Generalized Multidimensional Association Rules

ZHOU Aoying (周傲英), ZHOU Shuigeng (周水庚), JIN Wen (金 文)
and TIAN Zengping (田增平)

Department of Computer Science, Fudan University, Shanghai 200433, P.R. China

E-mail: {ayzhou,sgzhou,zptian}@fudan.edu.cn

Received December 14, 1998; revised October 22, 1999.

Abstract The problem of association rule mining has gained considerable prominence in the data mining community for its use as an important tool of knowledge discovery from large-scale databases. And there has been a spurt of research activities around this problem. Traditional association rule mining is limited to intra-transaction. Only recently the concept of N-dimensional inter-transaction association rule (NDITAR) was proposed by H.J. Lu. This paper modifies and extends Lu's definition of NDITAR based on the analysis of its limitations, and the generalized multidimensional association rule (GMDAR) is subsequently introduced, which is more general, flexible and reasonable than NDITAR.

Keywords multidimensional transaction database, data mining, N-dimensional inter-transaction association rules (NDITAR), generalized multidimensional association rules (GMDAR)

1 Introduction

One of the important problems in data mining[1] is discovering association rules from large-scale databases of transactions, where each transaction contains a set of items. First introduced by R. Agrawal *et al.* for market basket data analysis[2], association rules imply the association among items bought by the customers. For example,

R1: 90% of customers who bought bread also bought milk and jam.

Here the association among bread, milk, and jam is based on the result of statistic analysis rather than logic inference. Owing to its promising application perspective, association rule mining has become a top topic in the data mining area and attracted more and more research activities[3-6]. Association rule mining is now no longer limited to the transaction databases, it is also applied to relational databases, spatial databases and multimedia databases[7-9]. However, the semantics of traditional association rules introduced by R. Agrawal has not been changed, that is to say, association rules imply the co-occurrence of attribute items within the data records of transaction databases or any other types of databases. Therefore, such kind of association rule is intra-transactional.

Recently, H.J. Lu *et al.* proposed the N-dimensional inter-transaction association rule (NDITAR) while mining stock transaction data[10]. A typical 1-dimensional inter-transaction association rule is like

R2: If the prices of IBM and SUN go up, 80% of time Microsoft's will go up the next day.

Compared with $R1$, $R2$ is different in two aspects:

1) It implies association among items in different transaction records;
2) It deals with transaction records with time-dimensional attribute, i.e., the transactions are different records in the time dimension.

This work was supported by the National Natural Science Foundation of China, the NKBRSF of China, and the National Doctoral Subject Foundation of China.

Obviously, NDITAR is more general than the traditional association rules both semantically and formally, which makes the traditional association rule a specific case of NDITAR where there are no dimensional attributes with the transaction records. However, there are some limitations in the definition of NDITAR given by H.J. Lu. Thus, in this paper, we introduce the generalized multidimensional association rule (GMDAR) which is more general, flexible and reasonable than NDITAR.

The remainder of this paper is organized as follows. We first pinpoint the limitations of definition of NDITAR given by Lu in Section 2, then in Section 3 the generalized multidimensional association rule (GMDAR) is introduced, which modifies and extends the definition of NDITAR both semantically and formally. Section 4 is the conclusion remarks, meanwhile outlines the future research direction.

2 Problems with Lu's Definition of NDITAR

There are three major limitations in the definition of NDITAR introduced by Lu.

2.1 Only Quantitative Attributes Are Considered

In the definition of NDITAR given by Lu, for an arbitrary transaction record $T_i = (d_1, d_2, \ldots, d_n, E_i)$, $d_1, d_2, \ldots, d_n$ are all regarded as equi-interval values in the N-dimensional attribute space, such as the 1st day, the 2nd day, ..., in the time-dimensional attribute, and 1km, 2km, ..., in the distance dimensional attribute and so on. In such a case, the relationship between transactions can be represented by the relative differences in the values of dimensional attributes. However, the situation in the real world is not always so simple. For example, if we want to mine the commodity wholesale market data in order to predict the price trends, then there will be at least two dimensional attributes with each transaction record, maybe one is the trading *time*, and the other is the location of the wholesale market, say *city*. Thus there is a dimensional attribute pair (*city*, *time*) for each transaction record, where the *city* attribute is a categorical attribute that cannot be set to concrete quantitative values, over which arithmetic operations are carried out. There is another type of attribute, such as human's *age*, which is different from the *time* attribute even though it can also be represented by equi-interval integer, because the domain of *age* attribute is a finite set while *time*'s domain is an infinite ordered set. So while defining NDITAR, different types of dimensional attributes should be considered in order to make it more general semantically.

