PAPER

A Randomness Based Analysis on the Data Size Needed for Removing Deceptive Patterns

Kazuya HARAGUCHI[†a)], Mutsunori YAGIURA[††b)], *Members*, Endre BOROS[†††c)], *Nonmember*, *and* Toshihide IBARAKI[††††d)], *Fellow*

SUMMARY We consider a data set in which each example is an n-dimensional Boolean vector labeled as true or false. A pattern is a co-occurrence of a particular value combination of a given subset of the variables. If a pattern appears frequently in the true examples and infrequently in the false examples, we consider it a good pattern. In this paper, we discuss the problem of determining the data size needed for removing "deceptive" good patterns; in a data set of a small size, many good patterns may appear superficially, simply by chance, independently of the underlying structure. Our hypothesis is that, in order to remove such deceptive good patterns, the data set should contain a greater number of examples than that at which a random data set contains few good patterns. We justify this hypothesis by computational studies. We also derive a theoretical upper bound on the needed data size in view of our hypothesis.

key words: *frequent/infrequent item sets, association rules, knowledge discovery, probabilistic analysis*

1. Introduction

1.1 Background

Development of computer hardware technology enables us to save massive data at a low cost. In order to discover hidden meaningful knowledge from such data, various methodologies have been studied so far under the name of knowledge discovery, data mining, and so on.

A *data set* consists of *examples* drawn from the population of the considered phenomenon. One of the most challenging problems in the literature is to generate (or to enumerate) all *patterns*, substructures of examples, appearing frequently/infrequently in a given data set. This problem is often formulated as *frequent/infrequent pattern mining*, an important issue in data mining and bioinformatics (e.g., knowledge discovery from genome databases) [1], [6], [14]. (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.)

A "good" pattern in some sense may carry us useful information on decision making. However, its reliability as knowledge must heavily depend on the size of the data set; if the data set is too small, a pattern may be *deceptive* and thus may not serve as meaningful knowledge. While we can store a massive data set cheaply these days, data collection is still expensive in many application areas (e.g., weather data) [10]. In such areas, it is difficult to collect enough examples, and in this paper, we analyze the size of a data set needed for removing deceptive patterns, as an attempt to establish a criterion on the amount of examples needed for efficient knowledge discovery.

1.2 Preliminaries

Let us introduce the notations and terminologies used throughout this paper. Let $\mathbf{B} = \{0, 1\}$. We denote a data set by X. Each element in X, an example, is represented by a vector in $\mathbf{B}^n$, and is labeled either by 1 (*true*) or by 0 (*false*). We denote by X_1 (resp., X_0) the set of true (resp., false) examples in X, i.e., $X = X_1 \cup X_0$ with disjoint X_1 and X_0. We call the cardinality $|X|$ the *size* of a data set X. If $m_1 = |X_1|$ and $m_0 = |X_0|$, then we call X an (m_1, m_0)-data set.

For a vector $x \in \mathbf{B}^n$ and a subset $J \subseteq \{1, \ldots, n\}$, we denote by $x|_J = (x_j \mid j \in J)$ the sub-vector of x corresponding to the index set J. A pattern $r = (J, b)$ is defined by a subset $J \subseteq \{1, \ldots, n\}$ and a Boolean vector $b \in \mathbf{B}^{|J|}$. Given a pattern $r = (J, b)$ and a Boolean vector $x \in \mathbf{B}^n$, we say that r *appears* in x if $x|_J = b$ holds. Let us denote by $X(r)$ the set of examples in X in which r appears; i.e., $X(r) = \{x \in X \mid x|_J = b\}$. In particular, $\mathbf{B}^n(r)$ denotes the set of all binary vectors in which r appears. We define the *frequency* of r by $f(r, X) = |X(r)|/|X|$, i.e., by the proportion of examples of X in which r appears. Given a constant a ($0 \leq a \leq 1$), we call a pattern r *a-frequent* (resp., *a-infrequent*) in X, if $f(r, X) \geq a$ (resp., $f(r, X) \leq a$).

For given constants a_1, a_0 ($0 \leq a_1, a_0 \leq 1$), we call a pattern r an *(a_1, a_0)-pattern* in X, if $f(r, X_1) \geq a_1$ and $f(r, X_0) \leq a_0$. If a_1 is "large enough" and a_0 is "small enough" such a pattern describes a feature of the true examples in X. One could also consider (a_1, a_0)-patterns capturing the features of false examples (i.e., patterns r that

Manuscript received June 19, 2006.
Manuscript revised October 29, 2007.

[†] The author is with the Department of Information Technology and Electronics, Faculty of Science and Engineering, Ishinomaki Senshu University, Ishinomaki-shi, 986–8580 Japan.

[††] The author is with the Department of Computer Science and Mathematical Informatics, Graduate School of Information Science, Nagoya University, Nagoya-shi, 464–8603 Japan.

[†††] The author is with the Rutgers Center for Operations Research, Rutgers University, NJ, 08854–8003 USA.

[††††] The author is with the Department of Informatics, School of Science and Technology, Kwansei Gakuin University, Sanda-shi, 669–1337 Japan.

a) E-mail: kazuyah@isenshu-u.ac.jp
b) E-mail: yagiura@nagoya-u.jp
c) E-mail: boros@rutcor.rutgers.edu
d) E-mail: ibaraki@ksc.kwansei.ac.jp

DOI: 10.1093/ietisy/e91–d.3.781

satisfy $f(r, X_1) \leq a_1$ and $f(r, X_0) \geq a_0$ for "small" a_1 and "large" a_0). Since those could be obtained by interchanging the roles of true and false examples, we focus on "good" patterns for true examples in this paper.

