Discovering Interesting Exception Rules with Rule Pair

Einoshin Suzuki

Electrical and Computer Engineering, Yokohama National University, Japan
suzuki@ynu.ac.jp

Abstract. In this paper, we summarize a part of our 10-year endeavor for exception rule discovery. Our results mainly concern interestingness measure, reliability evaluation, practical application, parameter reduction, and knowledge representation.

1 Introduction

Currently, the difference between rule discovery and rule learning [6] is unclear. We might consider that the difference is due to the objectives of data mining (e.g. business) [5] and machine learning (e.g. accuracy) but will soon face exceptions. Typically, a method in rule discovery investigates many candidate rules and is not necessarily related with global performance in the example space while a method in rule learning investigates a smaller number of candidate rules and is often related with accuracy.

In rule discovery, a rule [2, 18] is a statement of a regularity in the form of "if *premise* then *conclusion*". Intuitively a rule can be divided into strong rules, each of which is a description of a regularity for numerous objects with few counterexamples, and weak rules, each of which describes a regularity for fewer objects. Interesting rules are relatively few in number, and are often found among weak rules. Efficient discovery of such interesting rules is in general challenging, since the number of rules typically increases exponentially as the number of objects that a rule covers decreases.

In order to cope with this problem, several researchers have proposed measures for evaluating interestingness of a discovered rule [1, 3, 7–19, 21, 32]. Such measures include various statistical criteria [3, 7, 10, 15, 18], actionability [1], and unexpectedness [8, 11, 12, 14, 16, 17, 19, 21, 32]. Among such measures, unexpectedness deserves special attention since it can uncover special situations that are not obvious.

Most of the studies [8, 11, 12, 14, 17, 19, 21, 32] in interesting-rule discovery attempt to achieve unexpectedness by finding a set of interesting exception rules, each of which represents a different regularity from domain knowledge or a set of strong rules. Consider the rule "using a seat belt is risky for a child", which represents exceptions to the well-known fact "using a seat belt is safe". This rule exhibited unexpectedness when it was discovered from car-accident data

years ago, and is still useful. It should be noted that an exception rule is often beneficial since it often differs from a basis for people's daily activity. For instance, suppose a species of poisonous mushrooms some of which are exceptionally edible. The exact description of their exceptions is highly beneficial since it enables exclusive possession of the edible mushrooms. Moreover, as Silberschatz and Tuzhilin commented [17], exceptions are often related with actionability. Consider an exception rule which describes uncured patients in spite of effective antibiotics, and another exception rule which describes cured patients against a mortal disease. The former can suggest precaution to the antibiotics, while the latter can suggest treatment to the mortal disease.

Discovery methods for interesting exception rules can be divided into two approaches from the viewpoint of background knowledge[1]. In a directed approach [11, 14, 17], a method is first provided with background knowledge typically in the form of rules, then the method obtains exception rules each of which deviates from these rules[2]. In an undirected approach [8, 12, 19, 21, 32], on the other hand, no background knowledge is provided. The target of discovery is typically a set of rule pairs each of which consists of an exception rule and its corresponding strong rule.

Despite its importance, one of the major difficulties for an undirected approach corresponds to its time complexity. Rule discovery is time-consuming when all candidates are searched. Compared with the directed approach, the undirected approach suffers from extra search for strong rules. Let the number of examples in a data set and the number of conditions in a premise be n and M respectively, then the time complexity of single-rule discovery is typically $O(n^{M+1})$, while rule-pair discovery requires $O(n^{2M+1})$. Other important issues concern interestingness measure, reliability evaluation, practical application, parameter reduction, and knowledge representation. In this paper, we summarize a part of our results for these issues.

2 Description of the Problem

We assume that an example e_i is a description about an object stored in a data set in the form of a record, and a data set contains n examples $e_1, e_2, \cdots, e_n$. An example e_i is represented by a tuple $< y_{i1}, y_{i2}, \cdots, y_{im} >$ where $y_{i1}, y_{i2}, \cdots, y_{im}$ are values for m discrete attributes. Here, a continuous attribute is supposed to be converted to a nominal attribute using an existing method such as presented in [4]. An event representing a value assignment to an attribute will be called an atom.

We define a conjunction rule as a rule of which premise is represented by a conjunction of atoms, and of which conclusion is a single atom. In this paper, we mainly consider the problem of finding a set of rule pairs each of which

[1] According to Silberschatz and Tuzhilin, these approaches can be named as subjective and objective [17] respectively.

[2] Several methods such as [14] provide additional search to find more interesting rules from the exception rules.

consists of an exception rule associated with a strong rule. Suppose a strong rule is represented by "if Y_μ then x", where $Y_\mu \equiv y_1 \wedge y_2 \wedge \cdots \wedge y_\mu$ is a conjunction of atoms and x is a single atom. Let $Z_\nu \equiv z_1 \wedge z_2 \wedge \cdots \wedge z_\nu$ be a conjunction of atoms and x' be a single atom which has the same attribute but a value different to the atom x, then the exception rule is represented by "if Y_μ and Z_ν then x'". The discovered pattern in our approach is, therefore, represented by a rule pair $r(x, x', Y_\mu, Z_\nu)$, where M is a user-prespecified parameter for the maximum number of atoms in a premise.

$$r(x, x', Y_\mu, Z_\nu) \equiv \begin{cases} Y_\mu & \rightarrow x \\ Y_\mu \wedge Z_\nu & \rightarrow x' \end{cases} \quad (1)$$
$$\mu,\ \nu \leq M$$

Our objective is to discover a set of (possibly) interesting rule pairs from a data set. The set is specified by assuming either an evaluation criterion or a set of constraints.

