Granule Based Intertransaction Association Rule Mining

Wanzhong Yang, Yuefeng Li, Yue Xu
Faculty of Information Technology
Queensland University of Technology, Brisbane, QLD 4001, Australia
E-mail: W2.Yang@qut.edu.au

Abstract

Intertransaction association rule mining is used to discover patterns between different transactions. It breaks the scope of association rule mining on the same transaction. Currently the FITI algorithm is the state of the art in intertransaction association rule mining. However, the FTTI introduces many unneeded combinations of items because the set of extended items is much larger than the set of items. Thus, we propose an alternative approach of granule based intertransaction association rule mining, where a granule is a group of transactions that meet a certain constraint. The experimental results show that this approach is promising in real-world industry.

1. Introduction

According to requirements of knowledge discovered in real applications, association rule mining can be divided into intra association mining and inter association mining.

Intra association mining developed very well and has obtained many significant solutions, e.g. Apriori based approaches and FTP based approaches [1] [3]. The inter association mining tries to find associations between items in different transactions.

Lu et al. [5] first proposed the concept of inter association mining and contributed E-Apriori and EH-Apriori algorithms to this area. To improve the performance, Feng et al. [2] presented a template model for this problem.

Moreover, Tung et al. [7] recently proposed the FITI (First Intratransaction Then Intertransaction) algorithm. In FITI, if the average size of the transactions is very large, the extended transactions should be very long. It generates many extra combinations of items because the set of extended items is much larger than the set of items. Thus, this method is very slow if the average size of the transactions is large.

To reduce the size of discovered knowledge in large databases, Pawlak [6] and Li and Zhong [8] used decision tables for association mining. Li et al. [4] also proposed the concept of granule mining for intra association mining recently. Granule mining is a new initiative that tries to improve the quality of discovered knowledge in databases.

Unlike patterns that are parts of transactions, *granules* describe common features of sets of transactions for selected sets of attributes. They are more general than patterns and contain more structured information. Another advantage of using granules in multidimensional databases is that they can explicitly describe selected dimensions.

In this paper, we present a granule-based method for inter association mining in order to provide an efficient solution. Here, a granule is a group of transactions that meet a set of constraints. This method can take advantages of granule mining to simplify patterns into granules for inter association mining. To justify the mining results, we also present the concept of *precision* to evaluate the effectiveness of inter association mining in this paper.

2. Inter Association Mining

Let $T = \{t_1, t_2, \ldots, t_n\}$ be a transaction database, and each transaction is a set of items. Tung et al. [7] used the sliding window and extended-items to describe the intertransaction. Each sliding window W can be viewed as a continuous ω (a fixed interval called *maxspan*, or *sliding_window_length*) sub-windows such that each sub-window contains only one transaction. Let e_i be an item, its occurrences in different transactions in a sliding window can be extended from $e_i(0)$ to $e_i(\omega)$, where 0, …, ω are positions of transactions in the window. The transactions in a sliding window W can be merged into a megatransaction (or extended transaction) by putting all of W's extended items in a collection. Hence, an inter itemset refers to a set of extended-items, and an inter association rule can be

1082-3409/07 $25.00 © 2007 IEEE
DOI 10.1109/ICTAI.2007.143

represented as $X \rightarrow Y$, where X and Y are both a set of extended-items and $X \cap Y = \emptyset$.

The definition of the support and confidence in inter association mining follows up the intra association mining. Let N be the number of megatransactions and, X and Y both be a set of extended-items and $X \cap Y = \emptyset$. Let T_{xy} be the set of megatransactions that contains X and also Y, and T_x be the set of megatransactions that contains X. We have

$$sup(X \rightarrow Y) = |T_{xy}| / N, \quad conf(X \rightarrow Y) = |T_{xy}| / |T_x| .$$

3. Granule mining

Formally a transaction database can be described as an information table ($\mathcal{D}$, $V^{\mathcal{D}}$), where $\mathcal{D}$ is a set of objects in which each object is a sequences of items, and $V^{\mathcal{D}} = \{a_1, a_2, \ldots, a_n\}$ is a set of selected items (or called attributes) for all objects in $\mathcal{D}$.

Decision tables are efficient for dealing with multiple dimensional databases in line with user constraints. Formally, users may use some attributes of a database; and they can divide these attributes into two target groups: condition attributes and decision attributes, respectively. We call the tuple ($\mathcal{D}$, $V^{\mathcal{D}}$, C, D) a decision table of ($\mathcal{D}$, $V^{\mathcal{D}}$) if $C \cap D = \emptyset$ and $C \cup D \subseteq V^{\mathcal{D}}$.

