A Framework of Numerical Basket Analysis

Takashi Washio, Atsushi Fujimoto and Hiroshi Motoda
The Institute of Scientific and Industrial Research
Osaka University
8-1, Mihogaoka, Ibarakishi, Osaka, 567-0047, Japan
washio@ar.sanken.osaka-u.ac.jp

Abstract

Basket Analysis is mathematically characterized and extended to search families of sets in this paper. These theories indicate the possibility of various new approaches of data mining. We demonstrate the potential through proposal of a novel approach QARMINT. It performs complete mining of generic QARs within a low time complexity which has not been well addressed in the past work. Its performance evaluation shows high practicality.

1. Introduction

Since an algorithm of Basket Analysis was proposed by Agrawal and Srikant [1], a large number of researches on more efficient Basket Analysis have been presented in the field of data mining. A basic principle underlying all of the algorithms is the bottom up building of candidate itemsets in a lattice under a downward closure property of itemsets, *i.e.*, "if any given itemset a is not large, any superset of a will also not be large." The most representative measure to introduce the downward closure property of the itemsets is "*support*," *i.e.*, occurrence frequency of an itemset in given transaction data. If an itemset a occurs more than a threshold value, *i.e.*, "*minimum support*," it is called a "*frequent itemset*." When two itemsets a_{k-1} and b_{k-1} sharing their $k-2$ elements are frequent, their *join* c_k is a candidate frequent itemset.

Some issues remain in the current Basket Analysis where transactions and itemsets are limited to finite Boolean sets. The aforementioned basic principle has wider applicability not limited to the search on the finite Boolean lattice, because it requires only a search space having (1) a join operation between two sets and (2) a downward closure property among sets. In spite of this wide applicability, the framework of the Basket Analysis has not been extended to address more generic tasks.

Another issue is the analysis of transaction data including items with numeric values such as "$Age : 32$" and "$NumCars : 2$." These items are called "*numeric items*" whereas the items having categorical values such as "$Married : Yes$" are called "*categorical items*." The clause of an item such as "Age" is called an "*attribute*." Some categorical item may be only a clause as "$Beer$" without its value. An association rule in which every numeric item has appropriate intervals of its value is called a "*quantitative association rule*" (QAR). An example QAR is "$\{Age : [30, 39]$ *and* $Married : Yes\} \Rightarrow \{NumCars : [2, 2]\}$" which states "a person who is thirties and married owns two cars." Since Srikant and Agrawal proposed an approach to mine QARs [2], number of studies on the QAR mining have been made. However, the problem to mine a complete set of QARs in generic form under representative mining measures is known to be *NP-complete* [4]. The state of the art has not addressed the complete mining of generic QARs within tractable time complexity.

In this paper, first, we extend the framework of the Basket Analysis to searching families of sets based on the mathematical characterization of the aforementioned basic principle. Second, we propose a novel approach and its implementation for complete mining of generic QARs within a low time complexity $O(N \log N)$ based on the extension where N is the number of transactions in data. This approach is called *QAR mining by Monotonic INTerval* (QARMINT) by the nature of its mining criterion. Its low time complexity in terms of the data amount is essential for mining large data. Third, its performance evaluation is presented to show practicality.

2. Extension of Basket Analysis

As mentioned earlier, the basic principle of the Basket Analysis requires only a search space having (1) a join operation between two sets and (2) a downward closure property among sets. The operation (1) introduces a structure on the search space. Let L be a "*family of sets*" in which ele-

ments are sets. Let a "*join*" of two elements a, b in L be an "*upper bound*" $a \cup b$ which follows the rules
Commutative Rule : $a \cup b = b \cup a$, and
Associative Rule **:** $a \cup (b \cup c) = (a \cup b) \cup c$.
L is called an "*upper semilattice*" when $a \cup b$ exists in L for any pair of elements a, b in L. Accordingly, the space searched by the join operation is an upper semilattice. On the other hand, the search space of the conventional Basket Analysis is a finite Boolean lattice where it is finite, an upper bound $a \cup b$ and a lower bound $a \cap b$ exist, and commutative rule, associative rule, absorption rule, distributive rule and complement rule must hold for the upper bound and the lower bound. The former search space is far less constrained than the latter space.