2.2 The NDITAR Is Formulated Too Rigidly

In [10], an NDITAR $X \Rightarrow Y$ (X, Y and $X \cup Y$ are frequent item-sets or event-sets) is strictly formulated by relative address *E-ADDR* ($X \cup Y$). Such a rigid relative address constraint on NDITAR will very possibly lead to the failure of mining some NDITARs which are available if a less strict constraint is imposed. For example, suppose the mining goal is $a(0) \Rightarrow b(1)$, $a(0) \Rightarrow b(2)$, and $a(0) \Rightarrow b(4)$. For a certain predefined support threshold, it is very likely that we cannot find frequent item-sets $\{a(0), b(1)\}$, $\{a(0), b(2)\}$ and $\{a(0), b(4)\}$. Nevertheless, if we loosen a little on the rule's formal constraint, for example, put the mining goal as $a(0) \Rightarrow b(n)$ ($n \leq 4$), then it is more possible that we can discover such a rule. Furthermore, if the mining goal is set to $a(n1) \Rightarrow b(n2)(\mid n1 - n2 \mid \leq 4)$, the possibility of success in mining this kind of association rule is greater than that of the previous two cases. On the other hand, relative address *E-ADDR* is not suitable for dealing with categorical attributes. Consequently, in order to enhance the practicability of mining NDITARs, a new and more flexible formal constraint of NDITAR must be adopted.

2.3 The *Support* and *Confidence* Definitions of NDITAR Are Not Reasonable Enough

According to the *support* and *confidence* definitions of NDITAR given in [10], the following consequences can be inferred:

1) *Support* is always $\leq$ *confidence*;
2) The implication of *support* for different frequent item-sets or event-sets is not consistent.

Here an example is given for explaining more clearly. Suppose there is a transaction database $TD = \{T_1(0,a), T_2(3,b), T_3(3,c)\}$. Based on Lu's definition, the support of $\{a(0), c(3)\}$ is 100% and the support of $\{a(0)\}$ is 33%. This is a very strange result. $\{a(0), c(3)\}$ and $\{a(0)\}$ both occur only once in the database TD, and the latter is a subset of the former, however their *supports* are quite different. It is obviously unreasonable. The cause lies in the definition of *support*. Furthermore, with such a *support* definition, the basis of *Apriori* algorithm no longer exists. Still, [10] applies an extended *Apriori* algorithm to mine one-dimensional inter-transaction stock data, which is unacceptable. So new definitions for *support* and *confidence* of NDITAR are necessary in order to eliminate the unreasonable aspects existing in the definition of NDITAR.

3 Generalized Multidimensional Association Rules (GMDAR)

Based on the analysis in the previous section of the problems with the definition of NDITAR introduced in [10], here we propose the generalized multidimensional association rule (GMDAR) for enhancing the generality, flexibility and reasonability of NDITAR. In fact, GMDAR is a modified and extended version of NDITAR with the following new features:

1) Taxonomies (*is-a* hierarchies) are introduced over the attributes of transactions;
2) Categorical attributes are considered in transaction records;
3) The *association constraint mode* is adopted to formulate association rules;
4) New *support* and *confidence* formulas are used.

Definition 1. *Let* $E = \{e_1, e_2, \ldots, e_u\}$ *be item set or event set,* $\{C_1, C_2, \ldots, C_k, D_1, D_2, \ldots, D_l\}$ $(l + k = N)$ *be attribute set, over which a taxonomies (is-a hierarchies) set* $\{H_i\}$ *is available. A transaction database consists of a series of transaction records* $(c_1, c_2, \ldots, c_k, d_1, d_2, \ldots, d_l, E_i)$, *where* $\forall m$ $(1 \leq m \leq k)$ $(c_m \in DOM(C_m))$, $\forall n$ $(1 \leq n \leq l)$ $(d_n \in DOM(D_n))$, $DOM(C_m)$ *and* $DOM(D_n)$ *are domains of* C_m *and* D_n *respectively, and* $E_i \in E$. C_m $(1 \leq m \leq k)$ *is categorical attribute (denoted as C-type attribute) and* D_n $(1 \leq n \leq l)$ *is quantitative value attribute with an unbounded domain (denoted as D-type attribute). We call a transaction database with* N *attributes* N*-dimensional transaction database, or multidimensional transaction database when* $N \geq 2$.

In Definition 1 we introduce two types of attributes. C-type is categorical attribute, D-type is quantitative value attribute with unbounded domain. Just as in [10], we take D-type attribute value as equi-interval value in its domain. So we also call this type of attribute infinite equi-interval value attribute. As for quantitative value attributes like human *age*, we still assign them into C-type attributes.

