Traversing Itemset Lattices with Statistical Metric Pruning

Shinichi Morishita
Graduate School of Frontier Sciences
University of Tokyo
7-3-1 Hongo, Bunkyo Ward
Tokyo 113-0033, Japan
moris@k.u-tokyo.ac.jp

Jun Sese
Graduate School of Frontier Sciences
University of Tokyo
7-3-1 Hongo, Bunkyo Ward
Tokyo 113-0033, Japan
sesejun@gi.k.u-tokyo.ac.jp

ABSTRACT

We study how to efficiently compute significant association rules according to common statistical measures such as a chi-squared value or correlation coefficient. For this purpose, one might consider to use of the Apriori algorithm, but the algorithm needs major conversion, because none of these statistical metrics are anti-monotone, and the use of higher support for reducing the search space cannot guarantee solutions in its the search space. We here present a method of estimating a tight upper bound on the statistical metric associated with any superset of an itemset, as well as the novel use of the resulting information of upper bounds to prune unproductive supersets while traversing itemset lattices. Experimental tests demonstrate the efficiency of this method.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining, Statistical Databases

1. MOTIVATING EXAMPLE

Association rules [2] have been widely studied in recent years. We call a set of items an *itemset*. Let D be a set of transactions wherein each transaction is an itemset. For instance, in Table 1 (A), each row except the first one represents an itemset, and each column denotes an item. "1" indicates the presence of the item in the row, while "0" indicates the item's absence. For example, the fourth row expresses {a,c,e}.

Let I be an itemset, and let $Pr(I)$ denote the ratio of the number of transactions that include I to the number of all transactions in D. We call $Pr(I)$ the *support* of I. In our running example, $Pr(\{a,b\}) = 25\%$ and $Pr(\{a,b,c\}) = 12.5\%$. An association rule has the form $I_1 \Rightarrow I_2$, where I_1 and I_2 are disjoint itemsets. The support of $I_1 \Rightarrow I_2$ is defined as $Pr(I_1 \cup I_2)$, while the confidence is $Pr(I_1 \cup I_2)/Pr(I_1)$. For instance, the support and the confidence of $\{a,b\} \Rightarrow \{c\}$ are 12.5% and 50%, respectively.

Agrawal et al. [2] addressed the problem of enumerating all association rules that have support and confidence values no smaller than the user-specified minimum thresholds. They developed important techniques and designed the well-known Apriori algorithm [3].

Apriori and its variants

An itemset I is *large* if its support $Pr(I)$ is no less than the user-specified minimum support. The strategy of Apriori is two-phased. First it enumerates all large itemsets, and then it derives association rules. For instance, from a large itemset, say $\{a,b,c\}$, Apriori tries to generate $\{a,b\} \Rightarrow \{c\}$. To speed up the expensive step of listing all large itemsets, Agrawal and Srikant developed a method of searching the lattice of itemsets with respect to itemset-inclusion. The strategy starts from the empty set and scans itemsets from smaller to larger in an incremental manner.

The support of an itemset is *anti-monotone* with respect to set-inclusion of itemsets; that is, for any $I \subseteq J$, $Pr(I) \geq Pr(J)$. Thus, whenever an itemset is not large with respect to a minimum support threshold, neither is any of its supersets. The Apriori algorithm uses this heuristics to effectively prune away a substantial number of unproductive itemsets. Ng et al. [15; 22] proposed a general class which they termed constraint association rules and to which the same pruning heuristics could be applied. Tsur et al. [25] used the anti-monotonicity to derive Datalog rules efficiently. Several other improvements of Apriori have also been reported [4; 7; 23].

Correlation

However, one drawback of the support-confidence framework is its weakness at expressing the notion of correlation [1; 6; 16]. Table 1 (B) shows some association rules derived from the database in Table 1 (A). Let us consider which rules are valuable. Statistically speaking, the first and the third rules do not make sense, because in each rule the assumptive itemset, say I, and the conclusive itemset, say C, are independent; that is, $Pr(I \cup C) = Pr(I) \times Pr(C)$. On the other hand, in the second rule, the assumption and the conclusion are highly and positively correlated.

This example suggests measuring the usefulness of an association rule by the significance of correlation between the assumption and the conclusion. For instance, the analysis

a	b	c	d	e
1	1	1	1	1
1	1	0	1	1
1	0	1	0	1
1	0	0	0	1
0	1	1	0	1
0	1	0	0	1
0	0	1	1	1
0	0	0	1	1

(A) Examples of Transactions

$I \Rightarrow C$	support	confidence	correlated ?
$\{x\} \Rightarrow \{y\}$ $(x, y \in \{a, b, c, d\}, x \neq y)$	25%	50%	No
$\{a, b\} \Rightarrow \{d\}$	25%	100%	Yes
$\{a\} \Rightarrow \{e\}$	50%	100%	No

(B) Examples of Association Rules

Table 1: Transactions and Association Rules

of scientific data calls for a method of discovering correlation among various phenomena. For this purpose, the chi-squared value is typically used because of its solid grounding in statistics. Other related statistical measures, such as correlation coefficient, entropy information gain, gini index and interclass variance have also been frequently used. There have been some arguments about the choice of statistical measures in statistics and artificial intelligence, but our proposal is independent of the choice, because we will present a general method that can handle all those standard statistical metrics in a uniform manner. But, for the moment, for the sake of simplicity, we continue this discussion by using the chi-squared value, and then we will present how to generalize the method.

2. ENUMERATION

Use of the chi-squared value instead of support and confidence motivates us to consider the following enumeration problem:

- **Enumeration Problem:** Enumerate all significant association rules that have chi-squared values no smaller than the user-specified minimum cutoff value, say at the 95% significance level.

To this problem, the application of the Apriori algorithm has been investigated [1; 6]. Brin et al. [6] proposed a method of enumerating large itemsets first and then selecting correlated itemsets. However, use of the support threshold may not always prune unproductive itemsets effectively. For instance, from the rules in Table 1 (B) we would like to select only the second rule, whose support is 25%, but the threshold of 25% does not discard itemsets of the form $\{x, y\}$ where $x, y \in \{a, b, c, d\}$, which are not correlated and therefore are irrelevant. In general, use of the lower threshold would generate many irrelevant itemsets, while the higher threshold may lose relevant itemsets.

To avoid the use of a support threshold, Aggarwal and Yu [1] proposed the generation of a strongly collective itemset that requires correlation among items of any subset. If an itemset is not strongly collective, neither is any of its supersets, and therefore Apriori's pruning strategy can be successfully used for enumerating strongly collective itemsets [22]. However, this constraint might be too restrictive to output some desired rules. In our running example, a and b are not correlated at all, and hence $\{a, b, d\}$ is not strongly collective. Thus we cannot derive "$\{a, b\} \Rightarrow \{d\}$."

These previous approaches present some difficulties of using the Apriori's strategy for the enumeration problem. We investigate another approach to the problem. Our idea is that we first select the conclusive itemset (say $\{d\}$), proceed to search the assumptive itemset (say $\{a, b\}$) that is significantly correlated with $\{d\}$, and derive "$\{a, b\} \Rightarrow \{d\}$." In real applications, we are often interested in a particular itemset C and want to find itemsets that are highly correlated with C. Thus restricting the enumeration problem to the following form makes sense.