The property (2) of a search space is defined in a more generic way than that by the conventional support. Let an "*inclusion relation*" $a \subseteq b$ be an "*ordered pair*" $\{a, b\}$ of two elements a, b where $\{a_1, b_1\} = \{a_2, b_2\}$ iff $a_1 = a_2$ and $b_1 = b_2$ hold. Let L be an "*ordered family of sets*" where some pairs of its elements have the inclusion relations. Give a property P of L where $P(a)$ means that $a (\in L)$ has the property P. Then the downward closure property P of L is defined as

$$a \subseteq b \Rightarrow P(b) \rightarrow P(a), \qquad (1)$$

where $a, b \in L$. When L is an upper semilattice, $a \subseteq b$ is given by $a \cup b = b$. In familiar settings, $P(a)$ is $sup(a) = |\{t | t \in D, a \subseteq t\}| \geq minsup$ where D is a database, t a transaction in D and $minsup$ minimum support.

Upon the above characterization, the basic principle of the Basket Analysis is known to be applicable to wider classes of problems whose search space has the upper semilattice structure and the generic downward closure property. We further extend the Basket Analysis to the search on families of sets. Given two families of sets A, B and two sets $a \in A, b \in B$, let $f(a, b)$ be a set function to map the pair of a and b onto a family of sets F. Under this definition of f, we define the following join operation.

$$A \cup B = \{c | c \in F = f(a, b), a \in A \text{ and } b \in B\}. \qquad (2)$$

A set of families of sets L is an upper semilattice, if $A \cup B \in L$ for any pair of families of sets $A, B \in L$. We also introduce an extended downward closure property. Given two families of sets A, B, let an inclusion relation $a \subseteq b$ be an ordered pair $\{a, b\}$ of $a \in A, b \in B$. Let L be an "*ordered set of families of sets*" where some pairs of sets a, b in some pairs of families of sets A, B have the inclusion relations. Then the downward closure property P of L is defined by Eq.(1) where $a \in A, b \in B$ and $A, B \in L$. When L is an upper semilattice, $a \subseteq b$ is given by $f(a, b) = \{b\}$. By these definitions, L is a search space of the Basket Analysis on families of sets.

3. Complete Mining of Generic QARs

3.1. QAR mining by Monotonic INTerval

We propose a novel approach called "*QAR mining by Monotonic INTerval* (QARMINT)" for complete mining of generic QARs within a low time complexity. The key ideas of QARMINT are to use the aforementioned extension of the Basket Analysis to families of sets and to introduce a "*Monotonic INTerval* (MINT)" measure having the downward closure property on hyper rectangles formed by numeric items.

First, we define some mining measures. Let a binary (p, q) be an item. (p, q) is called a numeric item if q is a closed interval on continuous number field, whereas (p, q) is called a categorical item if q is a categorical symbol. p stands for an attribute of (p, q). Let an itemset a be a set of items (p, q)s and a set of attributes of a $a_p = \{p | (p, q) \in a\}$. Given a pair of itemsets a and b, b supports a, when b is more or equally restrictive to a. It is represented as $a \subseteq b$, and defined as $\forall (p, r) \in a, \exists (p, s) \in b, r \supseteq s$ for a numeric item and $r = s$ or $s = null$ for a categorical item. $s = null$ means that any value is not admitted at s, and hence it is the most restrictive. Let a transaction t be a set of items and a data set D a collection of transactions ts. q of $(p, q) \in t$ is usually a point interval representing a unique value while q can be a finite interval in general. Let L be a set of families of sets where some pairs of sets a, b in some pairs of families of sets A, B have the inclusion relations $a \subseteq b$s. Given "*support*" of a as $sup(a) = |\{t | t \in D, a \subseteq t\}|$, a property P of L is that $P(a)$ is $sup(a) \geq minsup$. Then P is a downward closure property of L according to Eq.(1). Moreover, let $a \rightarrow c$ be a QAR where $a_p \cap c_p = \phi$. Then "*confidence*" of $a \rightarrow c$ is given by $conf(a \rightarrow c) = sup(b)/sup(a)$ where $b = \{(p, q) | (p, q) \in a \text{ or } (p, q) \in c\}$.