We usually assume that there is a function for every attribute $a \in V^{\mathcal{D}}$ such that a: $\mathcal{D} \rightarrow V_a$, where V_a is the set of all values of a. We call V_a the domain of a. C (or D) determines a binary relation $I(C)$ (or $I(D)$)on $\mathcal{D}$ such that $(d1, d2) \in I(C)$ if and only if $a(d1) = a(d2)$ for every $a \in C$, where $a(d)$ denotes the value of attribute a for object $d \in \mathcal{D}$. It is easy to prove that $I(C)$ is an equivalence relation, and the family of all equivalence classes of $I(C)$, that is a partition determined by C, is denoted by $\mathcal{D} / I(C)$ or simply by $\mathcal{D}/C$. The classes in $\mathcal{D}/C$ (or $\mathcal{D}/D$) are referred to C-*granules* (or D-*granule*).

For example, in the share market, a transaction contains different shares at the same day. To reduce the risk of investments, share-market experts usually consider a group of shares rather one or two shares based on the current performance of another group of shares. To help such investments, we can group shares into different industry categories. For instance, we may choose two industries: bank and insurance.

The mining process has two sub stages.

(1) Transform the transaction database into the form of a decision table;

(2) Generate C-granules and D-granules based users selected two industry categories;

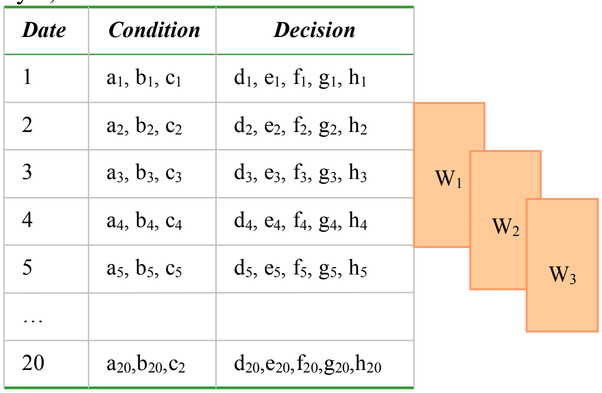
(3) Generate inter association rules between C-granules and D-granules.

The original transaction database records the data of ASX share transactions along the date dimension. The data includes attributes like *high*, *low*, *open* and *close*, which represent the price status in a day. To keep up the monotonic property, we assume the transactions are continuous and all records are complete filled. The empty records are instead of null value.

Since the mining object is transferred from the item to the group, a sliding window not only considers an interval (*sliding_window_length*), but also the number of attributes (we call *sliding_window_width*).

When transforming the transaction database to the decision table ($\mathcal{D}$, $V^{\mathcal{D}}$, C, D), let the banking shares be condition attributes C and the insurance shares be decision attributes D.

We can use the normal way for dealing with C-granules. We use the technique of sliding windows to generate D-granules, where *sliding_window_width* = $|D|$. Let $\mathcal{D}$ be all the transactions and *Va* refers to the profit gain of all shares in each transaction. *Va* includes three statuses: increased, neutral and loss, represented by 1, 0 and -1.



Date	*Condition*	*Decision*
1	a_1, b_1, c_1	d_1, e_1, f_1, g_1, h_1
2	a_2, b_2, c_2	d_2, e_2, f_2, g_2, h_2
3	a_3, b_3, c_3	d_3, e_3, f_3, g_3, h_3
4	a_4, b_4, c_4	d_4, e_4, f_4, g_4, h_4
5	a_5, b_5, c_5	d_5, e_5, f_5, g_5, h_5
…		
20	a_{20},b_{20},c_2	d_{20},e_{20},f_{20},g_{20},h_{20}

Figure 1. A decision table with sliding windows

In Figure 1 there are three bank shares a, b, c as condition attributes that represent *Westpac* bank, *ANZ* bank and *National* bank separately. Let a_i, b_i, c_i be the profit gain of bank shares on day i. The decision attributes d, e, f, g, h represent insurance shares *PMN*, *IAG*, *AMP*, *QBE*, *AXA*, where d_i, e_i, f_i, g_i, h_i refer to the profit gain of insurance shares on day i. The sliding windows only contains decision attributes, and the *sliding_window_width*=5 and *sliding_window_length*=3. The interval of the transactions decides the block of transactions in the sliding window, which would be used to generate D-granules for a same C-granule.

To describe the inter associations between condition granules and decision granules, we can extend the normal decision table into an extended decision table such that each condition granule is linked to all possible sub-windows in sliding windows. For example, Table 1 illustrates an extended decision table when we let *sliding_window_length* = 2.