3 Methods for Rule-pair Discovery

3.1 MEPRO with its Interestingness Measure

Our rule-pair discovery method MEPRO is based on the rule discovery system ITRULE [18]. The essential of ITRULE lies in its interestingness measure J, which corresponds to the quantity $J(x;\ y)$ of information compressed by a rule $y \rightarrow x$.

$$J(x;\ y) = \Pr(y)\ j(x;\ y) \quad (2)$$

$$\text{where } j(x;\ y) = \Pr(x|y) \log_2 \frac{\Pr(x|y)}{\Pr(x)} + \Pr(\overline{x}|y) \log_2 \frac{\Pr(\overline{x}|y)}{\Pr(\overline{x})} \quad (3)$$

We defined our measure of interestingness of a rule pair as a product ACEP (x, Y_μ, x', Z_ν) of J-measure of a strong rule and J-measure of an exception rule [19]. Our motivation was to obtain rule pairs each of which consists of rules with large J-measure values. We have proved that $J(x; Y_\mu) + J(x'; Y_\mu \wedge Z_\nu)$ is inappropriate as an evaluation index since it is dominated by one of $J(x; Y_\mu)$, $J(x'; Y_\mu \wedge Z_\nu)$ when it is large [19].

$$\text{ACEP}(x, Y_\mu, x', Z_\nu) \equiv J(x; Y_\mu) J(x'; Y_\mu \wedge Z_\nu) \quad (4)$$

We have then proposed a discovery algorithm which generates K rule pairs, where K is a user-specified parameter. In the algorithm, a discovery task is viewed as a search problem, in which a node of a search tree represents a rule pair $r(x, x', Y_\mu, Z_\nu)$. A depth-first search method with maximum depth D is employed to traverse this tree. Let $\mu = 0$ and $\nu = 0$ represent the state in which the premises of a rule pair $r(x, x', Y_\mu, Z_\nu)$ contain neither y_i nor z_i respectively, then we define that $\mu = \nu = 0$ holds in a node of depth 1, and as the depth

increases by 1, an atom is added to the premise of the general or exceptional rule. A node of depth 2 is assumed to satisfy $\mu = 1$ and $\nu = 0$; a node of depth 3, $\mu = \nu = 1$; and a node of depth l (≥ 4), $\mu + \nu = l - 1$ ($\mu, \nu \geq 1$). Therefore, a descendant node represents a rule pair $r(x, x', Y_{\mu'}, Z_{\nu'})$ where $\mu' \geq \mu$ and $\nu' \geq \nu$.

According to the following theorem, an upper-bound exists for the ACEP of this rule pair[19].

Theorem 1. *Let* $\mathrm{H}(\alpha) \equiv [\alpha/\{(1+\alpha)\Pr(\overline{x})\}]^{2\alpha}/\{(1+\alpha)\Pr(x)\}$, α_1 *and* α_2 *satisfy* $\mathrm{H}(\alpha_1) > 1 > \mathrm{H}(\alpha_2)$, *and* $ACEP = \mathrm{ACEP}(x, Y_{\mu'}, x', Z_{\nu'})$.
If $\mathrm{H}(\Pr(x', Y_\mu, Z_\nu)/\Pr(x, Y_\mu)) < 1$ *then,*

$$ACEP < \alpha_2 \Pr(x, Y_\mu)^2 \left\{ \log_2 \left(\frac{1}{1+\alpha_1} \frac{1}{\Pr(x)} \right) + \alpha_1 \log_2 \left(\frac{\alpha_1}{1+\alpha_1} \frac{1}{\Pr(\overline{x})} \right) \right\} \log_2 \frac{1}{\Pr(x')}$$

else

$$ACEP \leq \left\{ \Pr(x, Y_\mu) \log_2 \left(\frac{\mathrm{p}(x, Y_\mu)}{\Pr(x, Y_\mu) + \Pr(x', Y_\mu, Z_\nu)} \frac{1}{\Pr(x)} \right) + \Pr(x', Y_\mu, Z_\nu) \right.$$
$$\left. \cdot \log_2 \left(\frac{\Pr(x', Y_\mu, Z_\nu)}{\mathrm{p}(x, Y_\mu) + \Pr(x', Y_\mu, Z_\nu)} \frac{1}{\Pr(\overline{x})} \right) \right\} \Pr(x', Y_\mu, Z_\nu) \log_2 \frac{1}{\Pr(x')}$$

This upper bound was employed in our approach for a branch-and-bound method which guarantees the optimal solution and is expected to be time-efficient.

We have also introduced probabilistic constraints for eliminating rule pairs each of which has a large $\Pr(x'|Z_\nu)$ [20]. We have also considered unexpectedness from a different perspective and proposed a novel probabilistic criterion which mainly considers the number of counter-examples [22].

3.2 PADRE with its Simultaneous Reliability Evaluation

In rule discovery, generality and accuracy can be considered as frequently-used criteria for evaluating the goodness of a rule. In case of a conjunction rule $Y_\mu \rightarrow x$, these two criteria correspond to the probability $\Pr(Y_\mu)$ of the premise and the conditional probability $\Pr(x|Y_\mu)$ of the conclusion given the premise respectively [18]. Similar to [2], we specify two minimum thresholds θ_1^{S} and θ_1^{F} for generality and accuracy of the strong rule respectively. Two thresholds θ_2^{S} and θ_2^{F} are also specified for generality and accuracy of the exception rule respectively.

Consider the case in which the accuracy of a rule $Z_\nu \rightarrow x'$, which we call a reference rule, is large. In such a case, an exception rule can be considered as expected since it can be easily guessed from this rule. In order to obtain truly unexpected exception rules, we specify a maximum threshold θ_2^{I} for the accuracy of a reference rule.

We then proposed a method PADRE in which we specify thresholds θ_1^{S}, θ_1^{F}, θ_2^{S}, θ_2^{F}, θ_2^{I} for probabilistic criteria of a rule pair. Since a rule pair discovered from 10,000 examples exhibits different reliability from another rule pair discovered from 100 examples, it is inappropriate to use a ratio $\widehat{\Pr}(\cdot)$ in a data set as a probabilistic criterion. Therefore, we considered a true probability $\Pr(\cdot)$ for

each probabilistic criterion, and obtained a set of rule pairs each of which satisfies discovery conditions with the significance level δ [21, 28]. In the following, $\mathrm{MIN}(a, b)$ and $\mathrm{MAX}(a, b)$ represent the smaller one and the larger one of a and b respectively.

$$\begin{aligned}\Pr[\ \ & \Pr(Y_\mu) \geq \theta_1^{\mathrm{S}},\ \Pr(x|Y_\mu) \geq \mathrm{MAX}(\theta_1^{\mathrm{F}}, \widehat{\Pr}(x)),\ \Pr(Y_\mu Z_\nu) \geq \theta_2^{\mathrm{S}}, \\ & \Pr(x'|Y_\mu Z_\nu) \geq \mathrm{MAX}(\theta_2^{\mathrm{F}}, \widehat{\Pr}(x')), \Pr(x'|Z_\nu) \leq \mathrm{MIN}(\theta_2^{\mathrm{I}}, \widehat{\Pr}(x'))\ ] \geq 1 - \delta \end{aligned} \tag{5}$$

