<table>
<thead>
<tr>
<th>Title</th>
<th>Probabilistic Analyses on the Number of Reliable Rules and the Needed Data Size (Mathematical Models and Decision Making under Uncertainty)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Haraguchi, Kazuya; Yagiura, Mutsunori</td>
</tr>
<tr>
<td>Citation</td>
<td>数理解析研究所講究録 (2006), 1477: 223-231</td>
</tr>
<tr>
<td>Issue Date</td>
<td>2006-03</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/2433/48256">http://hdl.handle.net/2433/48256</a></td>
</tr>
<tr>
<td>Type</td>
<td>Departmental Bulletin Paper</td>
</tr>
<tr>
<td>Textversion</td>
<td>publisher</td>
</tr>
</tbody>
</table>

Kyoto University
有効要素数の確率的解析
(Probabilistic Analyses on the Number of Reliable Rules and the Needed Data Size)

原口 和也 (Kazuya Haraguchi) * 柳浦 亮憲 (Mutsunori Yagiura) †

Abstract

Suppose that we are given a data set of examples, where each example is an n-dimensional
Boolean vector and labeled either true or false. A pattern \( r = (J, b) \) is defined by a subset
\( J \subseteq \{1, \ldots, n\} \) of the \( n \) Boolean variables and a Boolean vector \( b \in \{0, 1\}^J \) of the variables in \( J \).
If \( r \) appears frequently in the true examples and infrequently in the false examples, we call \( r \) a
good rule. In this paper, we consider how many examples are needed for generating “reliable”
good rules, in the sense that they capture the essential properties of the data domain. Suppose
the random data domain where all examples in \( \{0, 1\}^n \) are uniformly distributed and labeled
at random. A small random data set may contain good rules superficially, although there is no
property in the data domain. Our claim is that the data set should contain sufficiently many
examples to avoid such deceptive good rules existing even in a random data set. We make
probabilistic analyses to estimate such amounts of examples, and show experimental studies
to justify our claim.

Keywords: frequent/infrequent item sets, association rules, knowledge discovery, probabilistic analysis.

1 Introduction

Assume that we are given a data set \( X \) of examples. Each example in \( X \) is an \( n \)-dimensional
Boolean vector, and is labeled either 1 (true) or 0 (false). We denote by \( X_1 \) (resp., \( X_0 \)) the set of
true (resp., false) examples in \( X \). (Hence, \( X = X_1 \cup X_0 \).) Let us denote \( B = \{0, 1\} \). A pattern
\( r = (J, b) \) is defined by a subset \( J \subseteq \{1, \ldots, n\} \) of the \( n \) Boolean variables and a Boolean vector
\( b \in B^J \) of the variables in \( J \). For a pattern \( r = (J, b) \) and a Boolean vector \( x \in B^n \), we say that
\( r \) appears in \( x \) if \( x|_J = b \) holds. Let \( X(r) \) denote the set of examples in \( X \) in which \( r \) appears;
\( i.e., \), \( X(r) = \{ x \in X | x|_J = b \} \). Note that \( B^J(r) \) is defined similarly. We define the frequency
\( f(r, X) = |X(r)|/|X| \), which is the ratio of examples in \( X \) in which \( r \) appears. For a constant
\( a \) (\( 0 \leq a \leq 1 \)), if \( f(r, X) \geq a \) (resp., \( f(r, X) \leq a \)) holds, then we call \( r \) an \( a \)-frequent (resp., an \( a \)-infrequent) pattern in \( X \).

The generation of frequent/infrequent patterns is an important issue in such fields as data
mining and bioinformatics (e.g., knowledge discovery from genome databases) [1, 4, 11]. (The
term “frequent/infrequent set” is widely used in the literature, but in order to avoid the confusion
with a simple set of elements, we use the term “pattern” in this paper.) It is well-known that
one can generate frequent/infrequent patterns in incremental polynomial time [1], and many fast
algorithms for this task have been proposed so far (e.g., [10]).