- **Item-wise Enumeration Problem:** For a fixed conclusion C, enumerate all significant association rules of the form $I \Rightarrow C$ that have chi-squared values no less than the user-specified minimum cutoff value, say τ.

Let $chi(I)$ denote the chi-squared value of rule $I \Rightarrow C$. Then, our goal is to develop an efficient way of enumerating $\{I \mid chi(I) \geq \tau\}$. One might wonder if the pruning strategy of Apriori is effective for this enumeration problem. Unfortunately, however, $chi(I)$ is not anti-monotone wrt set-inclusion, which is the major obstacle to the application of Apriori's pruning method.

We solve this problem as follows: We scan the lattice of itemsets beginning with smaller itemsets and continuing to larger ones. Suppose that we investigate an itemset I during the search. We develop a method of computing an upper bound, denoted by $u(I)$, on $\{chi(J) \mid I \subseteq J\}$ by using the convexity of the chi-squared function. If $u(I) < \tau$, for any superset J of I, $chi(J) \leq u(I) < \tau$, and hence we can safely prune $\{J \mid I \subseteq J\}$ at once.

There are more opportunities to prune the sub-lattice $\{J \mid I \subseteq J\}$ if the value of $u(I)$ is lower and closer to $\max\{chi(J) \mid I \subseteq J\}$. In general, however, it is difficult to calculate the maximum value unless we investigate all the supersets of I. Thus, it is a non-trivial question whether or not we can estimate a tight upper bound on $\{chi(J) \mid I \subseteq J\}$. We will present our algorithm, and empirically evaluate its effectiveness through a substantial number of experiments.

3. OPTIMIZATION

In many practical applications, we often face a situation in which a large number of association rules are generated. To resolve this problem, one may sort association rules according to their significance and screen out insignificant rules. Ideally, however, instead of computing too many rules, we

want to directly compute the most significant n rules. Furthermore, focusing on the most significant n rules might help us to abandon unimportant itemsets earlier in computation, and thereby accelerate the overall performance. Formally, the problem is defined as follows:

- **Optimization Problem:** For a fixed conclusion C, compute the optimal association rule of the form $I \Rightarrow C$ that maximizes the chi-squared value, or list the most significant n solutions.

If we treat the maximum number of items in an itemset as a variable, the problem is NP-hard (see Appendix), but in real applications, the maximum number is usually bounded by a constant, and hence the problem is tractable. To solve this optimization problem efficiently, we extend the idea for solving the item-wise enumeration problem. We also associate with an itemset I an upper bound $u(I)$ on $\{chi(J) \mid I \subseteq J\}$, but the difference is that during the scan of the itemset lattice, we always maintain the temporarily maximum (or n-th largest) chi-squared value among all the chi-squared values calculated so far, and set it to the cutoff value τ. If $u(I) < \tau$, no superset of I gives a chi-squared value greater than or equal to τ, and hence we can safely prune $\{J \mid I \subseteq J\}$.

The idea of pruning with the information of upper bounds is a standard technique in combinatorial optimization, but the novelty of our proposal is to utilize the idea during the traverse of the lattice of itemsets. There have been proposed some previous approaches to traverse a search tree of itemsets in a best-first manner [5; 20]. Bayardo considered various metrics including statistical ones and presented a general-purpose method of approaching the optimal solution by searching a support/confidence border [5]. But if we focus on statistical metrics such as a chi-squared value, we can further reduce the search space, and we only need to scan the convex hull of a support/confidence border. We will exploit this technique in this paper. The mining of optimized ranges or regions for numerical attributes has also been intensively studied [8; 10; 11; 12], but not with reference to the mining of optimal association rules over itemsets.

4. MAIN RESULTS

4.1 Correlation

DEFINITION 4.1. *Let $I \Rightarrow C$ be an association rule, D be a set of transactions, and n be the number of transactions in D. In the following contingency table, rows I and $\bar{I}$ show the number of transactions that do and do not contain I, respectively. Columns C and $\bar{C}$ correspond to the conclusion C. Each value in the last column (row, resp.) shows the summation of the two values in the second or third columns (rows) and in the same row (column).*

	C	$\bar{C}$	$\sum row$
I	O_{IC}	$O_{I\bar{C}}$	O_I
$\bar{I}$	$O_{\bar{I}C}$	$O_{\bar{I}\bar{C}}$	$O_{\bar{I}}$
$\sum column$	O_C	$O_{\bar{C}}$	n

$O_{I\bar{C}}$, for instance, represents the number of transactions that contain I but do not include C. O_I, for example, shows the number of transactions that contain I, which is equal to the sum of the values in the row, $O_{IC} + O_{I\bar{C}}$. For each pair $(i, j) \in \{I, \bar{I}\} \times \{C, \bar{C}\}$, we calculate expectation under the assumption of independence:

$$E_{ij} = n \times (O_i/n) \times (O_j/n).$$

The chi-squared value is the normalized deviation of observation from expectation; namely,

$$\sum_{i \in \{I, \bar{I}\}, j \in \{C, \bar{C}\}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

If we fix the conclusion C, O_C and $O_{\bar{C}}$ can be regarded as constants, and let m denote O_C. Also, let x and y denote O_I and O_{IC} respectively. The contingency table then becomes:

	C	$\bar{C}$	$\sum row$
I	$O_{IC} = y$	$O_{I\bar{C}}$	$O_I = x$
$\bar{I}$	$O_{\bar{I}C}$	$O_{\bar{I}\bar{C}}$	$O_{\bar{I}} = n - x$
$\sum column$	$O_C = m$	$O_{\bar{C}} = n - m$	n

Since n and m are independent of the choice of the assumption I, the values of x and y uniquely determine the chi-squared value. Thus we will refer to the chi-squared value as $chi(x, y)$; namely,

$$chi(x, y) = \sum_{i \in \{I, \bar{I}\}, j \in \{C, \bar{C}\}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

In the above definition, each O_{ij} must be non-negative, and hence

$$0 \le y \le x, \quad \text{and} \quad 0 \le m - y \le n - x.$$

Also, each E_{ij} must be greater than zero, and hence O_i must be greater than zero for each $i \in \{I, \bar{I}, C, \bar{C}\}$. Thus, $chi(x, y)$ is defined for $0 < x < n$ and $0 < m < n$. We will extend the domain of $chi(x, y)$ to include $(0, 0)$ and (n, m). Since

$$\lim_{x \to 0} chi(x, y) = \lim_{x \to n} chi(x, y) = 0,$$

we define

$$chi(0, 0) = chi(n, m) = 0.$$

Incidentally, it is often helpful to explicitly state that x and y are determined by I, and we define:

$$x = x(I), \quad \text{and} \quad y = y(I).$$

□

DEFINITION 4.2. *A function $f(x, y)$ is convex if for any (x_1, y_1) and (x_2, y_2) in the domain of f, and for any $0 \le \lambda \le 1$,*

$$\begin{aligned} & f(\lambda(x_1, y_1) + (1 - \lambda)(x_2, y_2)) \\ \le \quad & \lambda f(x_1, y_1) + (1 - \lambda) f(x_2, y_2). \end{aligned}$$ □

PROPOSITION 4.1. *$chi(x, y)$ is a convex function.*

$$\lim_{x \to 0} chi(x, y) = \lim_{x \to n} chi(x, y) = 0.$$

$chi(x, y)$ is minimum if $y = \frac{m}{n} x$.