Calculating (5) is difficult due to two reasons. First, obtaining a value of a true probability requires assumptions. Second, calculating (5) for a rule pair numerically is time-consuming since (5) contains five true probabilities. This method overcomes these difficulties by obtaining analytical solutions based on simultaneous reliability estimation of true probabilities. Let the number of examples in the data set be n, and $(n\Pr(xY_\mu Z_\nu),\ n\Pr(x'Y_\mu Z_\nu),\ n\Pr(\overline{x}\overline{x'}Y_\mu Z_\nu),\ n\Pr(xY_\mu \overline{Z_\nu}),\ n\Pr(\overline{x}Y_\mu \overline{Z_\nu}),\ n\Pr(\overline{x'Y_\mu} Z_\nu),\ n\Pr(x'\overline{Y_\mu} Z_\nu))$ follow a multi-dimensional normal distribution, then (5) is equivalent to (6) - (10) [21, 28].

$$G(Y_\mu,\ \delta,\ k)\widehat{\Pr}(Y_\mu) \geq \theta_1^{\mathrm{S}} \tag{6}$$

$$F(Y_\mu,\ x,\ \delta,\ k)\widehat{\Pr}(x|Y_\mu) \geq \theta_1^{\mathrm{F}} \tag{7}$$

$$G(Y_\mu Z_\nu,\ \delta,\ k)\widehat{\Pr}(Y_\mu Z_\nu) \geq \theta_2^{\mathrm{S}} \tag{8}$$

$$F(Y_\mu Z_\nu,\ x',\ \delta,\ k)\widehat{\Pr}(x'|Y_\mu Z_\nu) \geq \theta_2^{\mathrm{F}} \tag{9}$$

$$F'(Z_\nu,\ x',\ \delta,\ k)\widehat{\Pr}(x'|Z_\nu) \leq \theta_2^{\mathrm{I}} \tag{10}$$

where

$$G(a,\ \delta,\ k) \equiv 1 - \beta(\delta, k)\sqrt{\frac{1 - \widehat{\Pr}(a)}{n\widehat{\Pr}(a)}} \tag{11}$$

$$F(a,\ b,\ \delta,\ k) \equiv 1 - \beta(\delta, k)\varphi(a, b) \tag{12}$$

$$F'(a,\ b,\ \delta,\ k) \equiv 1 + \beta(\delta, k)\varphi(a, b)$$

$$\varphi(a,\ b) \equiv \sqrt{\frac{\widehat{\Pr}(a) - \widehat{\Pr}(a, b)}{\widehat{\Pr}(a, b)\{(n + \beta(\delta, k)^2)\widehat{\Pr}(a) - \beta(\delta, k)^2\}}} \tag{13}$$

where $\beta(\delta, k)$ represents a positive value which is related to the confidence region and is obtained by numerical integration. k represents the number of true probabilities each of which is satisfied by at least an example in the data set minus 1. We have also proposed an efficient discovery algorithm based on pruning.

3.3 Threshold Scheduling for PADRE

PADRE requires appropriate specification of values for five thresholds $\theta_1^{\mathrm{S}},\ \theta_2^{\mathrm{S}},\ \theta_1^{\mathrm{F}},\ \theta_2^{\mathrm{F}},\ \theta_2^{\mathrm{I}}$. A strict specification for a data set with a small number of exception rules can result in no discovery of exception rules. On the other hand, a loose specification for a data set with a large number of exception rules can result in discovery of many exception rules and the computation process is typically time-consuming. These problems come from the fact that there are five evaluation

criteria $\widehat{\Pr}(Y_\mu)$, $\widehat{\Pr}(Y_\mu Z_\nu)$, $\widehat{\Pr}(x|Y_\mu)$, $\widehat{\Pr}(x'|Y_\mu Z_\nu)$, $\widehat{\Pr}(x'|Z_\nu)$ of a discovered pattern.

For this problem, we have first invented a data structure which can manage discovered patterns with multiple criteria [23]. The data structure is based on a height-balanced tree. In order to realize flexible scheduling, we assign a tree for each index. A node of a tree represents a pointer to a discovered pattern, and this enables fast transformation of the tree. We show, in Figure 1, an example of this data structure which manages seven rule pairs $r1, r2, \cdots, r7$.

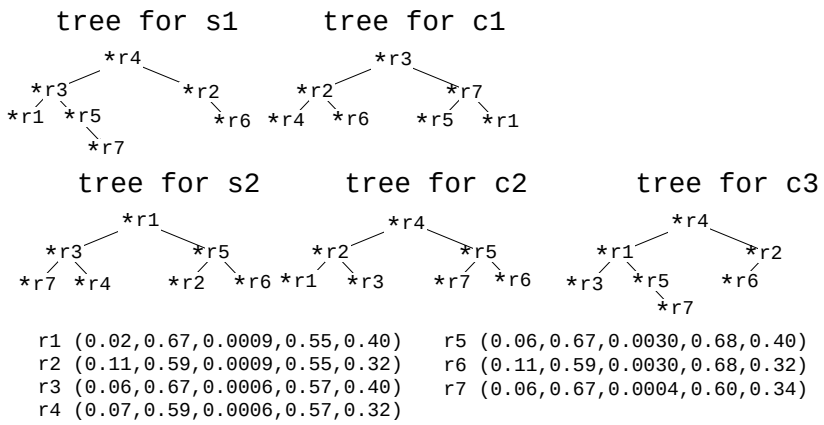

Fig. 1. Illustration based on proposed data structure, where an AVL tree is employed as a balanced search-tree, and keys are s1 ($\widehat{\Pr}(Y_\mu)$), s2 ($\widehat{\Pr}(Y_\mu, Z_\nu)$), c1 ($\widehat{\Pr}(x|Y_\mu)$), c2 ($\widehat{\Pr}(x'|Y_\mu, Z_\nu)$), c3 ($\widehat{\Pr}(x'|Z_\nu)$). Numbers in a pair of parentheses represent values of these indices of the corresponding rule pair. In a tree, $*r1, *r2, \cdots, *r7$ represent a pointer to r1, r2, $\cdots$, r7 respectively.

