequent itemsets should be identified. Identifying frequent itemsets is exactly searching of the space of exponential size of possible items and itemsets, and the number of possible infrequent itemsets is much greater than the number of frequent itemsets. That is, the amount of possible positive and negative itemsets becomes almost double as much.

In this paper, we proposed an algorithm of searching for both positive and negative itemsets of interest that should be given at the first stage for positive and negative association rules mining. This algorithm prepares preconditions for mining of all positive and negative association rules. And we showed by an example that the space searched is extremely reduced dealing with itemsets of interest.
References

[1] J.Ri. H.Kong, D.An. Itemsets of interest for negative association rules. https://arxiv.org/abs/1806.07084, arXiv:1806.07084, 2018.

[2] U.Ryang. H.Kong, C.Jong. Rare association mining for network intrusion detection. IJTPC, 12(12):13–17, 2016.

[3] M.S.Abadeh. M.Almasi. Rare-pears: A new multi-objective evolutionary algorithm to mine rare and non-redundant quantitative association rules. Knowledge-Based Systems, 89:366–384, 2015.

[4] R.Srikant. R.Agrawal. Fast algorithms for mining association rules. In: Proceedings of International Conference on Very Large Data Bases, pages 187–499, 1994.