We further introduce a novel class of mining measures on the hyperspace formed by multiple numeric attributes. Give itemsets a, b and c where $a \subseteq b$ and a property P of L where $P(a)$ and $P(b)$ are $a \subseteq c$ and $b \subseteq c$ respectively. Because the aforementioned definition of $a \subseteq b$ and $b \subseteq c$ which is $\forall (p, s) \in b, \exists (p, u) \in c, s \supseteq u$ for a numeric item and $r = s$ or $s = null$ for a categorical item, $\forall (p, r) \in a, \exists (p, u) \in c$, $r \supseteq u$ for a numeric item and $r = u$ or $u = null$ for a categorical item. Hence, $a \subseteq c$. Then P is a downward closure property of L according to Eq.(1). A mining measure to define intervals having this property from data D is called a "*Monotonic INTerval* (MINT)" measure. An advantage of MINT is that the optimum intervals for numeric items can be derived in low time complexity by the monotonicity.

An example of a MINT measure is the following "*denseness*". Given two numeric items $(p, q_i) \in t_i$ and $(p, q_j) \in t_j$ where $t_i, t_j \in D$, let Δ_p be a "*permissible range*" of p.

Then t_i and t_j are "*close*" on p if

$$inf(q_i) - sup(q_j) \leq \Delta_p \text{ and } inf(q_j) - sup(q_i) \leq \Delta_p. \tag{3}$$

Here, $sup(q)$ and $inf(q)$ are the upper bound and the lower bound of q. Given a projection mapping m_a of $t \in D$ to the space formed by all numeric attributes in a_p, and given a monotone hyper rectangular region R_a formed by intervals qs on all numeric attributes in a_p, let $D_a = \{t | t \in D, a \subseteq t, m_a(t) \in R_a\}$. When every $t \in D_a$ has another $t' \in D_a$ which is close on all numeric attributes in a_p, and all of such close pairs are mutually connected through the other close pairs in D_a, R_a is called a "*dense region*" of a. If any monotone hyper rectangular region $R'_a(\supseteq R_a)$ is not dense, R_a is called a "*maximal dense region* (*mdr*)" under given data D. Define a "*maximal dense interval* (*mdi*)" q_a of each numeric attribute p in a_p as the projection of the *mdr* R_a onto p. Consider another itemset b where $a_p \subseteq b_p$ and $\forall$ categorical items $(p, r) \in a, (p, r) = (p, s) \in b$. When $a_p = b_p$, the *mdi* q_b of each numeric attribute in b_p is identical with the *mdi* q_a of the numeric attribute in a_p, since $D_b = D_a$. When b_p has some attributes which is not in a_p, let the projection of the *mdr* R_b onto the space formed by all numeric attributes in a_p be $R_{b|a}$. Then $R_{b|a} \subseteq R_a$, since $D_b \subseteq D_a$. Accordingly, $q_b \subseteq q_a$ for each numeric attribute in a_p, and thus $a \subseteq b$. This concludes that the denseness measure that the interval of each numeric item is defined by its *mdi* is a MINT measure. The time complexity to derive *mdi*s of numeric attributes is $O(N^2)$ in the worst case because the pair wise evaluation of Eq.(3) is needed, while its practical complexity is $O(N \log N)$ as shown later.

Next, we define "*join*" operation of two families of sets A, B in L. Given $a(\in A)$ and $b(\in B)$, a join $F = f(a, b)$ is defined as follows. Let c be an itemset where $c_p = a_p \cup b_p$.

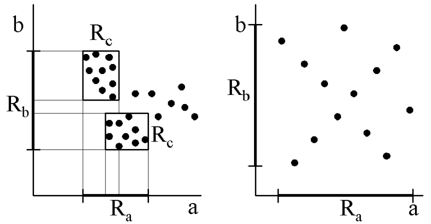
(1) Given $(p, r) \in a, (p, s) \in b$ for all categorical $p \in a_p \cap b_p$, let $(p, r) \in c$ if $r = s$ otherwise $(p, null) \in c$.

(2) Given $(p, r) \in a$ for all categorical $p \in c_p$ and $p \notin b_p$, let $(p, r) \in c$.

(3) Given $(p, s) \in b$ for all categorical $p \in c_p$ and $p \notin a_p$, let $(p, s) \in c$.

(4) Given an *mdi* of c, q_c, for all numeric $p \in c_p$, let $(p, q_c) \in c$.