In our method, a height-balanced tree is assigned to each evaluation criterion, and a node of a tree represents a pointer to a discovered pattern. Then we have proposed an algorithm which updates thresholds and discovers at most η rule pairs [23]. In this method, each time an $(\eta + 1)$-th rule pair is discovered, the worst rule pair in terms of the current criterion is deleted and the value of the criterion is updated according to the deleted rule pair. Each time this process occurs, the current criterion is replaced by another criterion.

3.4 Meta Pattern Study for Knowledge Representation

In the methods in sections 3.1 and 3.2, the discovered pattern is restricted to (1). We classified exception/deviation structures for discovery of interesting patterns based on a meta pattern and proposed an efficient algorithm which discovers all structures [27, 30].

In this study, we defined that a rule $u \rightarrow v$ satisfies

$$\widehat{\Pr}(u) \geq \theta_{\mathrm{S}} \text{ (generality) } \& \ \widehat{\Pr}(v|u) \geq \theta_{\mathrm{F}} \text{ (accuracy)} \tag{14}$$

Similarly, we also introduced a negative rule as $u \not\rightarrow v$, where θ_{I} is a threshold given by the user.

$$\widehat{\Pr}(u) \geq \theta_{\mathrm{S}} \ \& \ \widehat{\Pr}(v|u) \leq \theta_{\mathrm{I}} \tag{15}$$

In the study, an exception/deviation structure is defined as a rule triple $t(y,\ x,\ \alpha,\ \beta,\ \gamma,\ \delta)$, which represents the meta pattern, using literals $x,\ y,\ z$. A strong rule, an exception rule, and a reference rule are defined as $y \rightarrow x$, $\alpha \not\rightarrow \beta$, and $\gamma \rightarrow \delta$ respectively.

$$t(y,\ x,\ \alpha,\ \beta,\ \gamma,\ \delta) = (y \rightarrow x,\ \alpha \not\rightarrow \beta,\ \gamma \rightarrow \delta) \tag{16}$$

where each of α, β, γ, and δ represents a meta variable which is instantiated by variables x, y, and z, resulting definition of various exception/deviation structures[3]. Here $y \rightarrow x$ represents a rule and shows that $\widehat{\Pr}(y)$ and $\widehat{\Pr}(x|y)$ are greater than their respective thresholds. On the other hand, $\alpha \not\rightarrow \beta$ represents a negative rule and shows that $\widehat{\Pr}(\alpha)$ is greater than its threshold, and $\widehat{\Pr}(\beta|\alpha)$ is smaller than its threshold.

$$(\alpha, \beta, \gamma, \delta) \in \{(z, x, y, z), (z, x, z, y), (x, z, z, y), (y, z, z, x)\}. \tag{17}$$

Under appropriate assumptions, our exception/deviation structures can be classified into the eleven structures which are shown in Figure 2.

Association rule discovery [2] assumes a transaction data set, which has only binary attributes. Each attribute can take either "y" or "n" as its value, and most of the attribute values are "n". The sparseness of the data set allows to employ breadth-first search. Note that if breadth-first search were employed for an ordinary data set, the number of large item sets would be huge. In such a case, space efficiency would be so poor that any breadth-first algorithm would be impractical. We assume an ordinary data set and propose an algorithm which performs depth-first search for triplets of literals a, b, c. The number of atoms in a literal a is represented as $|a|$. We consider $|a|, |b|, |c| \leq M$ as the search restriction in the above algorithm. Time efficiency of this algorithm is $O(m^{3M})$, where m is the number of attribute in D. This is justified since this algorithm is complete in the sense that it discovers all rule triplets. This inefficiency is remedied by a novel pruning procedure [27, 30].

4 Experimental Evaluation

4.1 Time-efficiency of MEPRO

MEPRO has been applied to the "mushroom" data set (8,124 examples, 22 attributes, 2 classes) and the "Congressional Voting Records" data set (435

[3] Further investigation of other meta-patterns seems plausible.

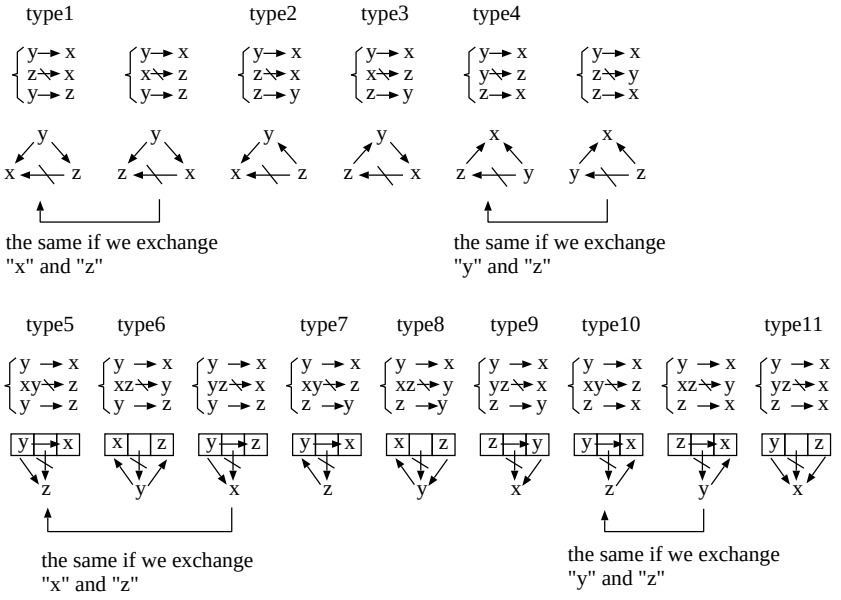

Fig. 2. Possible rule triplets. A rectangle on the top center for each triplet represents a conjunction of literals in the top right and left

examples, 16 attributes) in the UCI Repository and discovered exception rules each of which is interesting at least from the statistical point of view. The upper bound turned out to be useful since it achieved more than 5 times of speed-up for the "Congressional Voting Records" data set. The run time for the voting data set with the maximum depth 5 is approximately 66 seconds.

4.2 Effectiveness of PADRE

The first series of experiments are intended to investigate PADRE's performance and effectiveness of the sound pruning with different sets of thresholds using the "mushroom" data set. We first applied PADRE without its pruning method, and found it unacceptably inefficient: the computational time for a single trial is estimated more than 80 days under a typical condition. Therefore, we use PADRE with its pruning method, and compare the case of $\delta = 0.95$ with $\delta = 0$. Comparing these cases would justify the extra cost of calculating these coefficients.

