# An Integrated Approach for Mining Meta-Rules[1]

Feiyue Ye[12],Jiandong Wang[1],Shiliang Wu[2],Huiping Chen[1],Tianqiang Huang[1],Li Tao[1]

[1]College of Information Science and Technology, Nanjing University of Aeronautics and Astronautics, Postal code 210016, Nanjing, China {cjsyes8@pub.cz.jsinfo.net}

[2]Department of Computer Science and Technology, Jiangsu Teachers College of Technology, Postal code 213001,Changzhou, China {cjsyes8@pub.cz.jsinfo.net}

**Abstract:** An integrated approach of mining association rules and meta-rules based on a hyper-structure is put forward. In this approach, time serial databases are partitioned according to time segments, and the total number of scanning database is only twice. In the first time, a set of 1-frequent itemsets and its projection database are formed at every partition. Then every projected database is scanned to construct a hyper-structure. Through mining the hyper-structure, various rules, for example, global association rules, meta-rules, stable association rules and trend rules etc. can be obtained. Compared with existing algorithms for mining association rule, our approach can mine and obtain more useful rules. Compared with existing algorithms for meta-mining or change mining, our approach has higher efficiency. The experimental results show that our approach is very promising.

## 1 Introduction

Mining association rules is one of the important issues of data mining, and the key of mining association rules is mining frequent patterns. Now Apriori[1] and its enhanced algorithm, FP-growth[2] and CT-ITL[3] algorithm are some important ones on frequent pattern mining in the world. Those algorithms aim at methods and efficiency on mining association rules, but they only fit for mining strong association rules with

[1]The work was supported in part by the fund of the Natural Science Plan from University in Jiangsu Province, China, Number: 04KJB460033

average support in total. However, the strength of some association rules may change over time. To mining the association rules at changing data sets, some incremental updating algorithm for mining association rule [4] are put forward, but these algorithms mine still rules which are average support and confidence more than or equal to appointed threshold at whole, so they can not use to mine rules which are change over time and predict.

Researchers have put forward some changing mining algorithms [5-9]. Ref. [5] is concerned with a basic framework for mining change from rule sets. To find whether a set of association rules discovered in a time period is applicable in other time periods was discussed in [7]. An approach to find a decision tree change between two time periods was proposed in [6]. Ref. [9] is concerned with meta-mining. The method of mining changes in association rules using fuzzy decision trees was put forward in [8]. In these algorithms, only the algorithms that were mentioned in [5][7][8][9] can be used to mining change of association rule. However, these algorithms only consider the mining change of association rule based on rule sets that have been mined by some of algorithms for mining association rule, so they do not consider the integrated efficiency for the whole mining process.

This paper presents an integrated approach of mining association rules and meta-rules based on a hyper-structure, this approach is evidently different from the above mentioned algorithm. With this approach, we can mine various association rules and meta-rules, for example, stable association rules, trend association rules, etc. In this paper, a classification approach, based on neural network, to classify association rule sets is discussed and the corresponding experiment is performed.

## 2 Constructing Hyper-Structure

In this section, Hyper-Structure is constructed. We first define the problem of 1-frequent itemset projected database.

**Definition 1** Let $I = \{i_1, i_2, \cdots, i_m\}$ be a set of all item in transaction database $D_0$ that there are $N$ transactions, $X_{k'}$ is the transaction itemsets of the $k'$ -th transaction, $X_{k'} \subseteq I$ ,i.e., $X_{k'} = \{i_{k_1}, i_{k_2}, \cdots, i_{k_n}\}$ ,where $1 \le j < n, 1 \le k_j < k_{j+1} \le m$ . $X_{(n)}$ denotes that the set $X$ contains $n$ items. The number of transactions in $D_0$ containing itemset $X$ is called the support of $X$ ,denoted as $sup_0(X)$ .Given a minimum support threshold $s$,

if $sup_0(X) \geq s$ ,then $X$ is frequent in $D_0$ .Let $i'_{p_{j'}}$ be a frequent item in $D_0$ , called as 1-frequent item, where $1 \leq p_{j'} \leq m$ ,and Let $I' = \{i'_{p_1}, i'_{p_2}, \cdots, i'_{p_{m'}}\}$ be the set of all 1-frequent item, where $1 \leq m' \leq m$ and $I' \subseteq I$ .Thus the projection between $I'$ and $X_{k'}$ is $A_{k'}$ , and $A_{k'} = I' \cup X_i = \{i^{k'}_{q_1}, i^{k'}_{q_2}, \cdots, i^{k'}_{q_{n'}}\}$ , and then the transaction database that consists of $A_1, A_2, \cdots, A_N$ is called 1-frequent itemset projected database $A$.