Given $(p, r) \in a, (p, s) \in b$ and $(p, u) \in c$, $r = u$ or $u = null$ for each categorical $p \in a_p$ and $s = u$ or $u = null$ for each categorical $p \in b_p$ from (1) to (3). From (4) and the denseness being a MINT measure, $r \supseteq u$ for each numeric $p \in a_p$ and $s \supseteq u$ for each numeric $p \in b_p$. Accordingly, $a \subseteq c$ and $b \subseteq c$, and the join $F = f(a, b)$ gives the upper bounds of a, b. c may not be unique, since multiple *mdr*s R_cs can be derived. Also c may not exist, since $(p, null)$ can be obtained in (1), or the *mdr* R_c can not exist, *i.e.*, $q_c = null$ in (4). Figure 1 depicts these cases. In (a), R_c of the combined itemset c is multiple due to the lack of uniformity of $t \in D_c$, even if R_a and R_b of the original itemsets are unique respectively. In (b), R_c does not exist due to the low denseness of $t \in D_c$. Accordingly, F derived via $f(a, b)$ is a family of sets in general. Then we obtain the join operation $C = A \cup B$ by Eq.(2). From the above discussion, $\forall a \in A, \exists c \in C, a \subseteq c$, and thus $A \subseteq C$. Similarly, $B \subseteq C$. This indicates that the join $A \cup B$ gives the upper bound of A, B and an upper semilattice L.



(a) Multiple R_cs are derived. (b) No R_c is derived.

Figure 1. Derivation of R_c by join.

Based on this definition of join operation on families of sets with denseness and the definitions of support and confidence, the most of the standard algorithms of the Basket Analysis whose complexity is $O(N)$ can be applied to derive generic QARs from data.

3.2. Implementation

To assess the basic features of QARMINT, we used the standard Apriori-TID algorithm [1], since it is principally an algorithm running on memory, and its computational features are well known. Instead of hash tables, the trie data structure as depicted in Fig. 2 was used under lexicographically ordered itemsets. If any subsets of the joined set $c \in F = f(a, b)$ are not frequent according to a given $minsup$, c is pruned before its *mdr* R_c is computed. Moreover, after computing the *mdr* R_c, c is pruned if c is not frequent. The pruning by these checks are indicated by the slashed itemsets in Fig. 2. A difference from the original Apriori-TID algorithm is that the join of two itemsets a, b within a family depicted by a solid box is not allowed, and the itemsets cs obtained from a pair of families A, B belong to an identical family C. Another difference is that a join of a, b can generate multiple itemsets cs as depicted in a dashed box.

The most expensive process in QARMINT is to derive the *mdr* R_cs of joined itemset c. We introduce an iterative approach to reduce the required computation time. Given $c = \{(p_1, q_1), ..., (p_k, q_k)\}$, first, all transactions in D are sorted for each attribute p_i in c. This is $O(N \log N)$. Then, the *mdi*s on the number line of p_1 are computed from the transactions without taking into account the other attributes. When multiple *mdi*s are obtained, one of them is focused, and the transactions in the *mdi* is retained. Next, the identi-

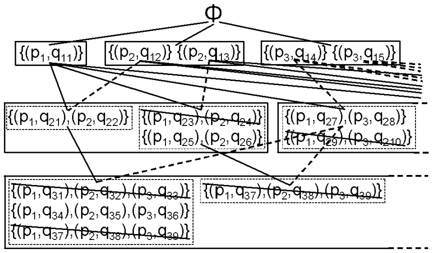

Φ
$\{(p_1,q_{11})\}$
$\{(p_2,q_{12})\}$ $\{(p_2,q_{13})\}$
$\{(p_3,q_{14})\}$ $\{(p_3,q_{15})\}$
$\{(p_1,q_{21}),(p_2,q_{22})\}$
~~$\{(p_1,q_{23}),(p_2,q_{24})\}$~~
$\{(p_1,q_{25}),(p_2,q_{26})\}$
$\{(p_1,q_{27}),(p_3,q_{28})\}$
~~$\{(p_1,q_{29}),(p_3,q_{210})\}$~~
~~$\{(p_1,q_{31}),(p_2,q_{32}),(p_3,q_{33})\}$~~
$\{(p_1,q_{34}),(p_2,q_{35}),(p_3,q_{36})\}$
~~$\{(p_1,q_{37}),(p_2,q_{38}),(p_3,q_{39})\}$~~
~~$\{(p_1,q_{37}),(p_2,q_{38}),(p_3,q_{39})\}$~~

Figure 2. Trie data structure.