It is well-known that one can find frequent/infrequent patterns in incrementally polynomial time [1], and many fast algorithms for this task have been proposed so far (e.g., [12]). By applying the previous algorithms to generate frequent/infrequent patterns, we can generate all (a_1, a_0)-patterns from the data set X in incrementally polynomial time; e.g., by taking the intersection of the set of a_1-frequent patterns in X_1 and that of a_0-infrequent patterns in X_0.

1.3 Description of Problems

We consider the problem of determining the data size needed to remove *deceptive* (a_1, a_0)*-patterns*. Let us define a *domain* by $D = (n, \rho, P_1, P_0)$, where n denotes the number of Boolean variables, $\rho \in [0, 1]$ denotes a probability, and $P_1, P_0 : \mathbf{B}^n \to [0, 1]$ denote probability distributions. Since P_1 and P_0 are probability distributions, it holds that

$$\sum_{x \in \mathbf{B}^n} P_1(x) = \sum_{x \in \mathbf{B}^n} P_0(x) = 1, \qquad (1)$$

and $P_1(x), P_0(x) \geq 0$ for any $x \in \mathbf{B}^n$. We make an assumption on the distribution of examples as follows;

Assumption 1: Given a domain $D = (n, \rho, P_1, P_0)$, an example (x, ω) is independently generated by the following steps:

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

Step 2: A vector x with label ω is drawn according to the distribution P_ω.

Now, given a data set X generated from a domain D, we expect that (a_1, a_0)-patterns in X carry important information about D. It is possible, however, that some of them are deceptive; they might be present as (a_1, a_0)-patterns in X only by chance, independently of the underlying structure of D. Obviously, such deceptive (a_1, a_0)-patterns would exist with high probability if m_1 and m_0 of an (m_1, m_0)-data set X are small, but the probability will diminish if m_1 and m_0 are sufficiently large. More precisely, let $E_D(m_1, m_0; a_1, a_0)$ denote the expected number of (a_1, a_0)-patterns found in an (m_1, m_0)-data set X generated from D. If $E_D(m_1, m_0; a_1, a_0) \gg E_D(m'_1, m'_0; a_1, a_0)$ holds for sufficiently large m'_1 and m'_0 with $m_1/m_0 = m'_1/m'_0$, then we conclude that X is not large enough and that it contains a lot of deceptive (a_1, a_0)-patterns. This will be experimentally studied in Sect. 3.

We then would like to estimate the sizes of m_1 and m_0, which guarantee that most of the (a_1, a_0)-patterns found in an (m_1, m_0)-data set are not deceptive. For this purpose, we introduce the random domain $R = (n, 1/2, U, U)$, where U denotes the uniform distribution with $U(x) = 1/2^n$ for all $x \in \mathbf{B}^n$. A data set generated from R is called a *random data set*. We consider that R has no particular structure, and any (a_1, a_0)-pattern found in a random data set is deceptive. By using the random domain, our hypothesis for the needed data size is summarized as follows.

Hypothesis 1: Let X be an (m_1, m_0)-data set generated from a given domain D. Then the probability of deceptive (a_1, a_0)-patterns found in X is (approximately) the same as the probability that an (m_1, m_0)-random data set Y contains (a_1, a_0)-patterns.

Therefore, we estimate experimentally and theoretically the sizes of m_1 and m_0, at which an (m_1, m_0)-random data set Y contains approximately no (a_1, a_0)-patterns; they will be used as the sizes of an (m_1, m_0)-data set X which contains no deceptive (a_1, a_0)-patterns.

The composition of this paper is as follows. After describing the related works in Sect. 2, we study such data sizes experimentally in Sect. 3, and derive their theoretical upper bounds by some probabilistic analysis in Sect. 4. Then in Sect. 5, we give the concluding remarks.

2. Related Works

If a pattern $r = (J, b)$ is a frequent pattern in X and there is no frequent pattern $r' = (J', b')$ with $J' \supset J$ and $b'|_J = b$, then we call r a *maximal frequent pattern*. If r is an infrequent pattern in X and there is no infrequent pattern $r' = (J', b')$ with $J' \subset J$ and $b|_{J'} = b'$, then we call r a *minimal infrequent pattern*. Boros et al. [4] showed that, given a family of $O(n^\varepsilon)$ maximal frequent patterns, it is NP-complete to decide whether X has any further maximal frequent patterns (for arbitrarily small fixed $\varepsilon > 0$), and that all minimal infrequent patterns can be enumerated in incremental quasi-polynomial time.

The problem of finding frequent patterns is closely related to that of *association rules*. An association rule is generally defined by two patterns $(r, r') = ((J, b), (J', b'))$ with $J \cap J' = \emptyset$; it represents that an example x with $x|_J = b$ is likely to attain $x|_{J'} = b'$.

An association rule (r, r') is usually evaluated by its *support* (which is the proportion of examples in X where both r and r' appear) 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 is a basic operation in finding association rules.

As the task of finding association rules from a huge data set is very time-consuming, Li et al. [9] and Toivonen [11] discussed the proper size of a randomly drawn subset X' of the original data set X such that $f(r, X')$ is close enough to $f(r, X)$ with a high probability, for all patterns r. While they consider the random sampling of a manageable size from the given huge data set, we consider the situation in which the size of the given data set is small, and discuss whether the extracted (a_1, a_0)-patterns are deceptive or not. This is the main difference between our approach and the existing ones.

3. Experimental Studies

3.1 Expected Number of (a_1, a_0)-Patterns

We derive the expected number of (a_1, a_0)-patterns in an (m_1, m_0)-data set generated from a domain $D = (n, \rho, P_1, P_0)$.

