Considerations on Rule Induction Procedures by STRIM and Their Relationship to VPRS

Yuichi Kato[1], Tetsuro Saeki[2], and Shoutarou Mizuno[1]

[1] Shimane University,
1060 Nishikawatsu-cho, Matsue City, Shimane 690-8504, Japan
ykato@cis.shimane-u.ac.jp
[2] Yamaguchi University,
2-16-1 Tokiwadai, Ube City, Yamaguchi 755-8611, Japan
tsaeki@yamaguchi-u.ac.jp

Abstract. STRIM (Statistical Test Rule Induction Method) has been proposed as a method to effectively induct if-then rules from the decision table. The method was studied independently of the conventional rough sets methods. This paper summarizes the basic notion of STRIM and the conventional rule induction methods, considers the relationship between STRIM and their conventional methods, especially VPRS (Variable Precision Rough Set), and shows that STRIM develops the notion of VPRS into a statistical principle. In a simulation experiment, we also consider the condition that STRIM inducts the true rules specified in advance. This condition has not yet been studied, even in VPRS. Examination of the condition is very important if STRIM is properly applied to a set of real-world data set.

1 Introduction

Rough Sets theory as introduced by Pawlak [1] provides a database called the decision table, with various methods of inducting if-then rules and determining the structure of rating and/or knowledge in the database. Such rule induction methods are needed for disease diagnosis systems, discrimination problems, decision problems, and other aspects, and consequently many effective algorithms for rule induction by rough sets have been reported to date [2–7]. However, these methods and algorithms have paid little attention to mechanisms of generating the database, and have generally focused on logical analysis of the given database.

In a previous studies [8, 9] we (1) devised a model of data generation for the decision table with if-then rules specified in advance, and proposed a statistical rule induction method and an algorithm named STRIM; (2) In a simulation experiment based on the model of the data generation, STRIM was confirmed to successfully induct the if-then true rules from different databases generated from the same specified rules[8]; (3) found that, when conventional methods [4, 6, 7] were used, significant rules could barely be inducted, and different rule sets were inducted from different sample data sets with the same rules; i.e.

M. Kryszkiewicz et al. (Eds.): RSEISP 2014, LNAI 8537, pp. 198–208, 2014.

Table 1. An example of a decision table

U	$C(1)$	$C(2)$	$C(3)$	$C(4)$	$C(5)$	$C(6)$	D
1	5	6	3	2	4	2	3
2	2	5	6	1	2	4	6
3	1	1	6	2	2	6	1
4	4	1	6	6	4	6	6
5	4	4	5	5	4	1	4
...	...	...	...	...	...	...	...
$N-1$	1	5	1	2	5	2	2
N	5	1	3	1	3	5	4

when using these methods, results were highly dependent on the sample set; (4) considered the data size of the decision table needed to induct the true rules with the probability of w [9], since conventional methods have not examined this problem.

This paper briefly summarizes STRIM and the conventional methods and investigates the principle of STRIM in more depth by using an example simulation experiment. Specifically, this paper summarizes STRIM as a two stage process. The first stage finds rule candidates, and the second arranges these candidates. We then consider the relationship between the principle of rule induction in the first stage and that in the conventional method, especially in VPRS (Variable Precision Rough Set) [5]. STRIM was developed independently from the conventional rough set theory. Our considerations show that STRIM can develop the notion of VPRS into a statistical principle, and the admissible classification error in VPRS corresponds to the significance level of the statistical test by STRIM. Consideration of the statistacl approach by Jaworski [10] is complex and difficult to understand, since he studies the confident intervals of accuracy and coverage of the rules inducted by VPRS.

We further examine the validity of the second process and the arrangement by STRIM, based on a statistical model which clearly shows the standard of the arrangement. In contrast, that achieved by VPRS is shown to be to not so clear, and the analyst studying the matter is required to make a decision. Both considerations in the two processes are also illustrated by the results of the simulation experiment, and seem to be useful for understanding and/or interpreting the results when analyzing real-world data sets.

2 Data Generation Model and Decision Table

Rough Sets theory is used for inducting if-then rules hidden in the decision table S. S is conventionally denoted $S = (U, A = C \cup \{D\}, V, \rho)$. Here, $U = \{u(i)|i = 1, ..., |U| = N\}$ is a sample set, A is an attribute set, $C = \{C(j)|j = 1, .., |C|\}$ is a condition attribute set, $C(j)$ is a member of C and a condition attribute, and D is a decision attribute. V is a set of attribute values denoted by $V = \bigcup_{a \in A} V_a$ and is characterized by an information function ρ: $U \times A \rightarrow V$. Table 1 shows the example where $|C| = 6$, $|V_{a=C(J)}| = M_{C(j)} = 6$, $|V_{a=D}| = M_D = 6$, $\rho(x = u(1), a = C(1)) = 5$, $\rho(x = u(2), a = C(2)) = 5$, and so on.