PROOF. See Appendix. □

4.2 Upper Bound

Let I be an arbitrary itemset. We now describe the method for estimating a tight upper bound of $\{chi(x(J), y(J)) \mid J \supseteq I\}$. With each $J \supseteq I$ we associate the point $(x(J), y(J))$, which we call a *stamp point*, and Figure 1 illustrates the resulting stamp points. The following theorem presents a method of calculating an upper bound.

THEOREM 4.1. *For any $J \supseteq I$,*

$$\begin{aligned} & chi(x(J), y(J)) \\ \leq\ & \max\{\ chi(y(I), y(I)),\ chi(x(I) - y(I), 0)\ \}. \end{aligned}$$

PROOF. Observe the following inequalities:

$$\begin{array}{ll} 0 \leq x(J) \leq x(I) & 0 \leq y(J) \leq y(I) \\ y(J) \leq x(J) & x(J) - y(J) \leq x(I) - y(I) \end{array}$$

Any stamp point $(x(J), y(J))$ is mapped onto the gray parallelogram in Figure 1, whose vertexes are $(0, 0)$, $(y(I), y(I))$, $(x(I), y(I))$, and $(x(I) - y(I), 0)$.

We continue to use the notations given in Definition 4.1, say m and n. Since

$$y(I) \leq m, \quad x(I) \leq n, \quad \text{and} \quad x(I) - y(I) \leq n - m,$$

$(x(I), y(I))$ and all the stamp points are mapped onto the quadrangle $(0, 0)$, $(y(I), y(I))$, (n, m), and $(x(I) - y(I), 0)$. It is known that any convex function is maximized at one of the vertexes on the boundary of a convex polygon [14]. From Proposition 4.1, both $(0, 0)$ and (n, m) minimize $chi(x, y)$. Thus, $(y(I), y(I))$ or $(x(I) - y(I), 0)$ must maximize $chi(x, y)$ among all the stamp points in $\{(x(J), y(J)) \mid J \supseteq I\}$, which completes the proof.

We also here present an alternative proof. Let L denote the line connecting $(0, 0)$ and (n, m). From Proposition 4.1, every stamp point on L minimizes $chi(x, y)$. Let $(x(J), y(J))$ be an arbitrary stamp point. If $(x(J), y(J))$ exists in the upper side of L. Draw the line from $(y(I), y(I))$ to $(x(J), y(J))$, and suppose that the line hits L at Q. Since Q minimizes $chi(x, y)$ and $chi(x, y)$ is a convex function,

$$chi(x(J), y(J)) \leq chi(y(I), y(I)).$$

On the other hand, if $(x(J), y(J))$ is in the lower side of L, we can similarly prove

$$chi(x(J), y(J)) \leq chi(x(I) - y(I), 0).$$

□

DEFINITION 4.3. *Let $u(I)$ denote*

$$\max\{chi(y(I), y(I)), chi(x(I) - y(I), 0)\}. \qquad \square$$

$u(I)$ is tight in the sense that there could exist $J \supseteq I$ such that $chi(x(J), y(J)) = u(I)$.

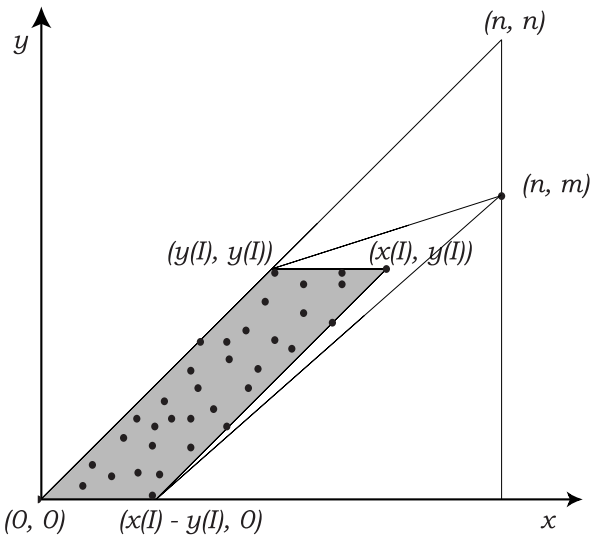


Figure 1: Stamp Points

4.3 Traversing Itemset Lattices

Item-wise Enumeration Problem

DEFINITION 4.4. *We treat the set of all itemsets as a lattice by regarding the set inclusion as the partial order, the union of itemsets as the least upper bound, and the intersection of itemsets as the greatest lower bound. The lattice is also called an itemset lattice.* □

DEFINITION 4.5. *Let τ be the user-specified minimum chi-squared value threshold. An itemset I is significant if*

$$chi(x(I), y(I)) \geq \tau.$$

An itemset I is promising if $u(I) \geq \tau$. □

Our goal is to compute the set of significant itemsets. For this purpose, the set of promising itemsets has three useful properties. First, any significant itemset is promising, and thereby the set of promising itemsets includes all significant itemsets. Second, if an itemset I is promising but may not be significant, it is worth searching the supersets of I, because there could be a significant superset of I. Third, if I is not promising, there is no point of searching the supersets of I, because any $J(\supseteq I)$ is not significant since $chi(x(J), y(J)) \leq u(I) < \tau$. Thus we calculate the set of promising itemsets to derive the set of significant itemsets.

DEFINITION 4.6. *An itemset is called a k-itemset if it contains k distinct items. Let P_k denote the set of all promising k-itemsets.* □

Since the set of all promising itemsets is $P_1 \cup P_2 \cup \ldots$, we construct P_k for each $k = 1, 2, \ldots$. We now introduce a candidate set for P_k that is useful for computing P_k efficiently.

DEFINITION 4.7. *We call an itemset I potentially promising if every proper subset of I (any subset smaller than I)*

τ is given by the user.
$Q_1 := \{I \mid I$ is a 1-itemset.$\}$; $k := 1$;
repeat begin
    If $k > 1$, generate Q_k from P_{k-1};
    Scan all the transactions to compute $u(I)$ and $chi(x(I), y(I))$ for each $I \in Q_k$;
    $P_k := \{I \in Q_k \mid u(I) \geq \tau\}$; $X := P_k$; $k++$;
end until $X = \phi$;
Return $\{I \in \cup_k P_k \mid chi(x(I), y(I)) \geq \tau\}$;

Figure 2: AprioriSMP for the Item-wise Enumeration Problem

$\tau := 0$;
$Q_1 := \{I \mid I$ is a 1-itemset.$\}$; $k := 1$;
repeat begin
    If $k > 1$, generate Q_k from P_{k-1};
    Scan all the transactions to compute $u(I)$ and $chi(x(I), y(I))$ for each $I \in Q_k$;
    $\tau := \max(\tau,\ \max\{chi(x(I), y(I)) \mid I \in Q_k\})$;
    $P_k := \{I \in Q_k \mid u(I) \geq \tau\}$; $X := P_k$; $k++$;
end until $X = \phi$;
Return τ with its corresponding itemset;

Figure 3: AprioriSMP for the Optimization Problem

is promising. Let Q_k denote the set of potentially promising k-itemsets. □

THEOREM 4.2. $Q_k \supseteq P_k$.