Then a question is whether it is permissible that there are only C-type attributes in the multidimensional transaction databases. As far as mining multidimensional inter-transaction association rules is concerned, we think it is unallowable for such a case. It is only due to the association existing among transaction attributes that the items or events in the transactions can be associated with each other. We can mine one-dimensional inter-transactional association rules from stock transaction data. It is only because the time attribute of stock transaction records is associative, just as yesterday is related to today, and today is followed by tomorrow. Conversely, the categorical attributes are generally not associative. Based

on such an assertion we suppose there is at least one D-type attribute in the transaction records when we deal with multidimensional inter-transaction association rules mining. As a matter of fact, when all dimensional attributes in the transaction databases are C-type attributes, what we can mine from the databases is the traditional association rules.

Definition 2. *For any transaction record T_i or event e_i in a transaction record, correspondingly there is a set of attribute values $(c_{1i}, c_{2i}, \ldots, c_{ki}, d_{1i}, d_{2i}, \ldots, d_{li})$, which we define as the status of the transaction record T_i or event e_i, and abbreviate it to s.*

Definition 3. *For any transaction set $\{T_1, T_2, \ldots, T_k\}$ or event set $\{e_1, e_2, \ldots, e_k\}$, there is a set of status $\{s_1, s_2, \ldots, s_k\}$, in which certain relationship must exist among all elements of the status set. This kind of relationship can be seen as a formal constraint on the transaction set or event set. We define such a relationship as status-constraint mode of the transaction set $\{T_1, T_2, \ldots, T_k\}$ or event set $\{e_1, e_2, \ldots, e_k\}$, and abbreviate it to SCM.*

Suppose there is a 2-dimension transaction database with attributes A_1 and A_2, which are C-type and D-type attributes respectively. Now there is an event set $e = \{a(\text{"}A\text{"},2), b(\text{"}B\text{"},4), c(\text{"}C\text{"},5)\}$. We give 3 SCMs as follows.

1) $a.A_1 = \text{"}A\text{"}, b.A_1 = \text{"}B\text{"}, c.A_1 = \text{"}C\text{"}, b.A_2 - a.A_2 = 2, c.A_2 - b.A_2 = 1$,

2) $a.A_1 = \text{"}A\text{"}, b.A_1 = \text{"}B\text{"}, c.A_1 = \text{"}C\text{"}, b.A_2 - a.A_2 \leq 3, c.A_2 - b.A_2 \leq 3$,

3) $\max(a.A_2, b.A_2, c.A_2) - \min(a.A_2, b.A_2, c.A_2) \leq 5$.

Obviously, the events set e conforms to all of the three SCMs above. However, the constraint strengths of the three SCMs are different. The advantage of SCM over E-$ADDR$ is its flexibility and the capability of coping with categorical attributes.

Definition 4. *Given an event-set $e = \{e_1, e_2, \ldots, e_k\}$ and its SCM, and a transaction set $T = \{T_1, T_2, \ldots, T_l\}$, if there is at least one subset T_c of T, and $T_c = \{T_{i1}, T_{i2}, \ldots, T_{ij}\}$ $(1 \leq ij \leq l)$ such that any transaction T_{ij} is got rid of from T_c, the following conditions will not be satisfied simultaneously.*

1) *For every event e_i in e, correspondingly there must be at least one transaction T_{ij} in T_c such that $e_i \in T_{ij}.E_{ij}$;*

2) *T_c conforms to SCM.*

Then we say T contains e in terms of SCM, and each distinct T_c stands for an occurrence of e in T. The total number of occurrences of e in T means the frequency of e occurring in T.

We give an example as follows.

Suppose there is a transaction database TD, which is one-dimensional and its dimensional attribute A is D-type, $TD = \{T_1(1,a,b), T_2(2,a,c), T_3(3,a,b,c), T_4(4,b,c), T_5(5,c)\}$, and event-set $e = \{(1,a),(2,b),(3,c)\}$ with an SCM which is defined as $b.A - a.A = 1$, $c.A - b.A = 1$. We can see that e is contained in T in terms of that SCM above, and there are two distinct subsets of T that contains e, which are $\{T_2, T_3, T_4\}$ and $\{T_3, T_4, T_5\}$ respectively. This indicates that the frequency of e occurring in T is 2.