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

$$\begin{aligned} c_1(r; D) &= \sum_{x \in \mathbf{B}^n(r)} P_1(x), \\ c_0(r; D) &= \sum_{x \in \mathbf{B}^n(r)} P_0(x). \end{aligned} \tag{2}$$

More generally, under the condition that m_1 true examples and m_0 false examples are generated, the probability that a pattern r is a_1-frequent in the m_1 true examples is:

$$\begin{aligned} &b_+(m_1, a_1, c_1(r; D)) \\ &\quad = \sum_{s=\lceil a_1 m_1 \rceil}^{m_1} \binom{m_1}{s} c_1(r; D)^s (1 - c_1(r; D))^{m_1 - s}, \end{aligned} \tag{3}$$

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

$$\begin{aligned} &b_-(m_0, a_0, c_0(r; D)) \\ &\quad = \sum_{s=0}^{s=\lfloor a_0 m_0 \rfloor} \binom{m_0}{s} c_0(r; D)^s (1 - c_0(r; D))^{m_0 - s}. \end{aligned} \tag{4}$$

Note that the product $b_+ b_-$ gives the *expectation* that r is an (a_1, a_0)-pattern in an (m_1, m_0)-data set generated from the domain D.

For a pattern $r = (J, b)$, let us call the cardinality $|J|$ the *level* of r. We denote by R_k the set of all possible patterns of level k ($1 \le k \le n$). Note that $|R_k| = 2^k \binom{n}{k}$ holds and that $|\mathbf{B}^n(r)| = 2^{n-k}$ holds for any $r \in R_k$. Let $E_D(m_1, m_0; a_1, a_0)$ be the expected number of (a_1, a_0)-patterns in an (m_1, m_0)-data set from the domain D, and $E_{D,k}(m_1, m_0; a_1, a_0)$ be the same number when their levels are restricted to k. From the linearity of expectations, they are formulated as follows:

$$\begin{aligned} E_D(m_1, m_0; a_1, a_0) &= \sum_{k=1}^{n} E_{D,k}(m_1, m_0; a_1, a_0) \\ &= \sum_{k=1}^{n} \sum_{r \in R_k} b_+(m_1, a_1, c_1(r; D)) \\ &\qquad \times b_-(m_0, a_0, c_0(r; D)). \end{aligned} \tag{5}$$

3.2 Real Data Sets

Table 1 Summary of ten data sets from UCI repository.

Data	n	m_1^*	m_0^*	$m_1^* + m_0^*$
AUS	14	307	383	690
BCW	13	239	444	683
BUPA	21	200	145	345
CAR	12	518	1210	1728
CRX	13	296	357	653
HABER	18	75	219	294
HEART	10	120	150	270
IONO	9	225	126	351
PIMA	15	268	500	768
TTT	12	626	332	958

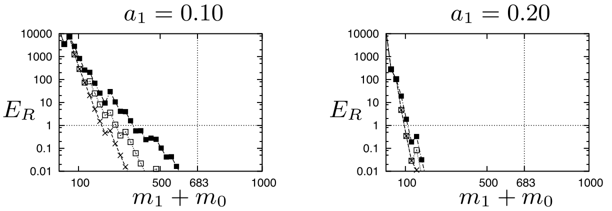


Fig. 1 $E_R(m_1, m_0; a_1, a_0)$ with $n = 13$ and $m_1/m_0 = 239/444$ corresponding to data set BCW. (Lines with points ×, □, ■ represent $a_0 = 0.00, 0.01, 0.02$, respectively.)

We take ten real data sets from UCI Repository [3]; i.e., AUS, BCW, BUPA, CAR, CRX, HABER, HEART, IONO, PIMA, TTT. In order to handle these data sets in our scheme, we modify them as follows:

- CAR is a four-labeled data set, and we modify it to a two-labeled data set; an example in CAR is labeled one of the four labels (i.e., `unacc`, `acc`, `good`, `v-good`). We treat those labeled `unacc` as false examples, and the rest as true examples.
- Some data sets contain examples with missing values or contradiction, and we exclude such examples.
- Finally, since the examples in some data sets are numerical and/or categorical vectors, we transform them into binary examples by the method used in [7].

For each real data set, let us denote by X_1^* and X_0^* the sets of true and false examples, respectively. We denote $X^* = X_1^* \cup X_0^*$, $m_1^* = |X_1^*|$ and $m_0^* = |X_0^*|$. Table 1 shows a summary of the binary data sets transformed from the ten real data sets.

3.3 E_R on Random Data Sets

We first compute the expected number of (a_1, a_0)-patterns in an (m_1, m_0)-random data set $E_R(m_1, m_0; a_1, a_0)$ by (2) to (5) with $D = R$. In order to compare this E_R with the expected number E_D on a real data set (where we write its domain by D) later, we adopt the number n of Boolean variables and the ratio $m_1/m_0 = m_1^*/m_0^*$ corresponding to the real data set, and test various m_1 and m_0 for all combinations of $a_1 \in \{0.10, 0.20\}$ and $a_0 \in \{0.00, 0.01, 0.02\}$.

Figures 1 and 2 show the computed $E_R(m_1, m_0; a_1, a_0)$ with the parameters corresponding to BCW and BUPA, respectively; i.e., $n = 13$ and $m_1 + m_0$ is changed with keeping $m_1/m_0 = 239/444$ for BCW, and $n = 21$ and $m_1/m_0 =$

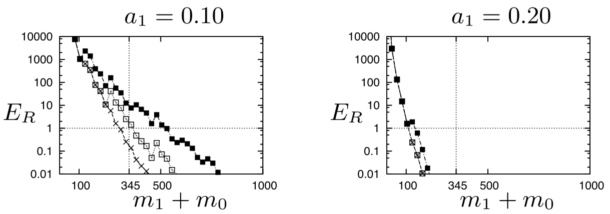


Fig. 2 $E_R(m_1, m_0; a_1, a_0)$ with $n = 21$ and $m_1/m_0 = 200/145$ corresponding to data set BUPA. (Lines with points ×, □, ■ represent $a_0 = 0.00, 0.01, 0.02$, respectively.)