PROOF. Otherwise, let I be an itemset such that $I \in P_k$ but $I \notin Q_k$. There must exist such a proper subset J of I that is not promising; that is, $u(J) < \tau$. Then, any superset of J is not significant, and hence none of all supersets of I is significant. This however implies that I is not promising, which contradicts the assumption that $I \in P_k$. □

We can derive Q_k solely from P_{k-1} without scanning all the transactions. To this end, we use the idea of the `apriori-gen` function of the Apriori algorithm [3]; that is, we select two members in P_{k-1}, say I_1 and I_2, such that I_1 and I_2 share $(k-2)$ items in common, and then check to see whether each $(k-1)$-itemset included in $I_1 \cup I_2$ belongs to P_{k-1}, which can be determined efficiently by organizing P_{k-1} as a hash tree structure. We repeat this process to create Q_k.

To compute P_k, we then scan all the transactions to calculate $u(I)$ and $chi(x(I), y(I))$ for each $I \in Q_k$, and set $\{I \in Q_k \mid u(I) \geq \tau\}$ to P_k. Figure 2 presents the overall algorithm, which we call *AprioriSMP* (Apriori with Statistical Metric Pruning).

Optimization Problem

To design an algorithm for the optimization problem, we slightly modify AprioriSMP for the item-wise enumeration problem in Figure 2. The key idea is that we use τ to store the temporarily maximum chi-squared value during the computation instead of the user-specified minimum chi-squared value. As before, an itemset I is defined *promising* if $u(I) \geq \tau$, because there is a possibility that a superset of I may give a chi-squared value no smaller than the temporarily maximum value τ.

Figure 3 presents the revision of AprioriSMP in Figure 2 according to the line outlined. We have underlined the three newly added statements. If X is empty, there is no promising itemset whose superset may give a chi-squared value larger than τ, which means that τ is guaranteed to be the maximum value. It is then relatively easy to modify the algorithm in Figure 3 so that it can list the most significant n itemsets.

4.4 Using Other Statistical Measures

There are other related statistical metrics commonly used for evaluating the correlation between the assumption and the conclusion of an association rule. We remark that it is relatively easy to modify AprioriSMP such that it can use three well-known statistical metrics: the entropy gain (mutual information), the gini index (mean squared error), and the correlation coefficient.

DEFINITION 4.8. *Recall the following contingency table in Definition 4.1.*

	C	$\bar{C}$	$\sum row$
I	$O_{IC} = y$	$O_{I\bar{C}}$	$O_I = x$
$\bar{I}$	$O_{\bar{I}C}$	$O_{\bar{I}\bar{C}}$	$O_{\bar{I}} = n - x$
$\sum column$	$O_C = m$	$O_{\bar{C}} = n - m$	n

Using the notations in the above table, we define the various statistical metrics.

Let $ent(p) = -p \ln p - (1-p)\ln(1-p)$. The entropy gain $Ent(x, y)$ is:

$$Ent(x, y) = ent(\frac{O_C}{n}) - \frac{O_I}{n} ent(\frac{O_{IC}}{O_I}) - \frac{O_{\bar{I}}}{n} ent(\frac{O_{\bar{I}C}}{O_{\bar{I}}}).$$

Let $gini(p) = 1 - p^2$. The Gini index $Gini(x, y)$ is:

$$Gini(x, y) = gini(\frac{O_C}{n}) - \frac{O_I}{n} gini(\frac{O_{IC}}{O_I}) - \frac{O_{\bar{I}}}{n} gini(\frac{O_{\bar{I}C}}{O_{\bar{I}}}).$$

Let $t \in D$. Let X be an itemset, and let X_t denote a variable such that $X_t = 1$ if $t \supseteq X$, and $X_t = 0$ otherwise. Let $\mu_I = O_I/n$ and $\mu_C = O_C/n$. The correlation coefficient $\rho(x, y)$ is:

$$\rho(x,y) = \frac{\sum_{t\in D}(I_t - \mu_I)(C_t - \mu_C)}{(\sum_{t\in D}(I_t - \mu_I)^2)^{1/2}(\sum_{t\in D}(C_t - \mu_C)^2)^{1/2}}.$$

□

Recall that in the design of the AprioriSMP algorithm, it is essential to use the convexity of the chi-squared function. It is known that $Ent(x, y)$ and $Gini(x, y)$ are convex functions [10; 17; 19]. Thus we can obtain AprioriSMP tailored to the entropy gain or the gini index by appropriately replacing the chi-squared function. Incidentally, $\rho(x, y)$ ranges from -1 to 1, and the absolute value of $\rho(x, y)$ shows the strength of the correlation between A and C. It is known that $chi(x, y) = n\rho(x, y)^2$ [24], and hence the optimal value of $|\rho(x, y)|$ can be immediately obtained from the optimal value of $chi(x, y)$.

Interclass Variance

It is interesting and also useful to associate with each transaction a numeric value such as height, weight, or balance. For instance, one may want to discover a significant association between an itemset of foods taken by people in a particular area and high blood pressure.

DEFINITION 4.9. *Let nu denote the numeric attribute of interest, and let $t[nu]$ denote nu's value associated with a transaction t.* □

We are interested in finding an itemset that is highly correlated with the numeric attribute *nu*. To measure the strength of the correlation, we can employ the *interclass variance*, which is frequently used in statistics.

DEFINITION 4.10. *Let D be the set of all transactions, and let X be a subset of D. Let $\mu(X)$ denote the averages of nu's values in X; namely,*

$$\mu(X) = \frac{\sum_{t\in X} t[nu]}{|X|}.$$

Let $D_I = \{t \in D \mid t \supseteq I\}$, and let $\bar{D_I} = D - D_I$. The interclass variance is

$$|D_I|(\mu(D) - \mu(D_I))^2 + |\bar{D_I}|(\mu(D) - \mu(\bar{D_I}))^2.$$

If D is fixed, the interclass variance is uniquely determined by $|D_I|$ and $\sum_{t\in D_I} t[nu]$. Let $x(I) = |D_I|$, and let $y(I) = \sum_{t\in D_I} t[nu]$. We will denote the above interclass variance by $var(x(I), y(I))$. □

Observe that if an itemset A is more correlated with the numeric item, the interclass variance gets larger. It is known that $var(x, y)$ is a convex function [19]. The following theorem presents a way of estimating a tight upper bound for $var(x(J), y(J))$ for any $J \supseteq I$.