### 3.1 Structure of Hyper-Structure Head Table

The hyper-structure head table contains two fields: item number field and pointer field. The pointer in pointer field points to a hash chain structure with the same number of items. The hyper-structure head table is created dynamically. The hyper-structure is illustrated in Fig.1.

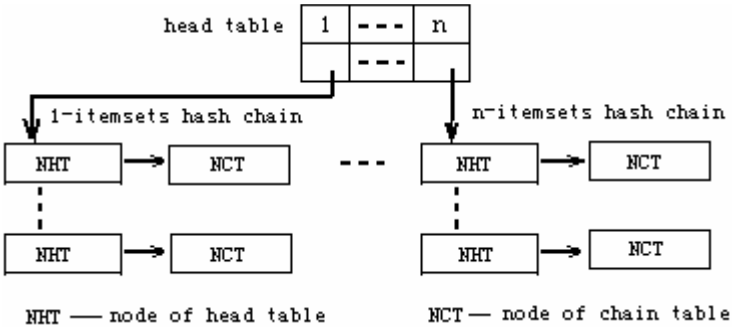


**Fig.1**. Hyper-structure

### 3.2 Chain Address Function

The hash function of the item $i_{k_j}$ (where $k_j$ is the item number) in 1-frequent itemsets is given below:

$$h(k_j) = k_j \qquad (1)$$

Let $B = \{q_1, q_2, \cdots, q_{n'}\}$ be a set of the item number in itemset $A_{k'} = \{i^{k'}_{q_1}, i^{k'}_{q_2}, \cdots, i^{k'}_{q_{n'}}\}$. The hash function of the multi-item itemsets is given below:

$$h(q_1, q_2, \cdots, q_{n'}) = (\sum_{i=q_1}^{q_{n'}} (2i+1) z_i) \bmod p' \qquad (2)$$

If $i \in B$ , then $z_i = i$ ,otherwise $z_i = 0$ ; $p'$ is the sum of the adjusted pattern of the multi-itemsets.

### 3.3 Chain Address Structure

The node structure of the heat table and the chain table of 1-frequent itemsets is illustrated in Fig.2 and Fig.3 respectively.

The chain address is produced according to formula (1) in Fig.2. The pointer points to the node structure of chain table.

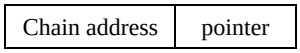


**Fig. 2.**The node structure of head table

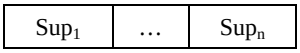


**Fig. 3.** The node structure of chain table

The node structure of head table of multi-itemsets is shown in Fig.4 and the node structure of the chain table of the multi-itemsets is shown in Fig.5:

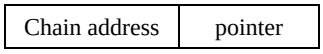


**Fig. 4** The node structure of head table

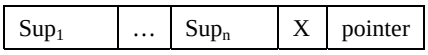


**Fig 5.** The node structure of chain table

The chain address is obtained from formula (2) in Fig.4 ,Here, "pointer" points to its chain table node; " $Sup_i$ " is used to record the count of itemsets $X$ which appears in $D_i$ ;The pointer points to the next chain table node in Fig.5.

### 3.4 The Algorithm of Constructing Hyper-Structure

**Algorithm 1:** the algorithm of constructing the hyper-structure

**Input:** A transaction database $D_0$

**Output:** Hyper-structure

**Method:**

1. scan transaction database $D_0$ to obtain the set of 1-frequent item and its projected database and maximal projected item $Max(|X|)$, and partition the projected database $A$ of $D_0$ into $D_1^A, D_2^A, \cdots, D_n^A$ according to time period $t_1, t_2, \cdots, t_n$ ;
2. construct the head table of hyper-structure from 1-item to $Max(|X|)$ ;
3. construct 1-frequent item hash chain;
4. *for* (i=1; n;i++) {

j=1;