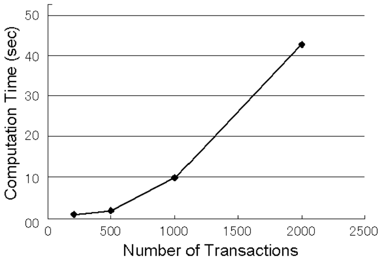


Figure 3. Time complexity.

cal process is applied to p_2, and this recursively continues in depth first search (DFS). After the *mdi*s on p_k is computed, the process continues again from p_1 until the *mdi* of every $p_i(i = 1, ..., k)$ converges. The *mdi*s always converge to these of the *mdr* R_c because the denseness is a MINT measure. After the convergence, the search is backtracked to the next *mdr* R_c. The computation of *mdi*s in each step requires $O(N)$ time at most. In the worst case, only one transaction is dropped in each step, and N steps required until the *mdi*s converge. Thus, $O(N^2)$. However, this does not likely occur. Practically, only a portion of the transactions are retained in each step. Let $0 < \alpha < 1$ be an expected rate of transactions retained in each step, m the required steps for convergence. The process to search an *mdr* R_c stops at the latest when the number of retained transactions $\alpha^m N$ becomes less than $minsup$. By solving the equation $minsup \simeq \alpha^m N$ with m, m is $O(\log N)$. Accordingly, the expected time complexity of this most expensive process is $O(N \log N)$.

4. Performance Evaluation

The performance of QARMINT has been evaluated through both artificial data and real bench mark data. Sets of artificial data have been generated under various conditions. The characteristics of the computation time is simlilar to the conventional Basket Analysis except for Δ_p and N. The time moderately increases when Δ_ps of all attributes are increased. This is because wider permissible ranges increases the number of *mdr*s. Figure 3 shows the dependency of the computation time on the number of transaction N. The curve almost follows the relation $O(N \log N)$.

The real bench mark data "*Labor relations Database*" in UCI Machine Learning Repository [3] was analyzed by QARMINT. It contains 57 instances, 8 numeric attributes and 8 categorical attributes and many missing values. We ignored the attributes of missing values in each instance, and transformed the data into transactions. Though the size of this data is quite small, we found many interesting QARs associated with the labor conditions under $minsup = 0.1$ and $\Delta_p = 0.1$ which is 10% of the maximum and minimum values of each p in the data. The following two are examples.

$sup = 35\%, conf = 65\%,$
$class : good, duration - years : [2, 2] \Rightarrow$
$working - hours : [33 - 40],$
$wage - increase - second - year(\%) : [4.0 - 5.8],$
$sup = 35\%, conf = 65\%,$
$class : good, duration - years : [3, 3] \Rightarrow$
$working - hours : [35 - 40],$
$wage - increase - second - year(\%) : [3.5 - 5.0].$

These rules indicate that the workers having longer duration contracts and evaluating their labor condition as $good$ admit longer working times and less wage increase. These evaluations indicate the sufficient tractability and the practical applicability of QARMINT.

5. Conclusion

The mathematical characterization and the extension of the Basket Analysis presented in this paper are expected to provide variety of new approaches of data mining. Their potential has demonstrated by a novel approach called QARMINT for complete mining of generic QARs within a low time complexity. We are implementing QARMINT in a more efficient algorithm and evaluating its performance in near future.

Acknowledgement

This research has conducted under the support of JSPS Kiban Kenkyuu (B) (2) 16300045.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. *Proc. of 20th Int. Conf. on Very Large Data Bases (VLDB)*, pages 487–499, 1994.

[2] R. Srikant and R. Agrawal. Mining quantitative association rules in large relational tables. *Proc. of 1996 ACM SIGMOD Int. Conf. on Management of Data*, pages 1–12, 1996.

[3] U. C. I. (UCI). *UCI Machine Learning Repository*. UCI, http://www.ics.uci.edu/ mlearn/MLRepository.html, 2004.

[4] J. Wijsen and R. Meersman. On the complexity of mining quantitative association rules. *Data Mining and Knowledge Discovery*, 2(3):263–281, September 1998.