THEOREM 4.3. *Divide $\{t \mid t \supseteq I\}$ into two disjoint subsets by a cutpoint $z \in \{t[nu] \mid t \supseteq I\}$, and define*

$$\begin{aligned} x_{<z} &= |\{t \mid t \supseteq I, t[nu] < z\}| \\ y_{<z} &= \sum\{t[nu] \mid t \supseteq I, t[nu] < z\} \\ x_{z\leq} &= |\{t \mid t \supseteq I, z \leq t[nu]\}| \\ y_{z\leq} &= \sum\{t[nu] \mid t \supseteq I, z \leq t[nu]\} \end{aligned}$$

For any $J \supseteq I$,

$$\begin{aligned} & var(x(J), y(J)) \\ \leq\ & \max \bigcup_{z\in\{t[nu]\mid t\supseteq I\}} \{var(x_{<z}, y_{<z}), var(x_{z\leq}, y_{z\leq})\} \end{aligned}$$

PROOF. This theorem can be proved according to the line of the proof of Theorem 4.1. First it is easy to show that

$$\bigcup_{z\in\{t[nu]\mid t\supseteq I\}} \{(x_{<z}, y_{<z}), (x_{z\leq}, y_{z\leq})\}$$

is a convex polygon that includes all the stamp points of the form $(x(J), y(J))$ for $J \supseteq I$. Now, $var(x, y)$ is a convex function, any convex function is maximized at one of the vertexes of a convex polygon, and thereby the objective inequality is proved. □

Because of the above theorem, it is straightforward to modify AprioriSMP to handle the interclass variance. Computing the upper bound is not costly, since the most time-consuming step is sorting $\{t[nu] \mid t \supseteq I\}$. To be more precise, we can compute $var(x_{<z}, y_{<z})$ and $var(x_{z\leq}, y_{z\leq})$ for each $z \in \{t[nu] \mid t \supseteq I\}$ by scanning the sorted list of $\{t[nu] \mid t \supseteq I\}$ just once.

5. EXPERIMENTAL RESULTS

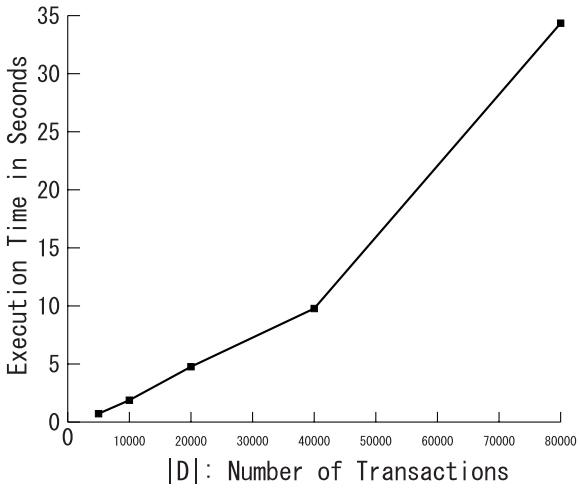
AprioriSMP has been implemented in C++. We evaluated the performance of AprioriSMP for the optimization problem on a single R10000 of SGI Origin 2000 with a CPU clock rate of 195MHz, 512MB of main memory, and running IRIX 6.5SE.

We generated a test dataset using the method introduced by Agrawal and Srikant [3]. Following [3], we use the following symbols:

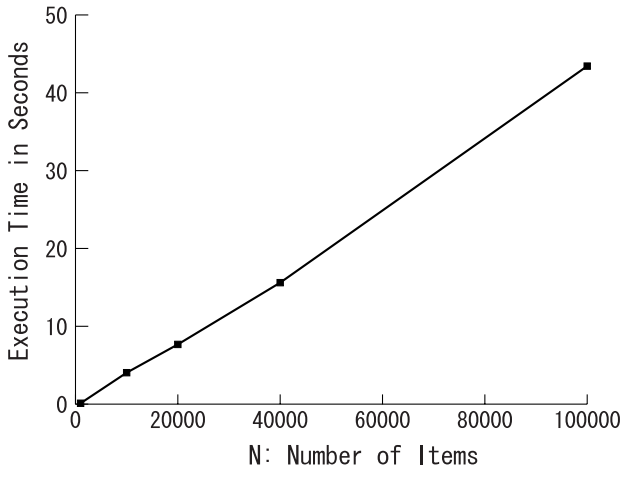
$\|D\|$:	the number of transactions
$\|T\|$:	the average size of transactions
$\|I\|$:	the average size of the maximal potentially large itemsets
$\|L\|$:	the number of maximal potentially large itemsets
N:	the number of items

We used parameters of $|T| = 20$, $|I| = 4$, and $|L| = 2000$.

To create an optimal association rule intentionally, we arbitrarily selected one maximal potentially large itemset, say X, and we doubled the probability that this itemset will be picked during the generation of the test dataset. The other itemsets are selected according to the method in [3].



(A)



(B)

Figure 4: Execution Times of AprioriSMP for the Optimization Problem

$\|D\|$	5000	10000	20000	40000	80000
$\|Q_1\|$	1000	1000	1000	1000	1000
$\|P_1\|$	522	935	976	383	576
$\|Q_2\|$	135981	436645	475800	73153	165600
$\|P_2\|$	3	3	3	3	3
$\|Q_3\|$	1	1	1	1	1
$\|P_3\|$	1	1	1	1	1

(A) Case when $N = 1000$

N	1000	10000	20000	40000	100000
$\|Q_1\|$	1000	10000	20000	40000	100000
$\|P_1\|$	407	145	97	99	261
$\|Q_2\|$	82621	10440	4656	4851	33930
$\|P_2\|$	3	3	3	3	3
$\|Q_3\|$	1	1	1	1	1
$\|P_3\|$	1	1	1	1	1

(B) Case when $|D| = 1000$

Table 2: Dramatic Effect of Statistical Metric Pruning for the Optimization Problem

We then selected an item c in X as the objective item, and thereby made $(X - \{c\}) \Rightarrow \{c\}$ the optimal association rule.

We have first observed the behavior of the execution time when we increase the number of transactions while we have fixed $N = 1000$. Figure 4 (A) shows the execution time scales almost linearly with the number of transactions. We have also observed the execution time for various numbers of items N while we have fixed $|D| = 1000$, which is illustrated in Figure 4 (B). The execution time also scales almost linearly with the number of items. Table 2 presents numbers of potentially promising itemsets $|Q_k|$ and numbers of promising itemsets $|P_k|$ that were generated during the computation. Note that $|Q_2| \gg |P_2| = 3$ in all cases, which shows the dramatic effect of the statistical metric pruning.

Although AprioriSMP does not use the information of supports of itemsets for pruning unproductive itemsets, one may want to know supports of itemsets generated during the computation. For each Q_k and each P_k, we kept the record of the minimum support, $\min\{Pr(I \cup \{c\}) \mid I \in Q_k (\text{or } P_k)\}$, and Table 3 presents those minimum supports. For instance, in Table 3 (A), when $|D| = 5000$, we see:

$$\begin{aligned} \min\{Pr(I \cup \{c\}) \mid I \in Q_1\} &= 0.02\% \\ \min\{Pr(I \cup \{c\}) \mid I \in P_1\} &= 0.02\% \\ \min\{Pr(I \cup \{c\}) \mid I \in Q_2\} &= 0\% \\ \min\{Pr(I \cup \{c\}) \mid I \in P_2\} &= 9.52\% \end{aligned}$$

Observe that AprioriSMP had to examine promising or potentially promising itemsets with very small supports in its earlier steps for computing the optimal itemset in the final step, even if the support of the optimal itemset is relatively high.