*do while* the scan for projected database $D_i^A$ is unfinished

{*forall* join the item in item number set $B$ of $A_j$ to generate itemset $X$ do {

calculate value of $h(q_1, q_2, \cdots, q_{n'})$ for $X$ ;

locate address of $h(q_1, q_2, \cdots, q_{n'})$ in $|X|$ -item hash chain;

*if* there has not been $X$ in corresponding address chain table in $|X|$ -item hash chain according to the value of $h(q_1, q_2, \cdots, q_{n'})$ *then*
{save $X$ to there and $sup_i = 1$ ;}
*else* { $sup_i = sup_i + 1$ ;}}
j=j+1;}

## 4 Association Rule and Meta-Rule Mining

The association rule mining problem can be decomposed into two subproblems (Agrawal et al., 1993), That is, Finding all frequent itemsets and using the frequent itemsets to generate the desired rules. The meta-rules mining problem is to find the change of the association rule over time. So the main mission of mining association rules and meta-rule is mining frequent itemsets. Having got a frequent itemset, we can further obtain all its subsets, and form the corresponding rule set. To mine the association rules and meta-rules from the hyper-structure, the association rule and the meta rule are defined below:

**Definition 2.** An association rule *r* is denoted by " $X_1 \Rightarrow X_2$ with caveat *c* ", where $X_1 \subset X$ and $X_2 \subset X$ ,and $X_1 \cap X_2 = \phi$ , *c* is support and confidence.

**Definition 3.** Let *min_sup* be a minimum support threshold, *min_conf* be a minimum confidence threshold, if rule *r* satisfies both $sup\,port(X_1 \Rightarrow X_2) \geq min\_sup$ and $confidence(X_1 \Rightarrow X_2) \geq min\_conf$ , then *r* is called strong rule. $R$ denotes the set of rule *r*.

**Lemma 1:** if itemsets $X$ is global frequent, then it is frequent at least in one segment $D_i$ （$1 \leq i \leq n$）.

When mining the meta-rule, we consider only the rules which are global frequent and are frequent in one time segment, According to **Lemma 1** the meta-rule is defined in the following:

**Definition 4**. Let the support and confidence of rule $X_1 \Rightarrow X_2$ from datasets $D_0, D_1, D_2, \cdots, D_n$ be $sup_0, sup_1, sup_2, \cdots, sup_n$ and $conf_0, conf_1, conf_2, \cdots, conf_n$ respectively, if $sup_0 \geq min\_sup$ and $conf_0 \geq min\_conf$ ,then the support meta-rule $m_s$ from $D_1, D_2, \cdots, D_n$ is given below:

$$X_1 \Rightarrow X_2 : \{\sup_1, \sup_2, \cdots, \sup_n\},$$

And the confidence meta-rule $m_c$ from $D_1, D_2, \cdots, D_n$ is given below:

$$X_1 \Rightarrow X_2 : \{conf_1, conf_2, \cdots, conf_n\}.$$

**Lemma 2:** all subsets of the frequent itemsets that appear in *n*-itemsets hash chain must also appear in corresponding 1-itemsets to (*n*-1)-itemsets hash chain.

**Apriori property:** all nonempty subsets of a frequent itemset must also be frequent.

According to above definition and lemma and property, the algorithm for mining association rules and meta-rules is given in the following:

**Algorithm 2:** the algorithm for mining association rules and meta-rules

**Input:** hyper-structure; dataset $D_0$ and dataset $D_1, D_2, \cdots, D_n$ ; support threshold $min\_sup_i$ $(i = 0,1,\cdots,n)$ and confidence threshold $min\_conf_i$ $(i = 0,1,\cdots,n)$

**Output:** association rules and meta-rules

**Method:**

Association rule set $R = \phi$ ;support meta-rule set $M_s = \phi$ ;confidence meta-rule set $M_c = \phi$ ;

**For** (i=2;n;i++) {

The pointer points to the first address of *i*-items hash chain;

*Do while* to search *i*-item hash chain is unfinished {

Search the chain table node at the address and obtain itemset $X$ and $sup_1, sup_2, \cdots, sup_n$ and $sup_0$ **;**

*If* $sup_0 \geq min\_sup_0$ *then* {

*If* i>2 *then* {

*Forall* the subsets $X_{(i-1)}$ and $X_{(i-2)}$ of the itemset $X$ *do* {

*Forall* the itemsets $X_{(i-1)}$ , $X_{(i-2)}$ , $X_{(i-2)}$ and $X_{(i-2)}$ to satisfy definition 2 are regarded as antecedent and consequent for forming association rules do {