Table 2 The (M_1^*, M_0^*) for real data sets with $a_1 = 0.10$. (A figure with an underline indicates that $M_1^* + M_0^* \le m_1^* + m_0^*$ holds.)

Data	a_0	M_1^*	M_0^*	$M_1^* + M_0^*$
AUS	0.00	101	124	$\underline{225}$
	0.01	121	149	$\underline{270}$
	0.02	161	199	$\underline{360}$
BCW	0.00	81	149	$\underline{230}$
	0.01	91	169	$\underline{260}$
	0.02	132	243	$\underline{375}$
BUPA	0.00	172	123	$\underline{295}$
	0.01	212	153	365
	0.02	273	197	470
CAR	0.00	71	164	$\underline{235}$
	0.01	81	189	$\underline{270}$
	0.02	122	283	$\underline{405}$
CRX	0.00	102	123	$\underline{225}$
	0.01	121	144	$\underline{265}$
	0.02	161	194	$\underline{355}$
HABER	0.00	81	234	315
	0.01	101	294	395
	0.02	151	439	590
HEART	0.00	83	102	$\underline{185}$
	0.01	103	127	$\underline{230}$
	0.02	143	177	$\underline{320}$
IONO	0.00	132	73	$\underline{205}$
	0.01	132	73	$\underline{205}$
	0.02	174	96	$\underline{270}$
PIMA	0.00	91	169	$\underline{260}$
	0.01	102	188	$\underline{290}$
	0.02	151	279	$\underline{430}$
TTT	0.00	161	84	$\underline{245}$
	0.01	161	84	$\underline{245}$
	0.02	242	128	$\underline{370}$

200/145 for BUPA. Each figure contains two cases corresponding to $a_1 = 0.10$ and 0.20, where the horizontal (resp., vertical) axis represents $m_1 + m_0$ (resp., E_R) and three curves correspond to different values of a_0. Note that the vertical axis is in the logarithmic scale. The E_R appears to be monotonically decreasing with $m_1 + m_0$ if we neglect small irregularities, and becomes less than 1 as $m_1 + m_0$ becomes larger than a certain point.

Among the examined values of m_1 (resp., m_0), let us denote by M_1^* (resp., M_0^*) the smallest value that attains $E_R \le 1$. Table 2 shows the observed (M_1^*, M_0^*) for the parameter values m_1 and m_0, which correspond to the ten real data sets, where we always use $a_1 = 0.10$. In this table, a real data set whose $M_1^* + M_0^*$ is underlined indicates that $M_1^* + M_0^* \le m_1^* + m_0^*$ holds; i.e., the data set contains a sufficient number of examples in view of our hypothesis.

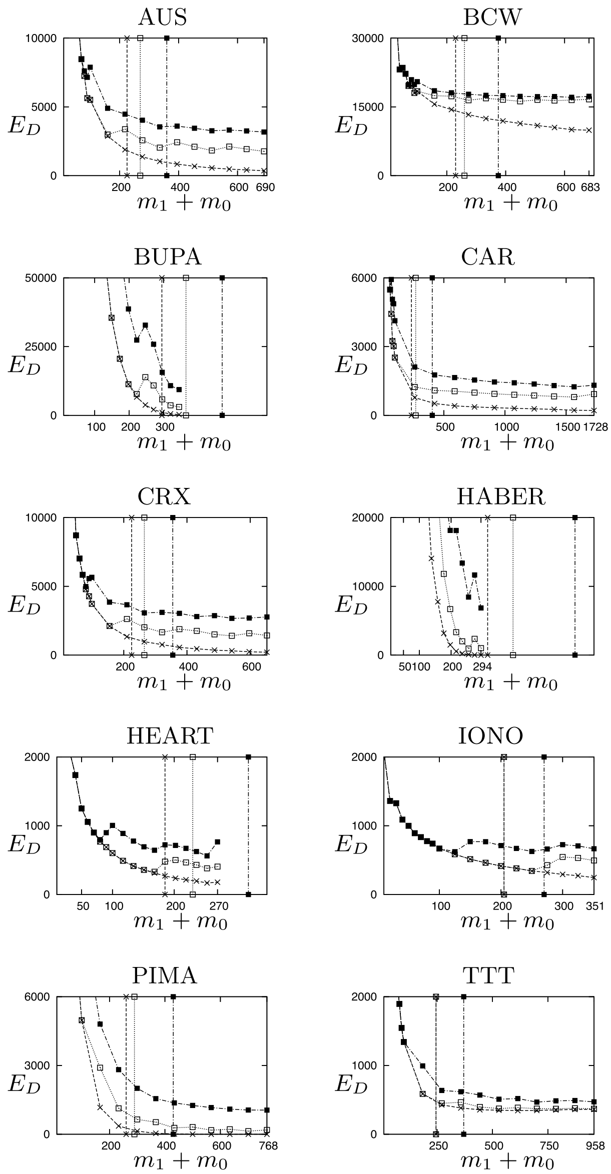


Fig. 3 Expected number E_D of (a_1, a_0)-patterns on real data sets with $a_1 = 0.10$. (Lines with points ×, □, ■ represent $a_0 = 0.00, 0.01, 0.02$, respectively. A broken line parallel to the vertical axis represents $M_1^* + M_0^*$.)