Table 1 shows the number of rule pairs found by PADRE and the number of nodes searched by PADRE with $\delta = 0.95$ and $\delta = 0$. In the experiments, we used $\theta_2^{\mathrm{F}} = 1.0,\ M = 3$. For the other parameters $\theta_1^{\mathrm{S}},\ \theta_2^{\mathrm{S}},\ \theta_1^{\mathrm{F}},\ \theta_2^{\mathrm{F}}$, standard values are settled to 0.25, 0.05, 0.70, 0.50 respectively, and we modified one of them in each trial. Considering the meaning of accuracy, we substituted $\widehat{\Pr}(c),\ \widehat{\Pr}(c'),\ \widehat{\Pr}(c')$ for $\theta_1^{\mathrm{F}},\ \theta_2^{\mathrm{F}},\ \theta_2^{\mathrm{I}}$ when $\widehat{\Pr}(c) > \theta_1^{\mathrm{F}},\ \widehat{\Pr}(c') > \theta_2^{\mathrm{F}},\ \widehat{\Pr}(c') < \theta_2^{\mathrm{I}}$ respectively. Here,

Table 1. Performance of PADRE with/without the reliability evaluation in the stopping criteria ($\delta = 0.95$ and $\delta = 0$ respectively). The mushroom data set is employed, and $\theta_2^{\mathrm{F}} = 1.0, \delta = 0.95, M = 3$. For the other parameters $\theta_1^{\mathrm{S}}, \theta_2^{\mathrm{S}}, \theta_1^{\mathrm{F}}, \theta_2^{\mathrm{F}}$, standard values are settled to 0.25, 0.05, 0.70, 0.50 respectively, and we modified one of them in each trial. A reduction rate represents the number of nodes with the reliability evaluation ($\delta = 0.95$) divided by the number of nodes without it ($\delta = 0$).

parameter	value	# of rules	# of nodes ($*10^6$)		reduction
			δ=0.95	δ=0	rate (%)
	(standard)	3	3.26	4.18	78.0
θ_1^{S}	0.05	3	6.32	7.78	81.2
	0.10	3	5.30	6.60	80.3
	0.15	3	4.40	5.44	80.8
	0.20	3	3.81	4.67	81.6
θ_2^{S}	0.01	7,445	20.20	23.93	84.4
	0.02	683	11.58	13.30	87.0
	0.03	3	7.28	8.51	85.4
	0.04	3	4.90	5.78	84.7
θ_1^{F}	0.50	63	3.86	4.75	81.2
	0.55	28	3.70	4.53	81.6
	0.60	26	3.60	4.42	81.5
	0.65	18	3.51	4.26	82.5
θ_2^{I}	0.55	6	3.35	4.23	79.1

$\widehat{\Pr}(\text{edibleclass} = e)$, $\widehat{\Pr}(\text{edibleclass} = p) < 0.55$. Thus, we did not investigate for cases $\theta_2^{\mathrm{I}} > 0.55$ since they are equivalent with the case $\theta_2^{\mathrm{I}} = 0.55$.

The Table shows that considering coefficient in the stopping criteria improves time efficiency approximately 20%. The amelioration is less effective when θ_2^{S} is modified, but is still more than 12%. We can safely conclude that considering coefficients $G(\cdot)$ in the stopping criteria is effective in these experiments for efficient discovery of exception rules. From the Table, we also see that the number of discovered rule pairs is below 100 unless θ_2^{S} is settled to either 0.02 or 0.01. This shows that output of PADRE can be typically inspected by a human unlike conventional rule-discovery methods. It should be also noted that the numbers of searched rule pairs are often moderate: they are below $1.00 * 10^7$ except for the two cases, and can be searched within at most a few days even with a personal computer.

In the second series of experiments, we have shown that PADRE can be used to justify absence of strong exceptions in a data set. Using the census data set with various settings on parameter values, PADRE finds rule pairs each of which shows intuitively mild exceptions on annual salary. We attribute this result to the very nature of the problem that exceptions in salaries are rare.

In the third series of experiments, we show effectiveness of PADRE in terms of discovery of interesting rules by briefly explaining our endeavor in a data mining contest with the meningitis data set [25]. The data set consists of 140

patients each of whom is described by 38 attributes and has been made public as a benchmark problem to the data mining community. Our method has discovered 169 rule pairs from a pre-processed version of this data set [25]. These rule pairs were inspected by Dr. Tsumoto, who is a domain expert, and each rule pair was assigned a five-rank score for the following evaluation criteria each of which were judged independently.

- validness: the degree that the discovered pattern fits domain knowledge
- novelty: the degree that the discovered pattern does not exist in domain knowledge
- usefulness: the degree that the discovered pattern is useful in the domain
- unexpectedness: the degree that the discovered pattern partially contradicts domain knowledge

For the scores, five and one represent the best score and the worst score respectively. We show the results classified by the attributes in the conclusions in Table 2.

Table 2. Average performance of the proposed method with respect to attributes in the conclusion. The column "#" represents the number of discovered rule pairs.

attribute	#	validness	novelty	unexpectedness	usefulness
(all)	169	2.9	2.0	2.0	2.7
CULTURE	2	1.0	1.0	1.0	1.0
C_COURSE	1	1.0	1.0	1.0	1.0
RISK	1	1.0	1.0	1.0	1.0
FOCAL	18	3.1	2.2	2.7	3.0
LOC_DAT	11	2.5	1.8	1.8	2.5
Diag2	72	3.0	1.1	1.1	2.6
CT_FIND	36	3.3	3.0	3.0	3.2
EEG_FOCUS	11	3.0	2.9	2.9	3.3
Course (G)	8	1.8	2.0	2.0	1.8
CULT_FIND	4	3.3	4.0	4.0	3.5
KERNIG	4	2.0	3.0	3.0	2.0
SEX	1	2.0	3.0	3.0	2.0

From the Table, we see that the average scores of the discovered rule pairs are high for several attributes in the conclusions. We inspected these rule pairs by grouping them with respect to the attribute in the conclusion, and found that these attributes can be classified into four categories. The first category represents attributes with the lowest scores, and includes CULTURE, C_COURSE, and RISK. We consider that attributes in this category cannot be explained with this data set, and investigation on them requires further information on other attributes. The second category represents attributes with higher scores

for validness and usefulness, and includes FOCAL, LOC_DAT, and Diag2. We consider that attributes in this category can be explained with this data set, and has been well investigated probably due to their importance in this domain. We regard them as one of important targets in discovery although one will often rediscover conventional knowledge. The third category represents attributes with approximately equivalent scores, and includes CT_FIND, EEG_FOCUS, and Course (G). We consider that attributes in this category can be explained with this data set, and has not been investigated well in spite of their importance in this domain. We regard them as one of the most important targets in discovery. The fourth category represents attributes with higher scores for novelty and unexpectedness, and includes CULT_FIND, KERNIG, and SEX. We consider that attributes in this category can be explained with this data set, but has been somewhat ignored. We consider that investigating these attributes using discovered rule sets can lead to interesting discoveries which might reveal unknown mechanisms in this domain in spite of their apparent low importance.