For constants \( a_1, a_0 \) (\( 0 \leq a_1, a_0 \leq 1 \)), if \( f(r, X_1) \geq a_1 \) and \( f(r, X_0) \leq a_0 \) hold, then we call \( r \)
an \((a_1, a_0)\)-good rule in \( X \). Such an \( r \) is considered to describe a feature in true examples (under
reasonable \( a_1 \) and \( a_0 \); e.g., \( a_1 \gg a_0 \)). When \( |X_1| \) and \( |X_0| \) are small, even a random data set \( X \)

*京都大学大学院 數值理学研究科 數理工学専攻 (Department of Applied Mathematics and Physics, Graduate School of
Informatics, Kyoto University) e-mail: kazuyah@amp.i.kyoto-u.ac.jp
†名古屋大学大学院 情報科学研究所 計算機理学専攻 (Department of Computer Science and Mathematical Informatics, Graduate School of Information Science, Nagoya University) e-mail: yagiura@nagoya-u.jp
may contain many good rules that have nothing to do with the inherent structure of $X$ and are deceptive. In this paper, we consider how many examples should be collected or sampled as the data set for generating "reliable" good rules, avoiding such deceptive ones.

We estimate such amounts through a probabilistic analysis on random data sets. Suppose a data domain where an example is drawn from $\mathcal{B}^n$ with the uniform probability (i.e., $1/2^n$), and is labeled either 1 or 0 at random. Essentially, there is no pattern that describes inherent information of the data set. However, if the given random data set $X$ contains insufficient examples, some patterns may happen to become good rules due to a bias peculiar to $X$; on the other hand, if $X$ contains sufficiently many examples, good rules will exist very rarely. We analyze upper bounds on the expected number of $(a_1, a_0)$-good rules in a random data set consisting of $m_1$ true examples and $m_0$ false examples, and show that it becomes close to 0 if $m_1$ and $m_0$ are larger than thresholds. We claim that such thresholds give rough estimates on the number of true and false examples needed to extract reliable good rules from a real data set. We then give some experimental results to justify our claim based on the random data analysis.

The problem is closely related to the problem of finding association rules. An association rule is defined by two patterns $(r, r') = ((J, b), (J', b'))$ with $J \cap J' = \emptyset$: it represents that an example $x$ with $x|_J = b$ tends to attain $x|_{J'} = b'$. Patterns in this paper may be regarded as special cases of association rules such that the labels of examples are attached to the original data set as the $(n+1)^{\text{st}}$ Boolean variable and the threshold is limited to $r' = \{(n+1), (1)\}$. An association rule $(r, r')$ is usually evaluated by support (which is the frequency of $r$ in $X$) and confidence (which is the frequency of $r'$ in $X(r)$), while we evaluate a pattern $r$ by its frequency in $X_1$ and infrequency in $X_0$. Thus, the generation of frequent patterns corresponds to finding association rules of a basic form. This task on a huge data set is too time-consuming, and Li et al. [7] and Toivonen [8] discussed the proper size of a subset of a huge data set $X$ from which frequent patterns are generated. Suppose a subset $X' \subseteq X$ of examples and a pattern $r$. They consider how many examples are needed as $X'$ for $f(r, X')$ to be close enough to $f(r, X)$ with high probability, if $X'$ is randomly selected from $X$.

While they consider the random sampling from the given data set to deal with the situation where the size of the given data set is huge (i.e., their objective is to approximate the given data set with a sample of manageable size), we consider the situation where the size of the given data set is small and our objective is to judge whether the extracted good rules are reliable or not. This is the main difference of our approach from the existing ones.

## 2 Probabilistic Analyses

### 2.1 Preliminaries

We first describe the assumption on the generation of examples.

**Assumption 1** The generation of examples is mutually independent. An example $(x, \omega)$ is generated by the following process:

**Step 1:** The label $\omega$ is set to 1 with probability $\rho$ ($0 \leq \rho \leq 1$), and to 0 otherwise (i.e., with probability $1 - \rho$).

**Step 2:** Let $P_1, P_0 : \mathcal{B}^n \rightarrow [0, 1]$ denote probability distributions. The vector $x$ is drawn according to the distribution $P_\omega$.

