Discovery of Fuzzy Rare Association Rules from Large Transaction Databases

Weimin Ouyang

Shanghai University of Political Science and Law, Shanghai, China

Oywm @shupl.edu.cn

Keywords: Data mining; Association rules; Rare association rules; Fuzzy rare association rules

Abstract. Rare association rules is an association rule which has low support and high confidence. In recent years, the discovery of rare association rules has got quite a lot of attention, which has become a hot topic in data mining research. However, current discovery algorithms for rare association rules are built on the binary valued transaction databases, which cannot deal with quantitative attributes. In this paper, we put forward a discovery algorithm for finding fuzzy rare association rules to handle quantitative attributes. Experiments on the synthetic data stream show that the proposed algorithm is efficient and scalable.

Introduction

Rare association rules is an association rule [1, 2] which has low support and high confidence [3, 4]. In recent years, the discovery of rare association rules has got quite a lot of attention, which has become a hot topic in data mining research. However, current discovery algorithms for finding rare association rules are built on the binary valued transaction databases, which cannot deal with quantitative attributes. In this paper, we put forward a discovery algorithm for finding fuzzy rare association rules to handle quantitative attributes.

The paper is organized as follows. The related work is described in section 2. The definitions for fuzzy rare association rules in large transaction databases are given in section 3. In section 4, we describe the algorithm to discover fuzzy rare association rules from large transaction database, and Section 5 describes our experimental results. The conclusion is made in the last section.

Related Work

Recently, the discovery for rare association rules has attracted many researches. However, the current discovery algorithms for finding rare association rules are designed for the binary attributes databases. There are two different types of rare association rules mining approaches: level-wise and tree based. Current rare item sets mining approaches which are based on level-wise exploration of the search space are similar to the Apriority algorithm.

Rarity [3], ARIMA [4], MS-Apriority [5], Apriority-Inverse [6] and Affirm [7] are five algorithms which discover rare item sets. All of them utilize level-wise algorithm similar to Apriority, which contains expensive candidate generation step and pruning step. In addition, these algorithms attempt to identify all possible rare item sets, and require a significant amount of execution time. Tsang et al. [8] proposed a RP-Tree algorithm to handle these issues. RP-Tree avoids the expensive item set generation and pruning steps by using a tree data structure to find rare patterns. However RP-Tree algorithm still needs a multiple pass approach. Up until now, to our best knowledge, there has been no research on fuzzy rare pattern mining from large transaction database.

Problem Definitions

Let I = {i_1, i_2, i_m} be a set of items. A transaction T = (Tid, x_1， x_2， …， x_n), xi∈ I, for $1 \le i \le n$, is a subset of I, while n is called the size of the transaction, and Tid is the unique identifier of the transaction. A non-empty subset of I is called item set. An item set containing k items is called k-item set.

Definition 1: Given a transaction database, and a user-defined maximum support threshold maxis, a user-defined minimum support threshold mines, and is a user-defined minimum confidence c, the rule X->Y is a rare association rules, if and only if sup(X$\cup$Y) < maxis, sup(X$\cup$Y) $\geq$ mines and cone(X,Y) $\geq$c.

Most of current discovery algorithms for finding rare association rules focused on the binary attributes databases. Since transaction data in real-world applications usually has quantitative values, traditional association rules for the binary attributes databases should be extended to fuzzy association rules for quantitative attributes transactions.

Generally, quantitative values consist of numerical values and categorical values. To transform into transaction data format, the domain of quantitative and categorical attributes should be partitioned into disjoint intervals, each of which is considered as an attribute or item [9]. However, a sharp boundary problem results from the disjoint intervals. In order to solve this sharp boundary problem [10], fuzzy set theory has been introduced in the process of mining quantitative association rules, which results in a new category of association rules called fuzzy association rules [11].