As Dr. Tsumoto admits, this is due to the fact that the structure of a rule pair is useful for discovery of interesting patterns. According to him, our method discovered the most interesting results in the data mining contest [31].

In the fourth series of experiments, our method has been applied to 1994 bacterial test data set (20,919 examples, 135 attributes, 2 classes) [26]. We have found that we need to consider distribution of attribute values and cause and effect relationships in order to discover interesting patterns from the data set. However, this application shows that our method is adequate in terms of efficiency in exception rule mining from a relatively large-scale data set. The execution time is less than 4 hours with the number of rules 500 and the attribute in the conclusions "CBPs" with the threshold scheduling in section 3.3.

4.3 Effectiveness of Threshold Scheduling for PADRE

Our method has been applied to four data sets, which have different characteristics with respect to the number of discovered rule-pairs. Experimental results clearly show that our method is effective for data sets with many rule pairs as well as for data sets with few rule pairs. We have also confirmed that specification of values for five thresholds can be replaced by an easier procedure of specifying the largest number of discovered rule pairs.

Due to space constraint, we only show results with the mushroom data set, where the edibility is the only attribute allowed in the conclusions. In the experiments, we settled the maximum number of discovered rule-pairs to $\eta = 500$, the maximum search-depth to $M = 5$, and the maximum search-depth to $M = 6$. A large number of rule pairs can be discovered from the mushroom data set. However, a user does not know this fact at analyzing this data set for the first time. We settled initial values of the thresholds loosely to $\theta_1^{\mathrm{S}} = 0.0004$, $\theta_2^{\mathrm{S}} = 10/8124$, $\theta_1^{\mathrm{F}} = 0.5$, $\theta_2^{\mathrm{F}} = 0.5$, $\theta_2^{\mathrm{I}} = 0.5$, and applied the proposed method. As the results, 500 rule pairs were discovered, and the final values of the thresholds were $\theta_1^{\mathrm{S}} = 0.154$, $\theta_2^{\mathrm{S}} = 0.00295$, $\theta_1^{\mathrm{F}} = 0.914$, $\theta_2^{\mathrm{F}} = 1.000$, $\theta_2^{\mathrm{I}} = 0.123$. In the

application, $5.00 * 10^6$ nodes were searched, and the number of rule pairs stored in the data structure (including those which were deleted) were 4069.

We have applied our method with stricter initial values $\theta_1^{\mathrm{S}} = 0.1,\ \theta_2^{\mathrm{S}} = 100/8124,\ \theta_1^{\mathrm{F}} = 0.7,\ \theta_2^{\mathrm{F}} = 0.9,\ \theta_2^{\mathrm{I}} = 0.5$ by varying the maximum number of rule pairs $\eta = 100, 200, \cdots, 500$. Table 3 shows numbers of searched nodes, numbers of rule pairs stored in the data structure (including those which were deleted), and the final values of the thresholds. We see, from the table, that the final values of the thresholds are valid, and the number of nodes are reduced to 44 - 56 %.

Table 3. Resulting statistics with respect to the maximum number η of rule pairs in the experiments with the mushroom data set, where initial values were settled as $\theta_1^{\mathrm{S}} = 0.1,\ \theta_2^{\mathrm{S}} = 100/8124,\ \theta_1^{\mathrm{F}} = 0.7,\ \theta_2^{\mathrm{F}} = 0.9,\ \theta_2^{\mathrm{I}} = 0.5$

η	# of nodes	# of rule pairs	θ_1^{S}	θ_2^{S}	θ_1^{F}	θ_2^{F}	θ_2^{I}
100	$2.28*10^6$	734	0.177	0.015	0.783	1.000	0.228
200	$2.51*10^6$	1250	0.211	0.015	0.809	1.000	0.240
300	$2.63*10^6$	1662	0.248	0.017	0.804	1.000	0.260
400	$2.71*10^6$	1893	0.241	0.015	0.804	1.000	0.272
500	$2.81*10^6$	2245	0.217	0.015	0.794	1.000	0.272

In order to investigate the effectiveness of updating values of thresholds, we have applied our method without deleting a node in the AVL trees with initial values $\theta_1^{\mathrm{S}} = 0.0004,\ \theta_2^{\mathrm{S}} = 10/8124,\ \theta_1^{\mathrm{F}} = 0.5,\ \theta_2^{\mathrm{F}} = 0.5,\ \theta_2^{\mathrm{I}} = 0.5$. In this experiment, $8.94 * 10^6$ nodes were searched, and $1.48 * 10^5$ rule pairs were discovered. We see that without updating values of thresholds, computational time nearly doubles, and the number of discovered rule-pairs is huge.

4.4 Evaluation of Meta Pattern Study for Knowledge Representation

In our experiments using 15 UCI data sets (car, nursery, postoperative, vote, breastcancer, mushroom, credit, abalone, diabetes, yeast, australian, shuttle, hepatitis, german, thyroid), we deleted attributes that have only one value in a data set. In applying our algorithm, the number of discretization bins was set to 4. Other parameters were set to $\theta_{\mathrm{S}} = 0.025$, $\theta_{\mathrm{F}} = 0.7$, $\theta_{\mathrm{I}} = 0.6$, and $M = 2$. Figure 3 summarizes the results of experiments.