6. CONCLUSION

We have discussed how to efficiently calculate significant association rules according to common statistical measures. We have shown that the Apriori algorithm combined with the novel technique of pruning via statistical metric presents an efficient solution to this problem. The major advantage of AprioriSMP is its avoidance of the use of higher support thresholds that has been believed to be requisite for the application of Apriori.

We are now improving our current implementation so that the computation time scales for millions of transactions. Actually, we store transactions in the format of a two-dimensional table that could be often sparse. We will be able to reduce the size of the sparse table just by representing the table in a condensed set format.

Finding correlation between itemsets would be applicable to various problems. We have been using this technique to the analysis of correlation between multiple genotypes and the objective phenotype of interest [21].

$\|D\|$	5000	10000	20000	40000	80000
Q_1	0.02%	0.06%	0.1%	0.17%	0.14%
P_1	0.02%	0.06%	0.1%	0.17%	0.14%
Q_2	0%	0%	0%	0%	0%
P_2	9.52%	9.29%	9.95%	9.77%	9.76%
Q_3	9.5%	9.29%	9.95%	9.77%	9.74%
P_3	9.5%	9.29%	9.95%	9.77%	9.74%

(A) Case when $N = 1000$

N	1000	10000	20000	40000	100000
Q_1	0%	0%	0%	0%	0%
P_1	0%	0%	0%	0%	0%
Q_2	0%	0%	0%	0%	0%
P_2	8.8%	8.0%	9.3%	9.3%	10.8%
Q_3	8.8%	8.0%	9.3%	9.3%	10.8%
P_3	8.8%	8.0%	9.3%	9.3%	10.8%

(B) Case when $|D| = 1000$

Table 3: Minimum Support

Acknowledgement

The first author is supported in part by a Grant-in-Aid for Scientific Research (B) 10480066 and a Grant-in-Aid for Scientific Research on Priority Areas 10143102 from the Ministry of Education, Science, and Culture Japan. We thank Akihiro Nakaya and Osamu Watanabe for their valuable input.

7. REFERENCES

[1] C. C. Aggarwal and P. S. Yu. A new framework for itemset generation. In *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 1-3, 1998, Seattle, Washington*, pages 18–24. ACM Press, 1998.

[2] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., May 26-28, 1993*, pages 207–216. ACM Press, 1993.

[3] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*, pages 487–499. Morgan Kaufmann, 1994.

[4] R. J. Bayardo Jr. Efficiently mining long patterns from databases. In *SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA*, pages 85–93. ACM Press, 1998.

[5] R. J. Bayardo Jr. and R. Agrawal. Mining the most interesting rules. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 15-18 August 1999, San Diego, CA USA*, pages 145–154. ACM Press, 1999.

[6] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, Tucson, Arizona, USA*, pages 265–276. ACM Press, 1997.

[7] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, Tucson, Arizona, USA*, pages 255–264. ACM Press, 1997.

[8] S. Brin, R. Rastogi, and K. Shim. Mining optimized gain rules for numeric attributes. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 15-18 August 1999, San Diego, CA USA*, pages 135–144. ACM Press, 1999.

[9] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama. Constructing efficient decision trees by using optimized association rules. In *Proceedings of the 22nd VLDB Conference*, pages 146–155, Sept. 1996.

[10] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama. Constructing efficient decision trees by using optimized numeric association rules. In *VLDB'96, Proceedings of 22th International Conference on Very Large Data Bases, September 3-6, 1996, Mumbai (Bombay), India*, pages 146–155. Morgan Kaufmann, 1996.

[11] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama. Data mining using two-dimensional optimized accociation rules: Scheme, algorithms, and visualization. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Quebec, Canada, June 4-6, 1996*, pages 13–23. ACM Press, 1996.

[12] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama. Mining optimized association rules for numeric attributes. In *Proceedings of the Fifteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 3-5, 1996, Montreal, Canada*, pages 182–191. ACM Press, 1996.

[13] M. R. Garey and D. S. Johnson. *Computer and Intractability. A Guide to NP-Completeness.* W. H. Freeman, 1979.

[14] R. Horst and H. Tuy. *Global Optimization – Deterministic Approaches.* Springer, 1996.

[15] L. V. S. Lakshmanan, R. T. Ng, J. Han, and A. Pang. Optimization of constrained frequent set queries with 2-variable constraints. In *SIGMOD 1999, Proceedings ACM SIGMOD International Conference on Management of Data, June 1-3, 1999, Philadephia, Pennsylvania, USA*, pages 157–168. ACM Press, 1999.

[16] B. Liu, W. Hsu, and Y. Ma. Pruning and summarizing the discovered associations. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 15-18 August*

1999, San Diego, CA USA, pages 125–134. ACM Press, 1999.

[17] Y. Morimoto, T. Fukuda, H. Matsuzawa, T. Tokuyama, and K. Yoda. Algorithms for mining association rules for binary segmentations of huge categorical databases. In *Proceedings of VLDB 1998*, pages 380–391, Aug. 1998.

[18] Y. Morimoto, H. Ishii, and S. Morishita. Efficient construction of regression trees with range and region splitting. In *Proceedings of the 23rd VLDB Conference*, pages 166–175, Aug. 1997.

[19] S. Morishita. On classification and regression. In *Proceedings of Discovery Science, First International Conference, DS'98 — Lecture Notes in Artificial Intelligence*, volume 1532, pages 40–57, Dec. 1998.

[20] S. Morishita and A. Nakaya. Parallel branch-and-bound graph search for correlated association rules. In *Proceedings of ACM SIGKDD Workshop on Large-Scale Parallel KDD Systems*, Aug. 1999.

[21] A. Nakaya, H. Hishigaki, and S. Morishita. Mining the quantitative trait loci associated with oral glucose tolerance in the oletf rat. In *Proc. of Pacific Symposium on Biocomputing*, pages 367–379, Jan. 2000.

[22] R. T. Ng, L. V. S. Lakshmanan, J. Han, and A. Pang. Exploratory mining and pruning optimizations of constrained association rules. In *SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA*, pages 13–24. ACM Press, 1998.

[23] J. S. Park, M.-S. Chen, and P. S. Yu. An effective hash based algorithm for mining association rules. In *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, San Jose, California, May 22-25, 1995*, pages 175–186. ACM Press, 1995.

[24] C. Silverstein, S. Brin, R. Motwani, and J. D. Ullman. Scalable techniques for mining causal structures. In *VLDB'98, Proceedings of 24rd International Conference on Very Large Data Bases, August 24-27, 1998, New York City, New York, USA*, pages 594–605. Morgan Kaufmann, 1998.