For the fuzzy association rules, the support of an item set can be counted as follows: for every transaction in the transaction database, take the fuzzy logic AND of the membership values of the items under consideration, and summate these numbers. Let the transaction database be D and an item set X = {x_1, x_2, x $_3$, x_k} $\subseteq$ I. The support of a transaction t$\in$D to the item set X can be defined as

$$f\ \sup(X,t) = \bigcap_{i=1}^{k} \mu_{x_i}(t) \tag{1}$$

If we take the fuzzy logic AND as the product, the support of X from the transaction database D is defined as

$$f\ \sup(X) = \sum_{t\in D} f\ \sup(X,t) = \sum_{t\in D} \prod_{i=1}^{k} \mu_{x_i}(t) \tag{2}$$

The discovery algorithm for finding fuzzy association rules proposed in literature [9] first transforms each quantitative value into a fuzzy set with linguistic terms using membership functions. It then calculates the scalar cardinality of each linguistic term on all the transaction data, and compute the support of item set, carry an iterative search approach to find large item set. Each item uses only the linguistic term with the maximum cardinality in later mining processes, thus making the number of fuzzy regions to be processed the same as the number of original items. The discovery process based on fuzzy counts is then performed to find fuzzy association rules from these large item sets.

Definition 2: the fuzzy support of X from the transaction database D is then defined as

$$f\ \sup(X) = \max_{x\in X} \sum_{t\in D} \prod_{i=1}^{k} \mu_{x_i}(t) \tag{3}$$

Definition 3: Given a transaction database, and a user-defined maximum fuzzy support threshold maxis, a user-defined minimum fuzzy support threshold mines, and is a user-defined minimum confidence c, the rule X ->Y is a fuzzy rare association rule, if the following conditions hold:

(1) X$\cap$Y = $\varnothing$; (4)

(2) sup(X $\cup$Y) <maxis, sup(X $\cup$Y) >=mines; (5)

(3) Font(X,Y) =sup(X$\cup$Y) /sup(X) $\geq$c. (6)

Mining Fuzzy Rare Association Rules from Large Transaction Databases

Our discovery algorithm first transforms each quantitative value into a fuzzy set with linguistic terms using membership functions, then computes the scalar cardinality of each linguistic term on all the transaction data. Each item uses only the linguistic term with the maximum cardinality in later mining processes, thus making the number of fuzzy regions to be processed the same as the number of original items. The algorithm therefore focuses on the most important linguistic terms, which can decrease the time complexity of this algorithm. The discovery process based on fuzzy counts is then performed to find fuzzy association rules from these large item sets. Notations used in our algorithm are described as Table 1. The detail of the proposed discovery algorithm is described as follows.

Table 1 Notations

Notation	**meaning**
n	The total number of transactions in database
m	The total number of items
Di	Di is the ith transaction in D, 1≤ i ≤n
Ig	The gth item, 1≤ g ≤m
R^{gk}	The kth region of Ig, 1≤ k ≤ | Ig |, where | Ig | is the number of fuzzy regions for item Ig
v_i^g	The quantitative value of item Ig in Di
f_i^g	The fuzzy set converted from v_i^g
f_i^{gk}	The membership value of v_i^g in region Rgk
$count^{gk}$	The scalar cardinality of region Rgk
miscount	The maximum count value among countgk values
$maxR^g$	The fuzzy region of item Ig with miscount
C_k	the set of candidate itemsets with k items
R_k	the set of large fuzzy itemsets with k items
maxs	The predefined maximum weighted fuzzy support value
mins	The predefined minimum fuzzy support value
wminiconf	The predefined minimum fuzzy confidence threshold

Algorithm: MFRAR (Mining Fuzzy Rare Association rules)

Input: A transaction database D, each consists of customer ID, the purchased items with their quantities, a set of membership functions, maximum fuzzy support threshold maxis, minimum fuzzy support threshold mines, minimum fuzzy confidence threshold c;

Output: A set of fuzzy rare association rules: FRAR;

(1)For the transaction database D, n is the number of transactions in D, D_i is the itch transaction in D;

(2)Transform the quantitative value v_i^g of each item set I^g appearing in D_i into a fuzzy set f_i^g represented as $(\frac{f_i^{g1}}{R^{g1}}+\frac{f_i^{g2}}{R^{g2}}+\cdots+\frac{f_i^{gl}}{R^{gl}})$ using given membership functions, where R^{gk} is the kith fuzzy region of item I^g, f_i^{gk} is v_i^g's fuzzy membership value in region R^{gk}, and l is the number of fuzzy regions for I^g(k=1,2,3,…, l).

(3)Calculate the scalar cardinality of each attribute region R^{gk} as $count^{gk}=\sum_{i=1}^{n} f_i^{gk}$.