Since $P_1$ and $P_0$ are probability distributions,

\[
\sum_{x \in \mathcal{B}^n} P_1(x) = \sum_{x \in \mathcal{B}^n} P_0(x) = 1. \tag{1}
\]

Consider a pattern $r$. Under the condition that a generated example is labeled 1 (resp., 0) in Step 1 of Assumption 1, the probability $c_1(r)$ (resp., $c_0(r)$) that $r$ appears in this new example is:

\[
c_1(r) = \sum_{x \in \mathcal{B}^n(r)} P_1(x) \quad \text{(resp., } c_0(r) = \sum_{x \in \mathcal{B}^n(r)} P_0(x) \text{).} \tag{2}
\]
More generally, under the condition that $m_1$ examples are labeled $1$ and $m_0$ examples are labeled $0$, the probability that $r$ is $a_1$-frequent in the $m_1$ true examples is:

$$U(m_1, a_1, c_1(r)) = \sum_{s=[a_1 m_1]}^{m_1} \binom{m_1}{s} c_1(r)^s (1 - c_1(r))^{m_1-s},$$

and the probability that $r$ is $a_0$-infrequent in the $m_0$ false examples is:

$$L(m_0, a_0, c_0(r)) = \sum_{s=0}^{m_0} \binom{m_0}{s} c_0(r)^s (1 - c_0(r))^{m_0-s}.$$

Note that $U$ (resp., $UL$) is also the expectation that $r$ is an $a_1$-frequent pattern in the true examples (resp., an $(a_1, a_0)$-good rule in the true and false examples).

For a pattern $r = (J, b)$, let us call the cardinality $|J|$ the size of $r$. We denote by $R_k$ the set of all possible patterns of size $k$ ($1 \leq k \leq n$). Note that $|R_k| = 2^{n-k}$ holds and that $|\mathcal{B}^n(r)| = 2^{n-k}$ holds for any $r \in R_k$. Let $E(n, m_1, a_1)$ (resp., $E^*(n, m_1, m_0, a_1, a_0)$) be the expected number of $a_1$-frequent patterns (resp., $(a_1, a_0)$-good rules), and $E_k(n, m_1, a_1)$ (resp., $E_k^*(n, m_1, m_0, a_1, a_0)$) be the expectations $U$ (resp., $UL$) of those of size $k$. From the linearity of expectations, they are computed as follows:

$$E(n, m_1, a_1) = \sum_{k=1}^{n} E_k(n, m_1, a_1)$$

$$= \sum_{k=1}^{n} \sum_{r \in R_k} U(m_1, a_1, c_1(r)),$$

$$E^*(n, m_1, m_0, a_1, a_0) = \sum_{k=1}^{n} E_k^*(n, m_1, m_0, a_1, a_0)$$

$$= \sum_{k=1}^{n} \sum_{r \in R_k} U(m_1, a_1, c_1(r))L(m_0, a_0, c_0(r)).$$

Suppose that the size $k$ of a pattern $r$ is large. From $|\mathcal{B}^n(r)| = 2^{n-k}$, $r$ appears in a small portion of vectors in $\mathcal{B}^n$, and thus may not be frequent in the given true examples. Then, it is anticipated that $E_k$ and $E_k^*$ with large $k$ are close to 0. On the other hand, a pattern $r$ of a small size $k$ appears in many vectors in $\mathcal{B}^n$, and thus may not be infrequent in the given false examples. It is therefore anticipated that $E_k^*$ with small $k$ is close to 0. Hence, if $E_k^*$ is close to 0 for all $k = 1, \ldots, n$, then $E^*$ will also be close to 0. In the next subsection, we show that it surely holds in the random data under some conditions.

For the analysis, we need the following assumption on $P_1$ and $P_0$.

Assumption 2 For any $x \in \mathcal{B}^n$, $P_1(x) \leq p$ and $P_0(x) \geq q$ hold for some constants $p$ and $q$.

From (1), it is implied that $p \geq 1/2^n$ and $q \leq 1/2^n$. Note that the above assumption enables us to cover various distributions including the uniform distribution (which is realized by setting $p = q = 1/2^n$).

### 2.2 Upper Bounds on $E_k$ and $E_k^*$

We first introduce some well-known bounds in the probability theory.

**Theorem 1 (Chernoff [3])** Given a positive integer $m$ and $0 \leq \mu \leq 1$, let $Y_i$ be a random variable taking the value as follows:

$$Y_i = \begin{cases} 1 - \mu & \text{with probability } \mu, \\ -\mu & \text{with probability } 1 - \mu. \end{cases}$$

...
and let $Y = \sum_{i=1}^{m} Y_i$. Then, for any $\beta > 1$,
$$\Pr(Y \geq (\beta - 1)\mu m) < (\exp(\beta - 1)\beta^{-\beta})^{\mu m}$$ \hfill (5)
holds.