[25] D. Tsur, J. D. Ullman, S. Abiteboul, C. Clifton, R. Motwani, S. Nestorov, and A. Rosenthal. Query flocks: A generalization of association-rule mining. In *SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA*, pages 1–12. ACM Press, 1998.

APPENDIX

To try to make this paper self-contained, we here present the proofs of two propositions used in this paper. Theorems closely related with the propositions have been proved in other publications.

A. PROOF OF PROPOSITION 4.1

$chi(x, y)$ is a convex function.

This property is also shown in [17]. Proofs for $Ent(x, y)$, $Gini(x, y)$, and $var(x, y)$ can be found in [9; 19], [17], and [18; 19] respectively. Technically, those proofs can be shown in a similar way.

Recall the following contingency table and the definition of $chi(x, y)$ given in Definition 4.1.

	C	$\bar{C}$	$\sum row$
I	$O_{IC} = y$	$O_{I\bar{C}}$	$O_I = x$
$\bar{I}$	$O_{\bar{I}C}$	$O_{\bar{I}\bar{C}}$	$O_{\bar{I}} = n - x$
$\sum column$	$O_C = m$	$O_{\bar{C}} = n - m$	n

$$chi(x, y) = \sum_{i \in \{I, \bar{I}\}, j \in \{C, \bar{C}\}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where

$$E_{ij} = n \times (O_i/n) \times (O_j/n).$$

To order to prove the convexity of $chi(x, y)$, it is sufficient to show that for any real numbers δ_1 and δ_2, and $V = \delta_1 x + \delta_2 y$,

$$\partial^2 chi(x, y)/\partial V^2 \geq 0. \tag{1}$$

Focus on the case when $i = I$, and define

$$f(x, y) = \sum_{i = I, j \in \{C, \bar{C}\}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

Then,

$$chi(x, y) = f(x, y) + f(n - x, m - y).$$

To prove (1), it is sufficient to show

$$\partial^2 f(x, y)/\partial V^2 \geq 0$$

$$\partial^2 f(n - x, m - y)/\partial V^2 \geq 0.$$

We will prove the former inequality. The latter can be proved in a similar way.

$$\begin{aligned} f(x, y) &= \frac{(O_{IC} - E_{IC})^2}{E_{IC}} + \frac{(O_{I\bar{C}} - E_{I\bar{C}})^2}{E_{I\bar{C}}} \\ &= \frac{(y - \frac{xm}{n})^2}{\frac{xm}{n}} + \frac{(x - y - \frac{x(n-m)}{n})^2}{\frac{x(n-m)}{n}} \\ &= \frac{(ny - mx)^2}{x} \cdot \frac{1}{m(n-m)} \end{aligned} \tag{2}$$

Define

$$f_1(x, y) = \frac{(ny - mx)^2}{x}.$$

Then,

$$\partial^2 f(x, y)/\partial V^2 \geq 0 \text{ iff } \partial^2 f_1(x, y)/\partial V^2 \geq 0.$$

We will prove the latter inequality for $\delta_1 \neq 0$ and $\delta_2 \neq 0$.

$$\begin{aligned}
\frac{\partial f_1(x,y)}{\partial V} &= \frac{\partial f_1}{\partial x}\frac{\partial x}{\partial V} + \frac{\partial f_1}{\partial y}\frac{\partial y}{\partial V} \\
&= \frac{m^2x^2 - n^2y^2}{\delta_1 x^2} + \frac{2(ny-mx)n}{\delta_2 x} \\
&= -\frac{n^2}{\delta_1}(\frac{y}{x})^2 + \frac{2n^2}{\delta_2}(\frac{y}{x}) + constant \\
\frac{\partial^2 f_1(x,y)}{\partial V^2} &= \frac{2n^2}{x^3}(\frac{y}{\delta_1} - \frac{x}{\delta_2})^2 \\
&\geq 0
\end{aligned}$$

The cases when $\delta_1 = 0$ or $\delta_2 = 0$ can be proved similarly.

$\lim_{x\to 0} chi(x,y) = 0$

$$\begin{aligned}
chi(x,y) &= f(x,y) + f(n-x, m-y) \\
&= (ny-mx)^2(\frac{1}{x} + \frac{1}{n-x})\frac{1}{m(n-m)} \\
&= (n\frac{y}{x} - m)^2 x \frac{n}{n-x}\frac{1}{m(n-m)}
\end{aligned}$$

Since $0 \leq y \leq x$, $\frac{y}{x} \leq 1$, and hence $\lim_{x\to 0} chi(x,y) = 0$.

$\lim_{x\to n} chi(x,y) = 0$

$$chi(x,y) = (n\frac{m-y}{n-x} - m)^2(n-x)\frac{n}{x}\frac{1}{m(n-m)}$$

Since $0 \leq m-y \leq n-x$, $\frac{m-y}{n-x} \leq 1$, $\lim_{x\to n} chi(x,y) = 0$.

$chi(x,y)$ is minimum if $y = (m/n)x$.

Since $chi(x,y) = f(x,y) + f(n-x, m-y)$,

$$chi(x,y) = chi(n-x, m-y).$$

From the convexity of $chi(x,y)$,

$$\begin{aligned}
chi(x,y) &= \frac{1}{2}chi(x,y) + \frac{1}{2}chi(n-x, m-y) \\
&\geq chi(n/2, m/2),
\end{aligned}$$

and hence $chi(x,y)$ is minimum at $(n/2, m/2)$. From the definition of $f(x,y)$ (see (2)),

$$f(n/2, m/2) = 0 \quad \text{and} \quad f(x, (m/n)x) = 0.$$

Since $chi(x,y) = f(x,y) + f(n-x, m-y)$,

$$chi(n/2, m/2) = chi(x, (m/n)x) = 0,$$

which completes the proof.

B. NP-HARDNESS OF THE OPTIMIZATION PROBLEM

The case when the statistical measure is the entropy gain is proved in [19], and the argument carries over to the optimization problem for the chi-squared function.

THEOREM B.1. *It is NP-hard to compute the optimal association rule of the form $I \Rightarrow C$, where C is a fixed itemset.*

PROOF. We reduce the difficulty of the problem to the NP-hardness of finding the minimum cover [13]. Let (V, E) be a hypergraph such that V denotes the set of vertexes and E shows the set of hyperedges. $e \in E$ is a subset of V that may contain more than two vertexes in it. A subset E' of E is called a *cover* if every vertex in V belongs to one of hyperedges in E; that is,

$$\cup_{e\in E'} e = V.$$

A *minimum cover* minimizes the number of hyperedges in it among all covers. It is known that computing a minimum cover is NP-hard [13].

In what follows, we construct a set of transactions such that the optimal association rule presents a minimum set cover of (V, E). First, with each hyperedge $e \in E$, we associate a unique item, denoted by i_e, that is newly created. We also introduce an item of special named c. We will denote the set of all items by S; that is,

$$S = \{i_e \mid e \in E\} \cup \{c\}.$$

We now present how to generate a set of transactions using one of the following three rules:

1. With each $v \in V$, associate a transaction t_v such that
$$t_v = S - \{i_e \mid v \in e\}.$$
Let $\bar{v}$ be an arbitrary vertex in the complement of e; that is, $\bar{v} \in V - e$. Then, the transaction $t_{\bar{v}}$ corresponding to $\bar{v}$ contains the item i_e. t_v contains c. This rule generates $|V|$ transactions.
2. With each $e \in E$, associate a new transaction t_e such that
$$t_e = S - \{i_e, c\}.$$
This rule creates $|E|$ transactions.
3. If $|E| < |V|$, generate $(|V| - |E| + 1)$ distinct transactions each of which is $S - \{c\}$.