3.4 E_D on Real Data Sets

Now, for the ten real data sets, we would like to know how the expected number of (a_1, a_0)-patterns $E_D(m_1, m_0; a_1, a_0)$ changes as m_1 and m_0 increase relative to the M_1^* and M_0^*. However, we cannot compute E_D exactly by (5) since we do not know the domain D of a real data set. (Furthermore, we do not know even whether the examples are generated according to Assumption 1 or not. In this experiment, however, we regard that they are.) In order to estimate $E_D(m_1, m_0; a_1, a_0)$ experimentally for $m_1 \le m_1^*$ and $m_0 \le m_0^*$, we randomly sample subsets $X_1 \subseteq X_1^*$ and $X_0 \subseteq X_0^*$ with $|X_1| = m_1$ and $|X_0| = m_0$, satisfying $m_1/m_0 = m_1^*/m_0^*$, and enumerate all (a_1, a_0)-patterns in $X = X_1 \cup X_0$. For each tested values of m_1 and m_0, we repeat this process τ times and estimate $E_D(m_1, m_0; a_1, a_0)$ by the average of the numbers of (a_1, a_0)-patterns. In this experiment, we use $\tau = 100$.

The results are shown in Fig. 3, where the vertical axis indicates E_D and the horizontal axis indicates the size $m_1 + m_0$. Note that the vertical axes in these figures are not in the logarithmic scale in contrast to Figs. 1 and 2. In this experiment, we always use $a_1 = 0.10$, and a_0 is set to 0.00, 0.01 and 0.02, respectively. Three curves correspond to different values of a_0, and a broken line parallel to the vertical axis represents the value that corresponds to $M_1^* + M_0^*$.

As shown in these figures, when the size of $m_1 + m_0$ is small, the (m_1, m_0)-data set contains (a_1, a_0)-patterns much more than the (m_1^*, m_0^*)-data set. In such cases, we conclude that the expected number E_D contains many deceptive (a_1, a_0)-patterns. Hypothesis 1 states that we need to satisfy $m_1 \geq M_1^*$ and $m_0 \geq M_0^*$ in order to remove deceptive (a_1, a_0)-patterns, and it surely holds in the results; all the curves are stabilized after passing the point of $m_1 + m_0 = M_1^* + M_0^*$. For AUS, for example, if $m_1 + m_0$ is small (e.g., less than 100), then more than 5.0×10^3 (a_1, a_0)-patterns exist, while there are substantially smaller number of (a_1, a_0)-patterns if $m_1 + m_0 \geq M_1^* + M_0^*$ holds, for all three cases of $a_0 = 0.00, 0.01, 0.02$. Furthermore, the expected numbers E_D do not change to a great extent in the range $m_1 + m_0 \geq M_1^* + M_0^*$.

For BUPA and HABER, $|X^*| = m_1^* + m_0^* < M_1^* + M_0^*$ holds for almost all tested parameter combinations. According to our hypothesis, the size of X^* is not large enough. In fact, E_D is still making a rapid change even if $m_1 + m_0$ is increased to the limit of $m_1^* + m_0^*$, and thus the given data set X^* appears to contain many deceptive (a_1, a_0)-patterns.

4. Upper Bounds on the Needed Data Size

4.1 Preliminaries

The determination of M_1^* and M_0^* by using (5) requires a nontrivial computational cost. To alleviate this, we derive an upper bound on $M_1^* + M_0^*$ in this section. For the derivation, we assume that any domain $D = (n, \rho, P_1, P_0)$ satisfies the following assumption.

Assumption 2: For any $x \in \mathbf{B}^n$, $P_1(x) \leq p$ 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 random domain R is realized by setting $p = q = 1/2^n$.

For a domain D satisfying Assumption 2, we show that an upper bound on $E_{D,k}(m_1, m_0; a_1, a_0)$ becomes sufficiently small (i.e., not more than ε, a small positive value) if k is in some range, either m_1 or m_0 is larger than some threshold, and a few other conditions hold. If an upper bound on $E_{D,k}(m_1, m_0; a_1, a_0)$ becomes sufficiently small for all $k = 1, \ldots, n$, then their sum $E_D(m_1, m_0; a_1, a_0) = \sum_k E_{D,k}(m_1, m_0; a_1, a_0)$ also becomes small; thus such thresholds on m_1 and m_0 can respectively be used as upper bounds on the needed numbers of true and false examples, M_1^* and M_0^*.

Note that $E_{D,k}$ with "large" k or "small" k cannot be large for the following reason. Consider a pattern r with level k in an (m_1, m_0)-data set X. If k is large (resp., small), then $|\mathbf{B}^n(r)| = 2^{n-k}$ tells that the r appears in a small (resp., large) portion of vectors in $\mathbf{B}^n$. Thus the r is unlikely to be a_1-frequent in the m_1 true examples (resp., a_0-infrequent in the m_0 false examples), and thus unlikely to be an (a_1, a_0)-pattern in X. Our analysis in the following is to refine this observation.

4.2 Probabilistic Analyses on $E_{D,k}$ and Bounds on M_1^* and M_0^*

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

Theorem 1: (Chernoff [5]) Given a positive integer m and $0 \leq \mu \leq 1$, let Q_i be a random variable taking the value as follows:

$$Q_i = \begin{cases} 1 - \mu & \text{with probability } \mu, \\ -\mu & \text{with probability } 1 - \mu, \end{cases} \tag{6}$$

and let $Q = \sum_{i=1}^m Q_i$. Then, for any $\beta > 1$,

$$\Pr(Q \geq (\beta - 1)\mu m) < (\exp(\beta - 1)\beta^{-\beta})^{\mu m} \tag{7}$$

holds.