Obtain the $sup_j (j = 0,1,\cdots,n)$ of antecedent and consequent of association rule respectively;

Calculate $conf_j (j = 0,1,\cdots,n)$ according to $\sup_j (j = 0,1,\cdots,n)$ ;

**If** $conf_0 \geq \min\_conf$ **then** add the rule to $R$ and add support meta-rule $m_s : (sup_1, sup_2, \cdots, sup_n)$ to set $M_s$ and add confidence meta-rule $m_c : (sup_1, sup_2, \cdots, sup_n)$ to set $M_c$ ;}}}

**else** {

search and obtain the corresponding $sup_j (j = 0,1,\cdots,n)$ of two item of $X$ ;

Calculate $conf_j(j=0,1,\cdots,n)$ according to $sup_j(j=0,1,\cdots,n)$;

**If** $conf_0 \geq \min\_conf$ **then** add the rule to rule sets and add support meta-rule $r:(sup_1,sup_2,\cdots,sup_n)$ to meta-rule set $M_s$ and add confidence meta-rule $r:(sup_1,sup_2,\cdots,sup_n)$ to meta-rule set $M_c$;}}

**else** {stop mining}}

## 5 Analysis of the Change Trend of Association Rules Using Meta-Rules

The types of change trend of association rules can be divided into several cases in the following:

**Stable rules:** These rules do not change a great deal over time. Stable rules are more reliable and can be trusted.

**Trend rules:** These rules indicate some underlying systematic trends.

**Irregular or random movements:** These characterize the sporadic motion of time series due to random or chance events.

**Cyclic movement or cyclic variation:** These refer to the cycles, that is, the long-term oscillations about a trend line or curve, which may or may not be periodic.

**Seasonal movements or seasonal variations:** These movements are due to events that recur annually.

In this paper, only forefront three types are considered.

Apparently, if the meta-rules are directly analyzed using the trend analysis method, each rule will be scanned and calculated repeatedly. It is necessary to first classify the meta-rules in order to improve the analysis efficiency. That is, the meta-rules are classified into several classes which donate a change trend of association rules. For example, the association degree is stable or decreased or increased over time, etc. Then each class meta-rules are respectively analyzed on demand. There are many existing classification methods, for example, the SVM method, neural network method, C4.5 algorithm, Bayesian classification, etc. Next we are going to discuss classification method for meta-rules using BP neural network.

Before using neural network for classification, some training samples need to be obtained. These samples should be formed according to the following cases:

1) The association degree of rule sets is stable movement over time, its output is $y_1$.

2) The association degree of rule sets is increased over time, its output is $y_2$ .
3) The association degree of rule sets is decreased over time, its output is $y_3$ .
4) The association degree of rule sets is random movements over time, its output is $y_4$ .

First, the BP network is trained by sample data sets to obtain its weight, and then trained BP network is applied to corresponding classification work. In the process of mining the meta-rule, the meta-rule which has already mined from hyper-structures will be imported into the BP network defined before so that the classified association rules can be obtained.

Generally speaking, after classification the usability of association rules is improved considerably, and they can be further analyzed expediently.

## 6 Experimental Result

In our experiments，the dataset is from a supermarket transaction database from October 1 in 1996 to May 31 in 1997, its original size is 50.6M, its worked size is 11.6M , and let it be $D_0$ , there are 47536 transactions in $D_0$ . We partition $D_0$ into $D_1, D_2, D_3, D_4$ according to time period from $t_1$ to $t_4$, and the partitioned result is shown in Table 3.

**Table 3.** Partitioned result

| Sub-database | $D_1$ | $D_2$ | $D_3$ | $D_4$ |
|---|---|---|---|---|
| Time period | $t_1$（96.10-96.11） | $t_2$（96.12-97.01） | $t_3$（97.02-97.03） | $t_4$（97.04-97.05） |
| The number of transaction | 10918 | 10456 | 13801 | 12361 |

By setting minimum support to 0.04% and minimum confidence to 50%, the association rule set $R_0$ and corresponding meta-rule set $M_s$ and $M_c$ are discovered in the hyper-structure that is constructed through scanning the sub-database $D_1$, $D_2$, $D_3$, $D_4$ respectively. There are 25 entry association rules in $R_0$. There are 25 entry support meta-rules and 25 entry confidence meta-rules in $M_s$ and $M_c$ respectively. The classified result of the support meta-rule set is showed in Table 4. The classified result of the confidence meta-rule set is omitted.