Let n denote the number of all the transactions. In what follows; we call a transaction t *positive* if t contains c, while we call t *negative* otherwise. Let m denote the number of all positive transactions, and then the number of all negative transactions is $n-m$. Note that all the positive transactions are generated by the first rule, and hence $|V| = m$. On the other hand, all the negative transactions are constructed by the second or the third rule. The third rule is introduced to guarantee that there are more negative transactions than positive ones; namely,

$$n - m > m.$$

Figure 5 presents a way of understanding the above construction of transactions and itemsets from a hypergraph in a visual manner. Figure 5 (A) shows a hypergraph. Its hyperedges are $\{e1, e2, e3, e4, e5, e6\}$. We regard each hyperedge as an itemset and each vertex as a transaction. In Figure 5(B), black points show positive transactions, and white points are negative transactions. For instance, black points enclosed in the hyperedge $e1$ are transactions that do not contain i_{e1}, The white point in e_1 shows the itemset $S - \{i_{e1}, i_c\}$, which again does not include i_{e1}. On the

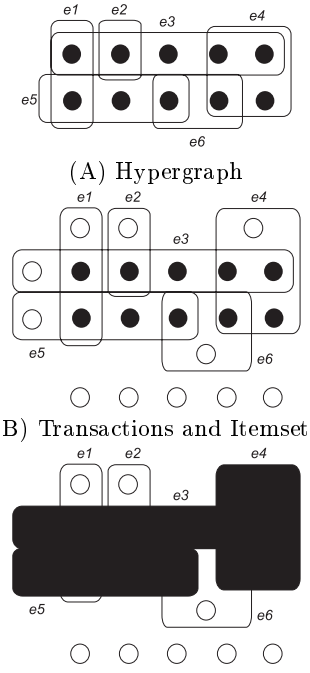


Figure 5: Creating Transactions and Itemsets from a Hypergraph

other hand, the complement of $e1$ includes all the transactions that contain i_{e1}. White points that are outside of all the hyperedges are created by the third rule. In Figure 5 (C), hyperedges $e3, e4$, and $e5$ are painted black, and the remaining eight white points correspond to the transactions that include all items in $\{i_{e3}, i_{e4}, i_{e5}\}$.

We will identify a set of hyperedges E' with the following itemset I':

$$I' = \{i_e \mid e \in E'\}.$$

Suppose that $I \Rightarrow \{c\}$ maximizes the chi-squared value. We will show that a minimum cover can be created from I by using the above equation. Put another way, computing the optimal itemset immediately presents a minimum cover.

Let E^* be a minimum cover, and let $I^* = \{i_e \mid e \in E^*\}$. For instance, in Figure 5 (A), $\{e3, e4, e5\}$ is a minimum cover, and its corresponding itemset is $\{i_{e3}, i_{e4}, i_{e5}\}$. Following the notations defined in Definition 4.1, let $x(I^*)$ denote the number of transactions that include I^*, and let $y(I^*)$ note the number of transaction that contain both I^* and $\{c\}$. Since E^* is a cover, for any vertex v, there must exist $e \in E^*$ such that $v \in e$. Then, t_v does not contain i_e, and therefore $t_v \not\supseteq I^*$. This implies that no positive transaction includes I^*, and hence $y(I^*) = 0$. For any $e \in E^*$, t_e does not include i_e, and $t_e \not\supseteq I^*$. On the other hand, for $e \notin E^*$, $t_e \supseteq I^*$, and $S \supseteq I^*$. Thus, $x(I^*) = n - m - |E^*|$. Let k denote $|E^*|$. Then,

$$x(I^*) = n - m - k.$$

In Figure 5 (C), the eight white points correspond to transactions including $\{i_{e3}, i_{e4}, i_{e5}\}$ that is generated from the minimum cover $\{e3, e4, e5\}$. Thus we have the following equation:

$$x(\{i_{e3}, i_{e4}, i_{e5}\}) = 21 - 10 - 3 = 8.$$

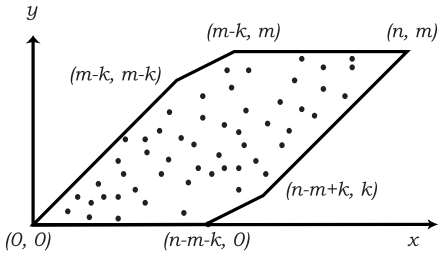


Figure 6: Stamp Points and NP-hardness

As we have done in the proof of Theorem 4.1, with each itemset I, we associate the stamp point $(x(I), y(I))$. Figure 6 shows all the stamp points. It is easy to see that all the stamp points are mapped onto the hexagon of $(0,0)$, $(m-k, m-k)$, $(m+k, m)$, (n, m), $(n-m+k, k)$, and $(n-m-k, 0)$. I^* is associated with $(n-m-k, 0)$. Since $chi(x, y)$ is a convex function, $chi(x, y)$ is maximized at one of the six vertexes. From Proposition 4.1, $chi(x, y)$ is minimum at $(0, 0)$ and (n, m). Since $chi(x, y) = chi(n - x, m - y)$,

$$\begin{aligned} chi(m+k, m) &= chi(n-m-k, 0) \\ chi(m-k, m-k) &= chi(n-m+k, k) \end{aligned}$$

To prove that $chi(x, y)$ is maximum at $(n - m - k, 0)$, it is sufficient to show that

$$chi(m+k, m) > chi(m-k, m-k).$$

Recall

$$chi(x, y) = \frac{(ny - mx)^2}{x(n-x)} \cdot \frac{n}{m(n-m)}.$$

Define

$$g(x, y) = \frac{(ny - mx)^2}{x(n-x)},$$

then it remains to show $g(m+k, m) > g(m-k, m-k)$.

$$\begin{aligned} g(m+k, m) &= \frac{m^2(n-m-k)}{m+k} \\ g(m-k, m-k) &= \frac{(n-m)^2(m-k)}{n-m+k} \end{aligned}$$

Then,

$$\begin{aligned} & g(m+k, m) - g(m-k, m-k) \\ = \; & \frac{k^2((n-m)^2 - m^2)}{(m+k)(n-m+k)} > 0, \end{aligned}$$

because $n - m > m$. We show that no itemset is associated with $(m+k, m)$. Let J be an arbitrary non-empty itemset. For any $i_e \in J$, e must contain at least one vertex, say v. Since t_v does not contain i_e, $t_v \not\supseteq J$, and hence $y(J) < m$.

Consequently, among all stamp points, $(n-m-k, 0)$ is the unique point that maximizes $chi(x, y)$. From every itemset associated with $(n-m-k, 0)$, we can immediately create a minimum cover. □