**Theorem 2** (Hoeffding [6]) For a positive integer $m$ and $0 \leq a \leq 1$, if $0 \leq \mu \leq a$, then
$$U(m, a, \mu) \leq \exp(-2m(a - \mu)^2).$$ \hfill (6)
If $a \leq \mu \leq 1$, then
$$L(m, a, \mu) \leq \exp(-2m(\mu - a)^2).$$ \hfill (7)

Variations of Theorem 1 are found in [2], for example.

Now we show two types of upper bounds on $E_k$ (and thus $E_k^*$) for “large” $k$.

**Theorem 3** For given parameters $n$, $a_1$ and $p$, and for any $\epsilon \in (0, 1]$, if $k \geq k_1$ and $m_1 \geq M_1$, then $E_k(n, m_1, a_1) \leq \epsilon$ holds, where
$$k_1 = n - \log_2 \frac{a_1}{e^2 p}, \quad M_1 = \frac{n \ln(2n) - \ln \epsilon}{a_1},$$
and $e$ denotes the base of the natural logarithm.

**Proof.** Let $r$ be a pattern of size $k \geq k_1$. From Assumption 2 and $|B^a(r)| = 2^{n-k}$, $c_1(r) \leq \min\{1, 2^{n-k}\}$ holds; now since $2^{n-k} \leq a_1/\epsilon^2 p$, $c_1(r) \leq 2^{n-k}p \leq a_1/\epsilon^2 < 1$ holds. Let $Z_i$ be a random variable taking the value as follows:
$$Z_i = \begin{cases} 1 & \text{with probability } 2^{n-k}p, \\ 0 & \text{with probability } 1 - 2^{n-k}p. \end{cases}$$ \hfill (8)
and let $Z = \sum_{i=1}^{m_1} Z_i$. Let $Y_i = Z_i - 2^{n-k}p$ and $Y = \sum_{i=1}^{m_1} Y_i = Z - 2^{n-k}pm_1$. Then, we have
$$E_k(n, m_1, a_1) = \sum_{r \in R_k} U(m_1, a_1, c_1(r)) \leq U(m_1, a_1, 2^{n-k}p) \times |R_k| = \Pr(Z \geq a_1 m_1) \times 2^k \binom{n}{k} = \Pr\left(\sum_{i=1}^{m_1} Z_i \geq a_1 m_1 \right) \times 2^k \binom{n}{k}.$$
From $k \geq k_1$, $a_1/(2^{n-k}p) \geq \epsilon^2 > 1$. By applying Theorem 1 with $m = m_1$, $\mu = 2^{n-k}p$ and $\beta = a_1/(2^{n-k}p)$, we have
$$E_k(n, m_1, a_1) \leq \left(\frac{2^{n-k}p}{a_1}\right)^{a_1 m_1} \times 2^k \binom{n}{k} \leq e^{a_1 m_1} \times (2n)^n.$$ \hfill (9)

The right hand side of (9) is not more than $\epsilon$ if and only if
$$m_1 \geq \frac{n \ln(2n) - \ln \epsilon}{a_1} = M_1.$$ \hfill (10)
\qed
Theorem 4 For given parameters $n$, $a_1$ and $p$, and for any $\epsilon \in (0,1)$ and any $t \in (0,a_1)$, if $k \geq k_1(t)$ and $m_1 \geq M_1(t)$, then $E_k(n, m_1, a_1) \leq \epsilon$ holds, where

$$k_1(t) = n - \log_2 \frac{a_1 - t}{p}, \quad M_1(t) = \frac{n \ln(2n) - \ln \epsilon}{2t^2}.$$

Proof. Let $r$ be a pattern of size $k \geq k_1(t)$. From Assumption 2 and $|B^s(r)| = 2^{n-k}$, $c_1(r) \leq \min\{1, 2^{n-k}p\}$ holds; now since $k \geq k_1(t)$, $2^{n-k}p \leq a_1 - t < a_1 \leq 1$. Thus, $c_1(r) \leq 2^{n-k}p$ and

$$U(m_1, a_1, c_1(r)) \leq U(m_1, a_1, 2^{n-k}p).$$

By applying (6) of Theorem 2 with $m = m_1$, $a = a_1$ and $\mu = 2^{n-k}p$,

$$U(m_1, a_1, 2^{n-k}p) \leq \exp(-2m_1(a_1 - 2^{n-k}p)^2),$$

and thus

$$E_k(n, m_1, a_1) \leq \exp(-2m_1(a_1 - 2^{n-k}p)^2) \times 2^k \binom{n}{k} \leq \exp(-2m_1t^2) \times (2n)^n. \tag{11}$$