ID	*Condition*	*Decision*
1	a_1,b_1,c_1	d_2,e_2,f_2,g_2,h_2
2	a_1,b_1,c_1	d_3,e_3,f_3,g_3,h_3
3	a_2,b_2,c_2	d_3,e_3,f_3,g_3,h_3
…		
39	a_{20},b_{20},c_{20}	$d_{21},e_{21},f_{21},g_{21},h_{21}$
40	a_{20},b_{20},c_{20}	$d_{22},e_{22},f_{22},g_{22},h_{22}$

Table 1. An extended decision table with *maxspan* = 2

The data compression is along the vertical direction in the extended decision table. Let $\mathcal{D}/C$ be the set of *C*-granules that refer to all classes of the profit situations for three bank shares. Let $\mathcal{D}/D$ be the set of *D*-granules that refer to all classes of the profit situations for five insurance shares. The inter association rule mining can be represented by mining granules now.

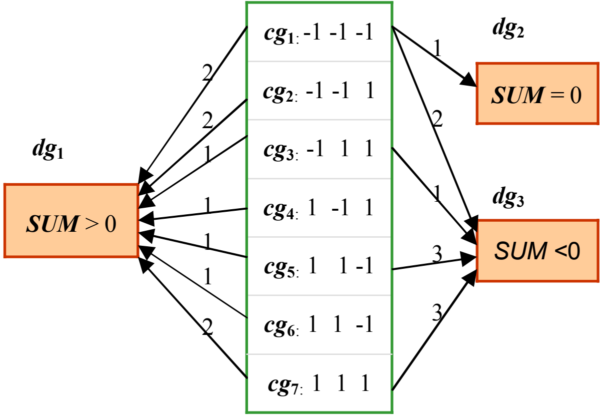


Figure 2. The association of *C* granules and *D* granules

It is hard to clearly understand the inter associations between condition granules and decision granules because of many duplicates. For this purpose we would like to represent the extended decision table as a 2-tier structure. The first tier contains all condition granules, the second tier contain decision granules and the inter associations are the links.

For the above example, people concern the gain of the group of shares, not only single share. Therefore, we can use a simple *SUM* measure to denote the gain information of a group of shares, where *SUM* > 0 means positive gain, *SUM* <0 means negative gain and *SUM* = 0 means no-gain.

Figure 2 depicts an example of a 2-tier structure, where we have seven condition granules that describe the possible changes of three bank shares; and have only three decision granules that describe the possible gains of buying five insurance shares after 1 or 2 days based on the changes of the three bank shares.

Formally, a set of items *X* is referred to as an *itemset* if $X \subseteq V^{\mathcal{D}}$. Let *X* be a itemset, we use **[***X***]** to denote the *covering set* of *X*, including all objects *d* such that $X \subseteq d$, i.e., $[X] = \{d \mid d \in \mathcal{D}, X \subseteq d\}$.

Let $\mathcal{D}/C = \{cg_1, cg_2, \ldots, cg_m\}$ and $\mathcal{D}/D = \{dg_1, dg_2, dg_3\}$. The decision rules in Figure 2 can be illustrated as follows:

$$cg_x \rightarrow dg_z$$
$$conf = |[cg_x \wedge dg_z]| \;/\; |cg_x|$$
$$support = |[cg_x \wedge dg_z]| \;/\; N$$

In Figure 2, there are twelve associations. If we set up the *min_sup* = 2, we have the following six inter association rules:

$cg_1 \rightarrow dg_1$ ($conf = 2/5$) $cg_1 \rightarrow dg_3$ ($conf = 2/5$)
$cg_2 \rightarrow dg_1$ ($conf = 2/2$) $cg_5 \rightarrow dg_3$ ($conf = 3/4$)
$cg_7 \rightarrow dg_1$ ($conf = 2/5$) $cg_7 \rightarrow dg_3$ ($conf = 3/5$)

5. Experiments

5.1. Basic experiments

In the ASX share market, there are 26 industries and almost 2000 companies. We take the ASX data of four industries from January 2005 to January 2007. We divide the data into two sections: a training set and a testing set. The first section contains over 260,000 transactions in 2005. The second section includes over 340,000 transactions in the other.

We choose two pairs of industries for the experiments: bank vs. insurance and food beverage & tobacco vs. retailing. In each pair, according to the yearly share volumes, we select the top three shares of one industry as condition granules and the top five products of another industry as decision granules.

ID	*B1*	*B2*	*B3*	*SUM>0*	*SUM=0*	*SUM<0*
1	-1	-1	-1	27	6	27
2	-1	-1	0	5	0	0
3	-1	-1	1	12	1	11
…						
15	1	1	1	33	5	29

Table 2. Bank vs. Insurance in 2005

Table 2 describes some samples for the first pair of industries in 2005 and the interval is one day. There are 15 condition granules. In the second pair of industries, there are 23 condition granules. The constraint-based decision granules are decided base on *SUM* > 0, *SUM* = 0 and *SUM* < 0. We choose three intervals for inter association mining. The intervals are one day, two days and three days.

5.2. Precision

When applying the inter association rule in the real data, we propose *Precision* as the criterion to evaluate the effectiveness of inter association rules.