The left hand-side of Figure 3 show that pruning is effective, since without pruning the number of searched nodes increase by 5 % ("nursery" and "diabetes") to 285 % ("mushroom"). This is due to the fact that a considerable number of nodes in a search tree tend to have small probabilities for their literals and are thus pruned. Numbers of discovered rule triplets per types reveal interesting tendencies. From the right-hand side of the figure, we see that types 3, 4, 6, 7, 10 are extremely numerous: they are more than $1 * 10^5$ in 11 data

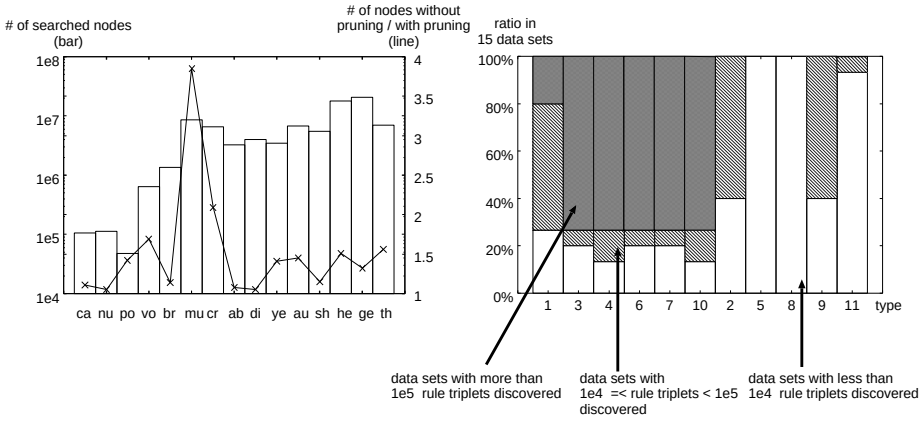

Fig. 3. Experimental results, where "1ec" represents 10^c

sets. Type 1 is also numerous since it is more than $1 * 10^5$ in 3 data sets. On the other hand, types 2 and 9 are modest in number: they never exceed $1 * 10^5$ in any data sets, and exceed $1 * 10^4$ in 9 data sets. Finally, types 11, 8, and 5 are rare in this order: type 11 exceeds $1 * 10^4$ in 1 data set, and types 8 and 5 never exceed $1 * 10^4$ in any data sets. Similar tendencies were observed for $M = 1$. Interestingly, we had anticipated the exceptionality of types 2, 5, 8, 9, and 11 as stronger than the other types.

5 Evaluation Scheme for Exception Rule/Group Discovery

Currently, the articles each of which proposes an exception rule/group discovery method, typically evaluates its method by decrease of discovered rules in number or a success in a case study. Decrease of discovered rules in number represents an important quantitative criterion which typically results in reduction of inspection costs for output of data mining. However, this evaluation criterion has a serious deficiency that it neglects interestingness of discovered rules.

In [29], we have proposed evaluation scheme for exception rule/group discovery. The evaluation scheme consist of seven criteria: generality, monotonicity, reliability, search range, interpretation of the evaluation measure, use of domain knowledge, and successes in real applications.

Generality represents possibility that a data mining method can be applied to a problem with a small amount of effort, and has been chosen as an evaluation criterion. We have settled the categories of generality as low (cannot be applied to many domains), middle (employ domain knowledge), and high (no domain knowledge required). Note that a method which employs domain knowledge belongs to the category "high" if it can be executed without domain knowledge.

If an exception rule is represented by $yz \rightarrow x$, the value of $\Pr(x|yz)$ typically differs from those of $\Pr(x|y)$ and $\Pr(x|z)$ considerably. For instance when the value of $\Pr(x|yz)$ is nearly 1, the values of $\Pr(x|y)$ and $\Pr(x|z)$ are often nearly 0. In such a case the exception rule $yz \rightarrow x$ is said to break the monotonicity of the rules $y \rightarrow x$ and $z \rightarrow x$ [4]. We have settled the categories of this measure as low (unknown), middle (one direction), and high (multiple directions).

As described in section 3.2, a pattern discovered from 100 examples and a pattern discovered from 10,000 examples differ in their reliability. The latter can be considered as relatively reliable, but the former might be an unreliable pattern which has occurred by chance. This kind of problem can be typically resolved by a statistical approach for reliability evaluation. We have settled the categories of this measure as low (none), middle (done for a criterion), and high (done for multiple criteria).

It is empirically known that an exception rule/group with good statistics is rare, and nothing is discovered with a small search range. We consider this fact important and have chosen search range as an evaluation criterion. We have settled the categories of this measure as low (partial), middle (all), and high (all but is adjusted by the size of the discovery problem).

If the evaluation measure has a clear interpretation, it would be easy to explain discovered exception rules/groups with it. An evaluation measure can be classified into theoretical and empirical, and the latter is often difficult to be interpreted. Since the meaning of discovered patterns is an important issue, we have chosen interpretation of the evaluation measure as an evaluation criterion. We have settled the categories of this measure as middle (empirical) and high (interpretable).

In discovering interesting exception rules/groups using domain knowledge, it is desirable that the use necessitates less efforts. However, we must be prudent since the use might hinder us from discovering interesting exception rules/groups. In order to handle these aspects, we have chosen use of domain knowledge as an evaluation criterion. We include intervention of a domain expert to a discovery system in this criterion. We have settled the categories of this measure as low (usage is possible), middle (usage is direct), and high (usage is sophisticated).

It should be noted that simply observing decrease of discovered patterns in number, even in a real application, cannot be regarded as a success. An application is judged as a success if a set of interesting patterns for a domain expert is discovered. We have settled the categories of this measure as low (none), middle (one domain), and high (multiple domains).

6 Conclusions

In this paper, we have summarized a part of our endeavor for discovery of interesting exception rules. Our objective was to discover a set of (possibly) interesting rule pairs from a give data set in table format. We believe that our results

[4] Thanks to A. Tuzhilin.

on interestingness measure, reliability evaluation, practical application, parameter reduction, and knowledge representation represent a step toward undirected discovery of interesting rules.

We believe that some sort of interestingness can be systematically captured by syntax. The fact that the structure of a rule pair is useful for discovery of interesting patterns represents a convincing example.