Theorem 2: (Hoeffding [8]) For a positive integer m and $0 \leq a \leq 1$, if $0 \leq \mu \leq a$, then the b_+ in (3) satisfies

$$b_+(m, a, \mu) \leq \exp(-2m(a - \mu)^2). \tag{8}$$

Similarly, if $a \leq \mu \leq 1$, then the b_- in (4) satisfies

$$b_-(m, a, \mu) \leq \exp(-2m(\mu - a)^2). \tag{9}$$

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

Now we derive two types of upper bounds on $E_{D,k}$ for "large" k.

Theorem 3: Suppose that D, m_1, m_0, a_1, a_0, k and $\varepsilon \in (0, 1]$ are given. If $k \geq K_+$ and $m_1 \geq M_1$, then $E_{D,k}(m_1, m_0; a_1, a_0) \leq \varepsilon$ holds, where

$$K_+ = n - \log_2 \frac{a_1}{e^2 p}, \tag{10}$$

$$M_1 = \frac{n \ln(2n) - \ln \varepsilon}{a_1}, \tag{11}$$

and e denotes the base of the natural logarithm.

Proof : Let r be a pattern of level $k \geq K_+$. From Assumption 2 and $|\mathbf{B}^n(r)| = 2^{n-k}$, we have $c_1(r; D) \leq \min\{1, 2^{n-k}p\}$, and since $2^{n-k} \leq 2^{n-K_+} = a_1/(e^2 p)$, we have $c_1(r; D) \leq 2^{n-k}p \leq a_1/e^2 < 1$. Let Z_i be a random variable taking the value as follows:

$$Z_i = \begin{cases} 1 & \text{with probability } 2^{n-k}p, \\ 0 & \text{with probability } 1 - 2^{n-k}p, \end{cases} \tag{12}$$

and let $Z = \sum_{i=1}^{m_1} Z_i$. Let $Q_i = Z_i - 2^{n-k}p$ and $Q = \sum_{i=1}^{m_1} Q_i = Z - 2^{n-k}pm_1$. Then, we have

$$\begin{aligned} &E_{D,k}(m_1, m_0; a_1, a_0) \\ &\quad = \sum_{r \in R_k} b_+(m_1, a_1, c_1(r; D))b_-(m_0, a_0, c_0(r; D)) \\ &\quad \le b_+(m_1, a_1, 2^{n-k}p) \times |R_k| \\ &\quad = \Pr(Z \ge a_1 m_1) \times 2^k \binom{n}{k} \\ &\quad = \Pr\left(Q \ge 2^{n-k}pm_1 \left(\frac{a_1}{2^{n-k}p} - 1\right)\right) \times 2^k \binom{n}{k}. \qquad (13) \end{aligned}$$

From $k \ge K_+$, it holds $a_1/(2^{n-k}p) \ge e^2 > 1$. By applying Theorem 1 with $m = m_1$, $\mu = 2^{n-k}p$ and $\beta = a_1/(2^{n-k}p)$, we have

$$\begin{aligned} E_{D,k}(m_1, m_0; a_1, a_0) &< \left(\frac{2^{n-k}pe}{a_1}\right)^{a_1 m_1} \times 2^k \binom{n}{k} \\ &\le \left(\frac{2^{n-k}pe}{a_1}\right)^{a_1 m_1} \times (2n)^k \\ &\le e^{-a_1 m_1} \times (2n)^n. \qquad (14) \end{aligned}$$

The right hand side of (14) is not more than ε if and only if

$$m_1 \ge \frac{n \ln(2n) - \ln \varepsilon}{a_1} = M_1. \qquad (15)$$

□

Another upper bound on $E_{D,k}$ for large k is given below. It depends on a parameter t and can bound $E_{D,k}$ for k with $k > K_+ - 3$.

Theorem 4: Suppose that D, m_1, m_0, a_1, a_0, k and $\varepsilon \in (0, 1]$ are given. If $k \ge K_+(t)$ and $m_1 \ge M_1(t)$ for some $t \in (0, a_1)$, then $E_{D,k}(m_1, m_0; a_1, a_0) \le \varepsilon$ holds, where

$$K_+(t) = n - \log_2 \frac{a_1 - t}{p}, \qquad (16)$$

$$M_1(t) = \frac{n \ln(2n) - \ln \varepsilon}{2t^2}. \qquad (17)$$

Proof : For an arbitrary $t \in (0, a_1)$, let r be a pattern of level $k \ge K_+(t)$. From Assumption 2 and $|\mathbf{B}^n(r)| = 2^{n-k}$, we have $c_1(r; D) \le \min\{1, 2^{n-k}p\}$, and since $k \ge K_+(t)$, we have $2^{n-k}p \le a_1 - t < a_1 \le 1$. Thus, $c_1(r; D) \le 2^{n-k}p$ and

$$b_+(m_1, a_1, c_1(r; D)) \le b_+(m_1, a_1, 2^{n-k}p). \qquad (18)$$

By applying (8) of Theorem 2 with $m = m_1$, $a = a_1$ and $\mu = 2^{n-k}p$, we have

$$\begin{aligned} &b_+(m_1, a_1, 2^{n-k}p) \\ &\quad \le \exp(-2m_1(a_1 - 2^{n-k}p)^2), \qquad (19) \end{aligned}$$

and hence

$$\begin{aligned} &E_{D,k}(m_1, m_0; a_1, a_0) \\ &\quad \le \exp(-2m_1(a_1 - 2^{n-k}p)^2) \times 2^k \binom{n}{k} \\ &\quad \le \exp(-2m_1 t^2) \times (2n)^n. \qquad (20) \end{aligned}$$

The right hand side of (20) is not more than ε if and only if

$$m_1 \ge \frac{n \ln(2n) - \ln \varepsilon}{2t^2} = M_1(t). \qquad (21)$$

□

Given D and a_1, the K_+ in Theorem 3 is a constant while $K_+(t)$ in Theorem 4 depends on the parameter t. The following corollary about the range of t is useful in obtaining an upper bound $E_{D,k} \le \varepsilon$ with $K_+(t) \le k \le K_+$ from Theorem 4.