The right hand side of (11) is not more than $\epsilon$ if and only if

$$m_1 \geq \frac{n \ln(2n) - \ln \epsilon}{2t^2} = M_1(t). \tag{12}$$

For given $n$, $a_1$ and $p$, $k_1$ in Theorem 3 is a constant while $k_1(t)$ in Theorem 4 depends on the parameter $t$. The following corollary about the range of $t$ helps in obtaining an upper bound $E_k \leq \epsilon$ with $k_1(t) \leq k \leq k_1$ by Theorem 4.

Corollary 1 For a nonnegative number $\ell$, if $0 < t \leq a_1(1 - 2^{\ell}/e^2)$, then $k_1 - k_1(t) \geq \ell$.

Proof. It directly comes from the definition of $k_1$ and $k_1(t)$. \qed

Now we show an upper bound on $E_k^s$ for "small" $k$.

Theorem 5 For given parameters $n$, $m_1$, $a_0$ and $q$, and for any $\epsilon \in (0,1)$ and any $s \in (0,1)$, if $k \leq k_0(s)$ and $m_0 \geq M_0(s)$, then $E_k^s(n, m_1, m_0, a_1, a_0) \leq \epsilon$ holds, where

$$k_0(s) = n - \log_2 \frac{a_0 + s}{q}, \quad M_0(s) = k_0(s) \ln(2n) - \ln \epsilon \quad 2s^2.$$

Proof. The proof is similar to that of Theorem 4. Let $r$ be a pattern of size $k \leq k_0(s)$. From Assumption 2, $|B^s(r)| = 2^n$ and $k \leq k_0(s)$, $c_0(r) \geq 2^{n-k}q \geq a_0 + s > a_0$ holds. By applying (7) of Theorem 2 with $m = m_0$, $a = a_0$ and $\mu = 2^{n-k}q$,

$$L(m_0, a_0, c_0(r)) \leq L(m_0, a_0, 2^{n-k}q) \leq \exp(-2m_0(2^{n-k}q - a_0)^2)$$

holds and hence we have

$$E_k^s(n, m_1, m_0, a_1, a_0) \leq \exp(-2m_0(2^{n-k}q - a_0)^2) \times 2^k \binom{n}{k} \leq \exp(-2m_0s^2) \times (2n)^{k_0(s)}. \tag{13}$$

The right hand side of (13) is not more than $\epsilon$ if and only if

$$m_0 \geq \frac{k_0(s) \ln(2n) - \ln \epsilon}{2s^2} = M_0(s). \tag{14}$$

\qed
Corollary 2 For given parameters \( n, a_1, a_0, p \) and \( q \), and for any \( t \in (0,a_2(1-1/e^2)] \) and any \( s \in (0,1) \), if \( s \leq q(a_1 - t)/p - a_0 \) holds, then \( k_0(s) \geq k_1(t) \) holds.

Proof. It directly comes from the definitions of \( k_1(t) \) and \( k_0(s) \).

Finally, \( E^* \) is sufficiently small under the conditions given in the following theorem.

Theorem 6 For given parameters \( n, a_1, a_0, p \) and \( q \), and for any \( t \in (0,a_2(1-1/e^2)] \), \( s \in (0,1) \) and \( \varepsilon \in (0,1] \), if \( m_1 \geq \max\{M_1, M_1(t)\} \) and \( m_0 \geq M_0(s) \), then

\[
E^*(n,m_1,m_0,a_1,a_0) \leq \sum_{k=1}^{[k_0(s)]} \varepsilon + \sum_{k=[k_0(s)]+1}^{[k_1(t)]} 2^k \binom{n}{k} + \sum_{k=[k_1(t)]}^{n} \varepsilon
\]

holds. Moreover, if \( s \leq q(a_1 - t)/p - a_0 \), then \( E^*(n,m_1,m_0,a_1,a_0) \leq ne \) holds.

For appropriate values of \( p, q, a_1 \) and \( a_0 \) (e.g., \( p \approx q \) and \( a_1 \gg a_0 \)), there exist \( s \) and \( t \) that satisfy the above conditions, and we can choose \( \varepsilon \) sufficiently small (e.g., \( \varepsilon = 2^{-n} \)), which shows that \( E^*(n,m_1,m_0,a_1,a_0) \) converges to 0.