(4)Find miscount= $MAX_{k=1}^{l}(count^{gk})$, where 1≤ g ≤m and l is the number of regions for item I^g. Let $maxR^g$ be the region with miscount for item I^g. The $maxR^g$ will be used to represent the fuzzy characteristic of item I^g in later mining processes.

(5)Calculate the fuzzy support of a region $maxR^g$ (for g=1 to m) fsup ($maxR^g$) = $maxcount^g$. Check whether the fuzzy support of a region $maxR^g$, for g=1 to m, is larger than or equal to the predefined minimum weighted support threshold sup. If the value of $maxcount^g$ is equal to or greater than sup, put $maxR^g$ in the large 1-itemsets L_1. That is R_1= { $maxR^g$ | $maxcount^g$ ≥ sup, 1≤ g ≤m }.

(6)If $R_1$is null, then exit the algorithm; otherwise, do the next step.

```
(7)Set k=1, where k is used to represent the number of items kept in the current large itemsets.
(8)Generate the candidate set $C_{k+1}$ from $R_k$ in a way similar to that in the apriori algorithm.
(9)Do the following substeps for each newly formed (k+1)-item sets s with items ($s_1$, $s_2$, … , $s_{k+1}$) in $C_{k+1}$:
  (9.1) Calculate the fuzzy value $f_i^s$ for of s in each transaction Di as:
        $f_i^s = MAX_{i=1}^{n}(f_i^s)$, and $f_i^s = f_i^{s_1} \wedge f_i^{s_2} \wedge \cdots f_i^{S_{k+1}}$.
  (9.2) calculate the scalar cardinality of s as $count^s = \sum_{i=1}^{c} f_i^s$.
  (9.3) If the fuzzy support of s $count^s$ < maxs and $count^s$ >=mints Then put s in $R_{k+1}$;
(10)  If $R_{k+1}$is null, then do the next step; otherwise, set k=k+1 and repeat Step 8 to Step 10.
(11)  $R = \cup_k R_k$;
(12)  For each itemset i in R Do {
(13)      For any X∪Y=i and X ∩ Y = ∅ Do {
(14)          If fsup(X∪Y)<maxs ∧ fsup(X∪Y)>=mins ∧
                  Conf(X->Y)≧c
(15)          Then FRAR = FRAR ∪ { X->Y };
(16)          }
(17)      }
```

Experiment

In this section, we describe the results of our experiments of the algorithm MFRAR for mining fuzzy rare association rules in this section. Our computing environment is i5-6700, 8G RAM, Windows 7 operating system. We implemented the algorithm MRSP-SW with C++. The synthetic data set for experiment is generated by Assocgen [2] program of IBM Almaden research center. The meanings of used parameters are showed in Table 2.

TABLE 2 PARAMETERS

Symbol	Meaning
D	Number of customers(size of database)
C	Average number of transactions per Customer
T	Average number of items per Transaction
S	Average length of maximal potentially large Sequences
I	Average size of Items in maximal potentially large sequences
N_S	Number of maximal potentially large Sequences
N_I	Number of maximal potentially large Itemsets
N	Number of items

We set parameters C=10, T=5, S=4, I=2.5, NS =500, NI =2500, N =10000, total number of customers D=100000, and the generated database is named as C10T5S4I25. The membership functions used in our experiment is shown as Fig. 1.

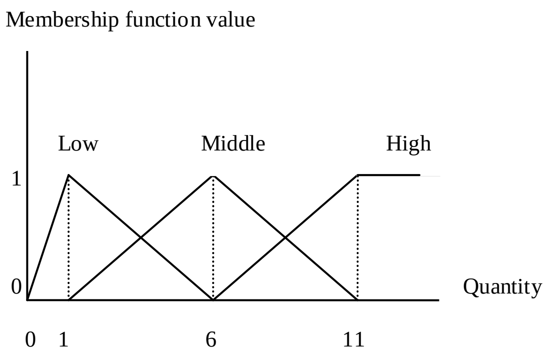

Figure 1. membership functions

Fig. 2 shows the algorithm executing time variance with minimum fuzzy support threshold mins decreasing from 1% to 0.2%. It demonstrates that the algorithm increases with the declining of mins.