Corollary 1: If we set $t = a_1(1 - \ell/e^2)$ for a constant $1 \le \ell < e^2$, then $K_+ - K_+(t) = \log_2 \ell$.

Proof : It directly comes from the definition of K_+ and $K_+(t)$. □

Note that $K_+ - K_+(t) < \log_2 e^2 < 3$ holds.

Now we turn to an upper bound on $E_{D,k}$ for "small" k.

Theorem 5: Suppose that D, m_1, m_0, a_1, a_0, k and $\varepsilon \in (0, 1]$ are given. If $k \le K_-(s)$ and $m_0 \ge M_0(s)$ hold for some $s \in (0, 1)$, then $E_{D,k}(m_1, m_0; a_1, a_0) \le \varepsilon$ holds, where

$$K_-(s) = n - \log_2 \frac{a_0 + s}{q}, \qquad (22)$$

$$M_0(s) = \frac{K_-(s) \ln(2n) - \ln \varepsilon}{2s^2}. \qquad (23)$$

Proof : The proof is similar to that of Theorem 4. For an arbitrary $s \in (0, 1)$, let r be a pattern of level $k \le K_-(s)$. From Assumption 2, $|\mathbf{B}^n(r)| = 2^{n-k}$ and $k \le K_-(s)$, we have $c_0(r; D) \ge 2^{n-k}q \ge a_0 + s > a_0$. By applying (9) of Theorem 2 with $m = m_0$, $a = a_0$ and $\mu = 2^{n-k}q$,

$$\begin{aligned} &b_-(m_0, a_0, c_0(r; D)) \\ &\quad \le b_-(m_0, a_0, 2^{n-k}q) \\ &\quad \le \exp(-2m_0(2^{n-k}q - a_0)^2) \qquad (24) \end{aligned}$$

holds and hence we have

$$\begin{aligned} &E_{D,k}(m_1, m_0; a_1, a_0) \\ &\quad \le \exp(-2m_0(2^{n-k}q - a_0)^2) \times 2^k \binom{n}{k} \\ &\quad \le \exp(-2m_0 s^2) \times (2n)^{K_-(s)}. \qquad (25) \end{aligned}$$

The right hand side of (25) is not more than ε if and only if

$$m_0 \ge \frac{K_-(s) \ln(2n) - \ln \varepsilon}{2s^2} = M_0(s). \qquad (26)$$

□

Recall that Theorems 3 and 4 hold for large k and Theorem 5 holds for small k. Then, if one of these theorems holds for every $k = 1, \ldots, n$, then we will have $E_{D,k} \le \varepsilon$ for all $k = 1, \ldots, n$ and hence, $E_D = \sum_k E_{D,k} \le n\varepsilon$. More precisely, if we choose parameters t and s so that $K_+(t) \le K_-(s)$ holds, and we have $m_1 \ge \max\{M_1, M_1(t)\}$ and $m_0 \ge M_0(s)$, then one of these theorems holds for every $k = 1, \ldots, n$. A sufficient condition for $K_+(t) \le K_-(s)$ to hold is given in the following corollary.

Corollary 2: If $t \in (0, a_1(1-1/e^2)]$ and $s \in (0, q(a_1-t)/p-a_0]$, then $K_-(s) \geq K_+(t)$ holds.

Proof : It directly comes from the definitions of $K_+(t)$ and $K_-(s)$. □

Finally, $E_D = \sum_k E_{D,k}$ becomes sufficiently small under the conditions given in the following theorem.

Theorem 6: Suppose that D, m_1, m_0, a_1, a_0 and $\varepsilon \in (0,1]$ are given. If $t \in (0, a_1(1-1/e^2)]$ and $s \in (0,1)$ satisfy $s \leq q(a_1-t)/p-a_0$, $m_1 \geq \max\{M_1, M_1(t)\}$ and $m_0 \geq M_0(s)$, then $E_D(m_1, m_0; a_1, a_0) \leq n\varepsilon$ holds.

Corollary 3: For appropriate values of p, q, a_1 and a_0 (e.g., $p \simeq q$ and $a_1 \gg a_0$), there exist t and s that satisfy the above condition $s \leq q(a_1-t)/p-a_0$. Then, if we take ε sufficiently small (e.g., $\varepsilon = 2^{-n}$), $E_D(m_1, m_0; a_1, a_0)$ converges to 0.

Corollary 4: The $\max\{M_1, M_1(t)\}$ and $M_0(s)$ in Theorem 6 are upper bounds on M_1^* and M_0^* in Sect. 3, respectively.

Let us consider the possibility of using $\max\{M_1, M_1(t)\}$ and $M_0(s)$ as estimates on M_1^* and M_0^* in Sect. 3, respectively. To see how close they are, we computed M_1, $M_1(t)$ and $M_0(s)$ on the random domain R for some combinations of (n, a_1, a_0), where we set t and s to the values that minimize $\max\{M_1, M_1(t), M_0(s)\}$ among all $t = \ell \times 10^{-3} \in (0, a_1(1-1/e^2)]$ and $s = \ell' \times 10^{-3} \in (0, a_1 - a_0 - t]$ with natural numbers ℓ and ℓ'. The obtained upper bounds, however, are not very tight; e.g., for $(n, a_1, a_0) = (13, 0.10, 0.00)$, $M_1 = 449.20$, $M_1(t) = 6036.04$ and $M_0(s) = 6029.60$, while $(M_1^*, M_0^*) = (81, 149)$ and $(102, 123)$ from the results for BCW and CRX in Table 2, respectively. It may indicate that the bounds $\max\{M_1, M_1(t)\}$ and $M_0(s)$ are not very accurate indicators of M_1^* and M_0^*.