3 Experimental Studies

In this section, we observe the expectation \( E^* \) for random data sets and the numbers of good rules in real data sets.

3.1 Real Data Sets

We take two real data sets from UCI Repository [9]; i.e., BCW and HEART. The examples in these data sets are numerical vectors, and we transform them into binary examples by the method used in [5]. For a data set, let us denote by \( X_t^* \) and \( X_t^\text{F} \) the sets of available true and false examples, respectively. We denote \( X^* = X_t^* \cup X_t^\text{F}, |X_t^*| = m_t^* \) and \( |X_t^\text{F}| = m_t^\text{F} \). BCW contains 239 true examples and 444 false examples with 13 Boolean variables (i.e., \( m_t^* = 239, m_t^\text{F} = 444, n = 13 \)), while HEART contains 120 true examples and 150 false examples with 10 Boolean variables (i.e., \( m_t^* = 120, m_t^\text{F} = 150, n = 10 \)).

3.2 \( E^* \) for Random Data

Let us observe the expectation \( E^* \) of good rules for random data sets. In the uniformly distributed random data domain, \( P_1(x) = P_2(x) = 1/2^n \) holds for all \( x \in B^n \). By using this, we can compute \( E^*(n,m_1,m_0,a_1,a_0) \) exactly from (2) to (4).

In order to compare \( E^* \) for random data sets to the numbers of good rules in real data sets later, we use \( n = 13 \) and \( 10 \), corresponding to BCW and HEART, respectively. We test all combinations of \( a_1 \in \{0.10,0.20\} \) and \( a_0 \in \{0.00,0.01,0.02\} \). For given \( n, a_1 \) and \( a_0 \), we examine the change of \( E^*(n,m_1,m_0,a_1,a_0) \) as \( m_1 \) and \( m_0 \) increase, where we use \( (m_1,m_0) \) with \( m_1/m_0 = m_t^*/m_t^\text{F} \) for each real data set.

We show \( E^*(n,m_1,m_0,a_1,a_0) \) for two combinations of parameters \( n \) and \( m_1/m_0 \) corresponding to BCW and HEART in Figs. 1 and 2, respectively. Each contains two graphs, where the left (resp., right) graph is for \( a_1 = 0.10 \) (resp., \( a_1 = 0.20 \)). In each graph, the horizontal (resp., vertical) axis represents \( m_1/m_0 \) (resp., \( E^* \)) and the three curves correspond to different values of \( a_0 \). Note that the vertical axis is the logarithmic scale.

\( E^* \) appears to be an approximately monotone decreasing function of \( m_1 + m_0 \). As observed from the figures, \( E^* \) is sufficiently small (i.e., less than 1) as \( m_1 + m_0 \) is larger than at most several hundred. Among the examined values of \( m_1 \) (resp., \( m_0 \)), let us denote by \( M_t^* \) (resp., \( M_t^\text{F} \)) the smallest value that attains \( E^* \leq 1 \). Table 1 shows \( (M_t^*, M_t^\text{F}) \) for each parameter combination examined in these figures.
Let us mention the values of \( \max\{M_1, M_1(t)\} \) and \( M_0(s) \) of Theorem 6 as upper bounds of \( M_1^* \) and \( M_0^* \), respectively. For given \( n, a_1, a_0 \), we compute \( M_1, M_1(t) \) and \( M_0(s) \) by setting \( p = q = 1/2^p \) and \( \varepsilon = 1/n \). We take \( t \) and \( s \) such that they minimize \( \max\{M_1, M_1(t), M_0(s)\} \) among those \( t = \ell \times 10^{-3} \in (0, a_1(1-1/e^2)] \) and \( s = \ell' \times 10^{-3} \in (0, a_1 - a_0 - t] \) for natural numbers \( \ell \) and \( \ell' \); in this experiment, we obtain such \( (t, s) \) by the simple enumeration. The obtained upper bounds are not very tight; e.g., if \( (n, a_1, a_0) = (13, 0.10, 0.00) \), then \( M_1 = 449.20, M_1(t) = 6036.04 \) and \( M_0(s) = 6029.60 \), while \( M_1^* = 95 \) and \( M_0^* = 167 \) from Table 1. It indicates that Theorem 1 and 2 do not always give tight bounds.