In share market, investors should be interested in the prosperous shares where *SUM* ≥ 0. Let $cg_x \rightarrow dg_z$

be an inter association rule discovered in training phase and $SUM(dg_z) > 0$, a positive gain.

Let S_{fst} be the number of transactions in the testing set that match cg_x. Let S'_{snd} be the number of dg_z with $SUM(dg_z) \geq 0$ that match cg_x , and S_{snd} be the number of dg_z with $SUM(dg_z) > 0$ that match cg_x.

We define P_N as *Non_ Negative_ Precision* where

$$P_N(cg_x \rightarrow dg_z) = (S'_{snd} \; / \; S_{fst} \;) * 100\% \;.$$

We also define P_P as *Positive_ Precision* where

$$P_P\,(cg_x \rightarrow dg_z) = (S_{snd} \,/\, S_{fst}) * 100\%.$$

In Figure 3 the pair is bank and insurance. All *Non_Negative_Precisions* are between 60% and 100%. All *Positive_Precisions* are greater than 10%. When the interval is one day, the positive percentage reaches 60%.

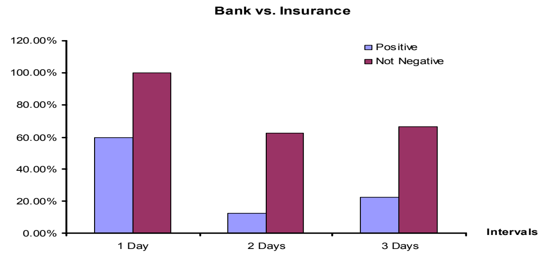


Figure 3. Precision for Bank vs. Insurance

5.3. Efficiency

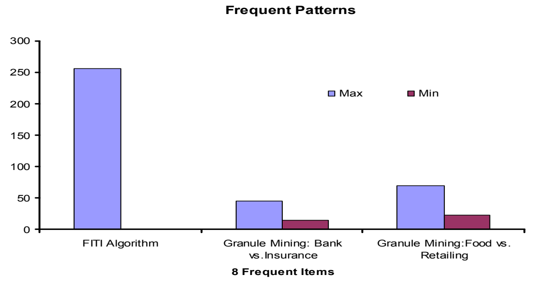


Figure 4. Frequent Patterns

Compared to the FITI algorithm, granule-based inter association mining makes long pattern mining possible and easier.In the FITI algorithm, the max frequent patterns of eight items in Figure 1 listed in Figure 4 $N_P = 2^8 = 256$. It expands the scope of the user requirement and generates many extra items. In the basic experiments, each pair includes eight different frequent items. In both pairs of industries, the minimum numbers of association rules are 15 and 23 separately; the maximum numbers of association rules are 45 and 69 separately. Our method obviously reduces the time and looks more efficient and applicable in the above example.

7. Conclusion

In this paper, we present granule based inter association mining to reduce the complexity of inter association mining. To compare with other methods, our method can reduce the width of sliding windows. It uses granules to replace extended item sets. Thus, we do not need to consider too many combinations of extended items. We also propose the concept of precision in order to evaluate the effectiveness of inter association mining. The experiments show that the proposed method is promising.

8. References

[1] Agraw, R., Imielinski, T., Swami, A., “Mining association rules between sets of items in large database”, *Proceedings of ACM-SIGMOD*, Montreal, Canada, 1993, pp. 207-216.

[2] Feng, L., Yu, J. X., Lu, H. and Han, J., “A template model for multidimensional inter-transactional association rules”, *The International Journal on Very Large Data Bases*, 11(2), pp. 153 -175, 2002.

[3] Han, J. and Kamber, M. *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2006.

[4] Li, Y., Yang, W. and Xu, Y., “Multi-Tier Granule Mining for Representations of Multidimensional Association Rules”, *6th IEEE International Conference on Data Mining* (ICDM 2006), Hong Kong, 2006, pp. 953-958.

[5] Lu, H., Han, J. and Feng, ., “Beyond intratransaction association analysis: mining multidimensional intertransaction association rules”, *ACM Transactions on Information Systems*, 18(4), pp.423 - 454, 2000.

[6] Pawlak, Z., “In pursuit of patterns in data reasoning from data, the rough set way”, *3rd International Conference on Rough Sets and Current Trends in Computing*, USA, 2002, pp. 1-12.

[7] Tung, A.K.H., Lu, H., Han, J. and Feng, L., “Efficient mining of intertransaction association rules”, IEEE Transactions on Knowledge and Data Engineering, 15(1), pp.43 – 56, 2003.

[8] Li, Y. and Zhong, N., “Interpretations of association rules by granular computing”, 3rd *IEEE International Conference on Data Mining* (ICDM 2003), USA, 2003, pp. 593-596.