It is left as our future work to derive tighter theoretical estimates of M_1^* and M_0^*.

5. Conclusion

In this paper, we considered how many examples are needed in a given data set in order to remove deceptive (a_1, a_0)-patterns. Our hypothesis is that the data set should contain a greater number of examples than that at which the probability of having (a_1, a_0)-patterns vanishes for the random data set. We justified the hypothesis by computational experiments in Sect. 3, and derived estimates of such number of examples by probabilistic analysis in Sect. 4.

Our future work includes the theoretical study of the hypothesis based on such theories as randomness and VC dimension [13] from learning theory, and an application of the hypothesis to other enumeration problems (e.g., graph mining).

Acknowledgments

We gratefully acknowledge very careful and detailed comments given by anonymous reviewers. We are indebted to Prof. Hiroshi Nagamochi and Prof. Liang Zhao of Kyoto University, Japan, for a number of helpful comments and suggestions. This work is partially supported by Grant-in-Aids for Scientific Research on Priority Areas “Comparative Genomics” and “New Horizons in Computing” from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] R. Agrawal, T. Imielinski, and A. Swami, “Mining association rules between sets of items in large databases,” Proc. ACM SIGMOD Conf. Management of Data, pp.207–216, Washington, USA, May 1993.

[2] N. Alon, J.H. Spencer, and P. Erdös, eds., The Probabilistic Method, John Wiley & Sons, 1992.

[3] A. Asuncion and D.J. Newman, UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science, 2007.

[4] E. Boros, V. Gurvich, L. Khachiyan, and K. Makino, “On the complexity of generating maximal frequent and minimal infrequent sets,” Proc. STACS 2002, pp.133–141, Antibes Juan-les-Pins, France, March 2002.

[5] H. Chernoff, “A measure of the asymptotic efficiency for tests of a hypothesis based on the sum of observations,” Annals of Mathematical Statistics, vol.23, pp.493–509, 1952.

[6] U. Fayyad, G. Piatetsky-Shapiro, and S. Padhraic, “From data mining to knowledge discovery in databases,” AI Magazine, vol.17, no.3, pp.37–54, 1996.

[7] K. Haraguchi, T. Ibaraki, and E. Boros, “Classifiers based on iterative compositions of features,” Proc. ICKEDS 2004, pp.143–150, Porto, Portugal, June 2004.

[8] W. Hoeffding, “Probability inequalities for sums of bounded random variables,” J. American Statistical Association, vol.58, pp.13–30, 1963.

[9] Y. Li and R.P. Gopalan, “Effective sampling for mining association rules,” Proc. 17th Australian Joint Conference on Artificial Intelligence, pp.391–401, Cairns, Australia, Dec. 2004.

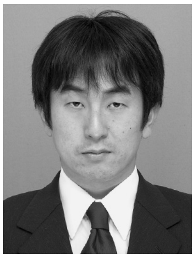
[10] N. Ramakrishnan and C. Bailey-Kellogg, “Gaussian process models of spatial aggregation algorithms,” Proc. IJCAI 2003, pp.1045–1051, Acapulco, Mexico, Aug. 2003.

[11] H. Toivonen, “Sampling large databases for association rules,” Proc. 22nd VLDB Conference, pp.134–145, Bombay, India, Sept. 1996.

[12] T. Uno, M. Kiyomi, and H. Arimura, “LCM ver 2: Efficient mining algorithms for frequent/closed/maximal itemsets,” Proc. IEEE ICDM’04 Workshop FIMI’04, Brighton, UK, Nov. 2004.

[13] V. Vapnik, ed., The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.

[14] G. Yang, “The complexity of mining maximal frequent itemsets and maximal frequent patterns,” Proc. 10th ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining, pp.344–353, Seattle, USA, Aug. 2004.



Kazuya Haraguchi received the B.E., the Master of Informatics and the Ph.D. degree in Informatics from Kyoto University, in 2000, 2002 and 2006, respectively. He is currently with the Department of Information Technology and Electronics, Faculty of Science and Engineering, Ishinomaki Senshu University. His interest includes algorithms, optimization, machine learning and their applications.

Mutsunori Yagiura received the B.E., M.E. and Ph.D. degrees in Engineering from Kyoto University, in 1991, 1993 and 1999, respectively. He is currently with the Department of Computer Science and Mathematical Informatics, Graduate School of Information Science, Nagoya University. His research interest includes metaheuristics, algorithms, optimization, computational complexity and their applications.

Endre Boros is a Professor of Operations Research at Rutgers University, studied mathematics and operations research at the Eötvös Loránd University, Budapest Hungary, and received his Doctorate in 1985. His main areas of research include combinatorial optimization, game and graph theory, logical analysis of data (LAD), and the theory of boolean functions and data mining.

Toshihide Ibaraki received the B.E., M.E., and Ph.D. degrees in Engineering from Kyoto University, in 1963, 1965 and 1970, respectively. After retiring from Kyoto University in March 2004, he is currently with Department of Informatics, School of Science and Technology, Kwansei Gakuin University. His interest includes algorithms, optimization, computational complexity and their applications.