### 3.3 Number of Good Rules in Real Data

For given real data sets and \( (a_1, a_0) \), we would like to observe how the number of good rules changes as \( m_1 \) and \( m_0 \) increase and compare its tendency with the values of \( M_1^* \) and \( M_0^* \) of the last subsection. In order to simulate the situation where we have smaller number of examples than the original data set, we randomly sample \( X_1 \subseteq X_1^* \) and \( X_0 \subseteq X_0^* \) with \( |X_1| = m_1, |X_0| = m_0 \) and \( m_1/m_0 = m_1^*/m_0^* \), and generate \((a_1, a_0)\)-good rules in \( X = X_1 \cup X_0 \). We repeat this process \( \tau \) times and take the average \( N \) of the numbers of \((a_1, a_0)\)-good rules. In this experiment, we use \( \tau = 100 \).

We show \( N \) for BCW and HEART in Figs. 3 and 4, respectively. Note that the vertical axes in these figures are not the logarithmic scale in contrast to Figs. 1 and 2. In each graph of the figure for HEART, the three curves overlap.
Table 1: $(M_{1}^*, M_{0}^*)$ for various parameter combinations corresponding to real data sets

<table>
<thead>
<tr>
<th>$a_1$</th>
<th>$a_0$</th>
<th>BCW</th>
<th>HEART</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.00</td>
<td>95</td>
<td>167</td>
</tr>
<tr>
<td>0.01</td>
<td>0.01</td>
<td>131</td>
<td>233</td>
</tr>
<tr>
<td>0.02</td>
<td>0.02</td>
<td>154</td>
<td>277</td>
</tr>
<tr>
<td>0.20</td>
<td>0.00</td>
<td>48</td>
<td>79</td>
</tr>
<tr>
<td>0.01</td>
<td>0.01</td>
<td>48</td>
<td>79</td>
</tr>
<tr>
<td>0.02</td>
<td>0.02</td>
<td>60</td>
<td>101</td>
</tr>
</tbody>
</table>

Figure 3: The number $N$ of good rules for data set BCW

For data set BCW, we see that the number $N$ of generated good rules does not change much when $m_1 + m_0 \geq M_1^* + M_0^*$, while it drastically decreases when $m_1 + m_0 < M_1^* + M_0^*$, for all combinations of $(a_1, a_0)$. We observe that if $m_1$ and $m_0$ are small (e.g., 10 to 30), then more than $10^4$ superficial good rules are extracted, while $X^*$ contains at most 2000 real good rules.

For data set HEART, $N$ becomes 0 with $(m_1, m_0)$ which are smaller than $(M_1^*, M_0^*)$. Some data sets do not contain good rules although they have some regularities or structures; for example, suppose such a data set $X = \mathbb{B}^n$ and each example $x \in X$ is labeled $\omega$ by the parity function:

$$\omega = \begin{cases} 
1 & \text{if } \sum_{j=1}^{n} x_j \text{ is odd}, \\
0 & \text{otherwise}.
\end{cases} \quad (15)$$

Clearly, $X$ does not have $(a_1, a_0)$-good rules under reasonable $(a_1, a_0)$ although it has the rule of (15). In the case of such data sets, it is important to detect that there is no good rule of our definition. Figs. 5 and 6 show that, for these data sets, it is sufficient for us to have $M_1^*$ true examples and $M_0^*$ false examples in order to see that there is no good rule.

The above results justify our claim on the size of the data set needed to generate reliable good rules. However, $\max\{M_1, M_1(t)\}$ and $M_0(s)$ are not very good as upper bounds of $M_1^*$ and $M_0^*$, respectively. It is our future work to derive tighter upper bounds.

4 Conclusion

In this paper, we consider how many examples are needed in data sets for extracting reliable good rules. Our claim is that the data set should contain examples more than the value such that a random data set has good rules very rarely. We derive a required amount of true examples as $\max\{M_1, M_1(t)\}$ and of false examples as $M_0(s)$. We then show some computational studies to justify our claim.
Acknowledgments

We are indebted to Prof. Toshihide Ibaraki of Kwansei Gakuin University, Japan, and Prof. Endre Boros of Rutgers University, USA, for a number of helpful comments and suggestions. This work was supported by Grant-in-Aid for Scientific Research on Priority Areas “Comparative Genomics” from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References