References

1. G. Adomavicius and A. Tuzhilin, "Discovery of Actionable Patterns in Databases: The Action Hierarchy Approach", *Proc. Third Int'l Conf. Knowledge Discovery and Data Mining (KDD)*, 1997, pp. 111–114.
2. R. Agrawal *et al.*, "Fast Discovery of Association Rules", *Advances in Knowledge Discovery and Data Mining*, eds. U.M. Fayyad *et al.*, AAAI/MIT Press, Menlo Park, Calif., 1996, pp. 307–328.
3. S. Brin *et al.*, "Dynamic Itemset Counting and Implication Rules for Market Basket Data", *Proc. ACM SIGMOD Int'l Conf. Management of Data*, 1997, pp. 255–264.
4. J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and Unsupervised Discretization of Continuous Features", *Proc. Twelfth Int'l Conf. Machine Learning (ICML)*, Morgan Kaufmann, San Francisco, 1995, pp. 194–202.
5. U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery: An Overview", *Advances in Knowledge Discovery and Data Mining*, eds. U. M. Fayyad *et al.*, AAAI/MIT Press, Menlo Park, Calif., 1996, pp. 1-34.
6. J. Fürnkranz, "Separate-and-Conquer Rule Learning", *Artificial Intelligence Review*, **13**, 1999, pp. 3–54.
7. R. Gras, "*L' Implication Statistique*", La Pensée Sauvage, France, 1996 (in French).
8. F. Hussain *et al.*, "Exception Rule Mining with a Relative Interestingness Measure", *Knowledge Discovery and Data Mining, LNAI 1805 (PAKDD)*, Springer, Berlin, 2000, pp. 86–97.
9. M. Klemettinen *et al.*, "Finding Interesting Rules from Large Sets of Discovered Association Rules", *Proc. Third Int'l Conf. Information and Knowledge Management (CIKM)*, 1994, pp. 401–407.
10. W. Klösgen, "Explora: A Multipattern and Multistrategy Discovery Approach", *Advances in Knowledge Discovery and Data Mining*, eds. U. M. Fayyad *et al.*, AAAI/MIT Press, Menlo Park, Calif., 1996, pp. 249–271.
11. B. Liu *et al.*, "Finding Interesting Patterns Using User Expectations", *IEEE Trans. Knowledge and Data Eng.*, **11**, 1999, pp. 817–832.
12. B. Liu, W. Hsu, and Y. Ma, "Pruning and Summarizing the Discovered Associations", *Proc. Fifth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD)*, 1999, pp. 125–134.
13. J. A. Major and J. J. Magano, "Selecting among Rules Induced from a Hurricane Database", *Proc. AAAI-93 Workshop Knowledge Discovery in Databases*, 1993, pp. 28–44.
14. B. Padmanabhan and A. Tuzhilin, "A Belief-Driven Method for Discovering Unexpected Patterns", *Proc. Fourth Int'l Conf. Knowledge Discovery and Data Mining (KDD)*, AAAI Press, Menlo Park, Calif., 1998, pp. 94–100.
15. G. Piatetsky-Shapiro, "Discovery, Analysis, and Presentation of Strong Rules", *Knowledge Discovery in Databases*, eds. G. Piatetsky-Shapiro and W. J. Frawley, AAAI/MIT Press, Menlo Park, Calif., 1991, pp. 229–248.

16. G. Piatetsky-Shapiro and C. J. Matheus, "The Interestingness of Deviations", *AAAI-94 Workshop on Knowledge Discovery in Databases*, Tech Rep. WS-94-03, AAAI Press, Menlo Park, Calif., 1994, pp. 25–36.
17. A. Silberschatz and A. Tuzhilin, "What Makes Patterns Interesting in Knowledge Discovery Systems", *IEEE Trans. Knowledge and Data Eng.*, **8**, 1996, pp. 970–974.
18. P. Smyth and R. M. Goodman, "An Information Theoretic Approach to Rule Induction from Databases", *IEEE Trans. Knowledge and Data Eng.*, **4**, 1992, pp. 301–316.
19. E. Suzuki and M. Shimura, "Exceptional Knowledge Discovery in Databases Based on Information Theory", *Proc. Second Int'l Conf. Knowledge Discovery and Data Mining (KDD)*, AAAI Press, Menlo Park, Calif., 1996, pp. 275–278.
20. E. Suzuki, "Discovering Unexpected Exceptions: A Stochastic Approach", *Proc. Fourth International Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery (RSFD)*, pp. 225–232, 1996.
21. E. Suzuki, "Autonomous Discovery of Reliable Exception Rules", *Proc. Third Int'l Conf. Knowledge Discovery and Data Mining (KDD)*, AAAI Press, Menlo Park, Calif., 1997, pp. 259–262.
22. E. Suzuki and Y. Kodratoff, Discovery of Surprising Exception Rules Based on Intensity of Implication, Principles of Data Mining and Knowledge Discovery, LNAI 1510 (PKDD), Springer (1998) 10–18.
23. E. Suzuki, "Scheduled Discovery of Exception Rules", *Discovery Science, LNAI 1721 (DS)*, Springer, Berlin, 1999, pp. 184–195.
24. E. Suzuki and S. Tsumoto, "Evaluating Hypothesis-Driven Exception-Rule Discovery with Medical Data Sets", *Knowledge Discovery and Data Mining, LNAI 1805 (PAKDD)*, Springer, Berlin, 2000, pp. 208–211.
25. E. Suzuki and S. Tsumoto, Evaluating Hypothesis-driven Exception-rule Discovery with Medical Data Sets, Knowledge Discovery and Data Mining, LNAI 1805 (PAKDD), Springer (2000) 208–211.
26. E. Suzuki, "Mining Bacterial Test Data with Scheduled Discovery of Exception Rules", *Proc. Int'l Workshop of KDD Challenge on Real-world Data (KDD Challenge)*, Kyoto, Japan, 2000, pp. 34–40.
27. E. Suzuki and J.M. Żytkow, Unified Algorithm for Undirected Discovery of Exception Rules, Principles of Data Mining and Knowledge Discovery, LNAI 1910 (PKDD), Springer (2000) 169–180.
28. E. Suzuki: "Undirected Discovery of Interesting Exception Rules", International Journal of Pattern Recognition and Artificial Intelligence, Vol. 16, No. 8, pp. 1065–1086, 2002.
29. E. Suzuki: "Evaluation Scheme for Exception Rule/Group Discovery", Intelligent Technologies for Information Analysis, Springer (accepted for publication).
30. E. Suzuki: "Unified Algorithm for Undirected Discovery of Exception Rules", International Journal of Intelligent Systems (accepted for publication).
31. S. Tsumoto *et al.*, "Comparison of Data Mining Methods using Common Medical Datasets", *ISM Symp.: Data Mining and Knowledge Discovery in Data Science*, 1999, pp. 63–72.
32. N. Yugami, Y. Ohta, and S. Okamoto, "Fast Discovery of Interesting Rules", *Knowledge Discovery and Data Mining, LNAI 1805 (PAKDD)*, Springer, Berlin, 2000, pp. 17–28.