Definition 5. *If a set of events $e = \{e_1, e_2, \ldots, e_k\}$ is associative in terms of status-constraint-mode SCM at certain pre-specified minimum frequency in the database, then we call this SCM association-constraint mode, and abbreviate it to ACM.*

Definition 6. *Suppose there is a multidimensional transaction database D_N, in which association rules are mined, and a taxonomies (is-a hierarchies) set $\{H_i\}$ exists among the attributes of the transactions in database. The minimum support and confidence are pre-specified as $sup_{\min}$ and $conf_{\min}$ respectively. Then a generalized multidimensional association rule $X \Rightarrow Y$ can be defined by a quintuple $GMDAR(X, Y, acm, sup, conf)$, where*

1) *X, Y and $X \cup Y$ are frequent event-sets conforming to association constraint mode acm and $X \cap Y = \emptyset$. No event in $X \cup Y$ is the ancestor of any other event that has similar status to the former event;*

2) *sup and $conf$ are support and confidence of $X \cup Y$, and*

$$sup = | X \cup Y | / | D_N | \geq sup_{\min} \tag{1}$$

$$cof = sup(X \cup Y)/sup(X) \geq cof_{\min} \quad (2)$$

Here, $| D_N |$ is the cardinality of multidimensional transaction database D_N, and $|X \cup Y|$ represents the total occurrence of $X \cup Y$ in D_N.

Clearly, our definitions of *support* and *confidence* are similar to those of the traditional association rules, which will avoid the problems coming with the *support* and *confidence* definitions of Lu's NDITAR. And the introduction of *association constraint mode* makes GMDAR mining more flexible and practicable. The user can choose freely an appropriate *ACM* to mine GMDARs. Based on Definition 6, a theorem about GMDAR can be obtained, which underlies the mining of GMDARs with Apriori algorithm.

Theorem. *If multidimensional event-set* $e = \{e_1, e_2, \ldots, e_k\}$ *that conforms to a certain ACM is a frequent event-set in multidimensional transaction database* D_N*, then any subset of* e *that conforms to the same ACM is also a frequent event-set in* D_N*.*

4 Conclusion and Future Work

N-dimensional inter-transaction association rule (NDITAR) is a generalization and extension of the traditional association rule. It is also a new challenge to the data mining community. This paper gives a more generalized association rule formula by modifying and extending the definition of NDITAR introduced by Lu[10]. We call the newly defined NDITAR generalized multidimensional association rule (GMDAR). GMDAR is more general, flexible and reasonable than NDITAR. Owing to the complexity of mining multidimensional association rules, developing efficient and effective algorithms to mine GMDAR is our future research direction.

References

[1] Chen M S, Han J, Yu P S. Data mining: An overview from database perspective. *IEEE Transactions on Knowledge and Data Engineering*, Dec. 1996, 8(6): 866–883.

[2] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In *Proc. the ACM SIGMOD Conference on Management of Data*, Washington, D. C., May 1993.

[3] Agrawal R, Srikant R. Fast algorithms for mining association rules. In *Proc. the 20th Int. Conference on Very Large Databases*, Santiago, Chile, Sept. 1994.

[4] Srikant R, Agrawal R. Mining generalized association rules. In *Proc. the 21st Int. Conference on Very Large Databases*, Zurich, Switzerland, Sept. 1995.

[5] Srikant R, Agrawal R. Mining quantitative association rules in large relational tables. In *Proc. the 1996 ACM SIGMOD International Conference on Management of Data*, Montreal, Canada, June 1996.

[6] Cheung D W, Ng V T, Fu A W, Fu Y. Efficient mining of association rules in distributed databases. *IEEE Transactions on Knowledge and Data Engineering*, Dec. 1996, 8(6): 911–922.

[7] Fu Y, Han J. Meta-rule-guided mining of association rules in relational databases. In *Proc. the 1st Int. Workshop on Integration of Knowledge Discovery with Deductive and Object-Oriented Databases, (KDOOD'95)*, Singapore, Dec. 1995, pp.39–46.

[8] Koperski K, Han J. Discovery of spatial association rules in geographic information databases. In *Advances in Spatial Databases.* In *Proc. the 4th Symposium, SSD'95.* (Aug.6–9, Portland, Maine). Springer-Verlag, Berlin, 1995, pp.47–66.

[9] Zaiane O R *et al.* Multimedia-miner: A system prototype for multimedia data mining. In *Proc. 1998 ACM-SIGMOD Conf. Management of Data*, Seattle, Washington, June 1998, pp.581–583.

[10] Lu H, Han J, Feng L. Stock movement and N-dimensional inter-transaction association rules. In *Proc. 1998 SIGMOD'98 Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'98)*, Seattle, Washington, June 1998.