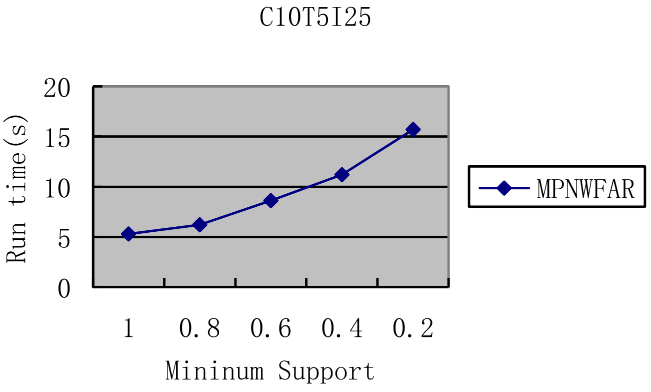

Figure 2. Execution times

To test the scalability of algorithm we increase the number of transactions from 50,000 to 150000, with mins=1%. The results are shown in Fig. 3. The executing time is increased almost linearly with the increasing of dataset size. It can be shown that our algorithm has a good scalable performance.

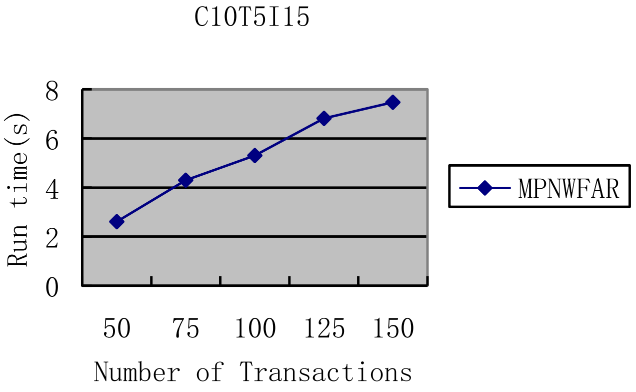

Figure 3. Scale-up: Number of transactions

Conclusions

In this paper, we made a fuzzy extension of crisp item to rare association rules in order to deal with quantitative items, and put forward a corresponding discovery algorithm for finding fuzzy rare association rules from large transaction databases. The experiments showed that the proposed algorithm is efficient and scalable.

References

[1] Agrawal R,Srikant R. Fast Algorithms for Mining Association Rules In the Proc. of the 20th International Conference on VLDB. Santiago, 1994. 487~499.

[2] Agrawal R,Srikant R. Mining association rules. In the Proc.1995 Int Conf. on Data Engineering, Taibei, Taiwan, March 1995.

[3] Szathmary L, Napoli A, Valtchev P, “Towards rare itemset mining”, In the Proc. of the 19th IEEE International Conference on Tools with Artificial Intelligence, vol. 1, pp. 305-312, 2007.

[4] Adda M, Wu L, Feng Y, “Rare itemset mining”, In the Proc. of the Sixth International Conference on Machine Learning and Applications, pp.73-80, 2007.

[5] Liu B, Hsu W, Ma Y, “Mining association rules with multiple minimum supports”, In: Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 337–341, 1999.

[6] Koh Y.S, Rountree N, “Finding Sporadic Rules Using Apriori-Inverse”, In the Proc. of PAKDD 2005, vol. 3518, pp.97-106, 2005.

[7] Torino L, Sibelius G, Barolo C, “A fast algorithm for mining rare itemsets”. In the Proc. of the Ninth International Conference on Intelligent Systems Design and Applications, pp.1149-1155, 2009.

[8] Tsang S, Koh Y.S, Dobbie G, “RP-Tree: Rare Pattern Tree Mining”, In the Proc. of DaWaK 2011, vol. 6862, pp. 277-288, 2011.

[9] Srikant R, Agrawal R. Mining Quantitative Association Rules in Large Relational Tables. Proc. of ACM SIGMOD Intl. Conference on Management of Data, 1996-06.

[10] A. Gyenesei, “A fuzzy approach for mining quantitative association. rules,” Technical Report of Turku Centre for Computer Science, no. 336,. March 2000.

[11]T. P. Hong, K. Y. Lin and S. L. Wang, “Mining fuzzy association rules from quantitative transactions", Soft Computing, Vol. 10, No. 10, pp. 925-932, 2006.