**Table 4.** Classified result

| Type of rules | Stable | Increased | Decreased | Random |
|---|---|---|---|---|
| Number of rules | 3 | 0 | 3 | 19 |

By observing the classified result we can discover that classified result is correct. The supports of 3 entries “Stable” rules are all more than support threshold from $t_1$ to $t_4$ and support change is little. The support of 3 entries “Decreased” rule falls gradually from $t_1$ to $t_3$, and the rules are perished in $t_4$. In “Random” column, except 1 entry rule appears both $t_1$ and $t_3$, the rest only appear in one time period. So the association rules obtained from $R_0$ is not all usable and trend analysis is necessary. Experimental result indicates that the usability of association rules obtained by our approach is improved considerably.

## 7 Discussion

Existing algorithm for mining association rule [1-4] can only mine strong association rules with average support and confidence in total. Our approach is distinctly different from above mentioned algorithm, as not only it can mine strong association rules with average support and confidence in total, but also can mine more various association rules. Experimental result appears that our approach can mine the association rules that have better usability.

Compared with existing algorithm of meta-mining or change mining [8-9], our approach has higher efficiency. In existing algorithms of meta-mining or change mining, first the data sets are partitioned into several subset according to time segments, and then each subset is mined by existing algorithm for mining association rule to gain association rule sets, finally, the meta-rule sets or trend rules are mined from association rule sets. With the approach proposed in this paper, global strong rule sets, meta-rule sets and the classification of the meta-rule sets can be obtained by scanning database only twice, so our approach has lower I/O spending than above mentioned algorithm.

In addition, our algorithm will generate less numbers of 2-itemsets than Apriori algorithm, because the 2-itemsets is generated by linking all 1-frequent item in Apriori algorithm. However, in our algorithm, the 2-itemsets is generated by linking 1-frequent item in intersection between 1-frequent itemset and each transaction. In our algorithm, the confidence of association rule can be calculated by obtaining the count of corresponding itemsets from the hyper-structure directly.

## 8 Conclusion

In this paper, we put forward an integrated approach for mining association rules and meta-rules. With the approach, the association rules and meta-rules can be mined for time serial database effectively, Not only has it high efficiency, but also has more powerful mining capability, It has some distinct advantages compared with existing algorithms. A formalized expression of meta-rules for time serial databases is given, thus meta-rules can be denoted and processed expediently, this offer a new pass for expression or re-mining of meta-rules.

## References


[1] Agrawal R. and Srikant R.: Fast algorithms for mining association rules. In VLDB'94(1994), 487-499.

[2] Han J., Pei J. and Y.Yin.: Mining frequent patterns without candidate generation. In SIGMOD'00(2000), 1-12

[3] Yudho G.S., Raj P., Gopalan.: CT-ITL: Efficient Frequent Item Set Mining Using a compressed Prefix Tree with Pattern Growth. 14th Australasian Database Conference(ADC2003)(2003).

[4] Yang M., Sun Z.H., Song Y.Q.: Fast Updating of Globally Frequent Itemsets. Journal of Software,(8)( 2004)1189-1196.

[5] Spiliopoulou M., Roddick J.F.: Higher order mining: modelling and mining the results of knowledge discovery, Conf. on Data Mining Methods and Databases for engineering, Finance, and Other Fields, WIT Press, Southampton, UK( 2000)309–320.

[6] Liu B., Wynne H., Heng S.H. *et al*.: Mining Changes for Real-life Applications, in The 2nd International Conference on Data Warehousing and Knowledge Discovery, UK(2000).

[7] Bing L., Wynne H. and Ming Y.: Discovering the Set of Fundamental Rule Changes, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2001).

[8] Wai H. A., Keith C.C.: Mining changes in association rules: a fuzzy approach. Fuzzy Sets and Systems, In Press, Corrected Proof, Available online 11 September 2004(2004).

[9] Abraham T., Roddick J. F.: Incremental Meta-mining from Large Temporal Data Sets, Advances in Database Technologies, Proceedings of the 1st International Workshop on Data Warehousing and Data Mining (DWDM'98)(1999),41-54.