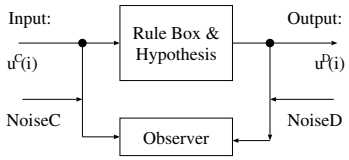

Fig. 1. Relationship between the condition attributes' value and the decision attribute's value

Table 2. Hypothesis with regard to the decision attribute value

Hypothesis 1	$u^C(i)$ coincides with $R(k)$, and $u^D(i)$ is uniquely determined as $D = d(k)$ (uniquely determined data).
Hypothesis 2	$u^C(i)$ does not coincide with any $R(d)$, and $u^D(i)$ can only be determined randomly (indifferent data).
Hypothesis 3	$u^C(i)$ coincides with several $R(d)$ $(d = d1, d2, ...)$, and their outputs of $u^C(i)$ conflict with each other. Accordingly, the output of $u^C(i)$ must be randomly determined from the conflicted outputs (conflicted data).

STRIM considers the decision table to be a sample data set obtained from an input-output system including a rule box (Fig. 1) and a hypothesis regarding the decision attribute values (Table 2). A sample $u(i)$ consists of its condition attributes values of $|C|$-tuple $u^C(i)$ and its decision attribute $u^D(i)$. $u^C(i)$ is the input into the rule box, and is transformed into the output $u^D(i)$ using the rules contained in the rule box and the hypothesis. For example, specify the following rules in the rule box as true rules to be inducted by STRIM introduced in 3:

$R(d)$: if Rd then $D = d$, $(d = 1, ..., M_D = 6)$,

where $Rd = (C(1) = d) \wedge (C(2) = d) \vee (C(3) = d) \wedge (C(4) = d)$. Generate $u^C(i) = (v_{C(1)}(i), v_{C(2)}(i), ..., v_{C(|C|)}(i))$ of $u(i)$ $(i = 1, ..., |U| = N)$ by use of random numbers with a uniform distribution, and then $u^D(i)$ is determined using the rules specified in the rule box and the hypothesis. For example, Hypothesis 1 is applied to $u(3)$ and $u(4)$, Hypothesis 2 is to $u(1)$, $u(2)$, $u(N-1)$, $u(N)$ and Hypothesis 3 is to $u(5)$ in Table 1. In contrast, $u(i) = (u^C(i), u^D(i))$ is measured by an observer, as shown in Fig. 1. Existence of *NoiseC* and *NoiseD* makes missing values in $u^C(i)$, and changes $u^D(i)$ to create another values of $u^D(i)$, respectively. This model is closer to the real-world system. However, Table 1 is an example generated by this specification without both noises, for a plain explanation of the system. Inducting if-then rules from the decision table then identifies the rules in the rule box, by use of the set of inputs-output $(u^C(i), u^D(i))$ $(i = 1, ..., |U| = N)$.

3 Summaries of Rule Induction Procedures by STRIM

STRIM inducts if-then rules from the decision table through two processes in separate stages. The first stage process is that of statistically discriminating and

Table 3. An example of a condition part and corresponding frequency of their decision attribute values

trying $CP(k)$	$C(1)$	$C(2)$	$C(3)$	$C(4)$	$C(5)$	$C(6)$	$f = (n_1, n_2, ..., n_6)$	z
1	1	0	0	0	0	0	(469, 240, 275, 238, 224, 226)	12.52
2	2	0	0	0	0	0	(238, 454, 245, 244, 232, 219)	12.12
3	3	0	0	0	0	0	(236, 213, 477, 271, 232, 222)	13.36
4	0	0	0	0	0	6	(289, 277, 300, 255, 296, 296)	0.97
5	1	1	0	0	0	0	(235, 10, 12, 2, 7, 8)	30.77
6	1	2	0	0	0	0	(41, 47, 41, 41, 40, 49)	1.06
7	1	3	0	0	0	0	(46, 52, 57, 60, 51, 40)	1.46
8	1	0	0	0	0	6	(84, 36, 52, 38, 34, 39)	5.95
9	2	1	0	0	0	0	(46, 35, 42, 51, 37, 37)	1.73
10	2	2	0	0	0	0	(8, 227, 6, 11, 4, 6)	30.47
11	2	3	0	0	0	0	(49, 43, 44, 55, 49, 44)	1.30
12	0	0	0	0	6	6	(52, 50, 60, 46, 52, 43)	1.54
13	1	1	0	0	1	0	(38, 1, 3, 1, 0, 1)	12.61
14	1	1	0	0	2	0	(38, 3, 1, 0, 3, 3)	11.81
15	1	1	0	0	5	0	(49, 3, 2, 1, 1, 1)	14.22
16	1	1	0	0	6	0	(46, 0, 3, 0, 0, 1)	14.48
17	1	1	0	0	0	3	(45, 4, 3, 0, 3, 1)	12.97
18	2	2	0	5	0	0	(1, 45, 1, 3, 1, 0)	13.90
19	2	2	0	0	1	0	(0, 55, 0, 2, 0, 1)	16.15
20	2	2	0	0	2	0	(0, 38, 1, 1, 2, 2)	12.61
21	2	2	0	0	4	0	(4, 37, 1, 2, 0, 1)	12.00

separating the set of indifferent data from the set of uniquely determined or conflicted data in the decision table (See Table 2). Specifically, assume $CP(k) = \bigwedge_j$ ($C(j_k) = v_j$) ($\in V_{C(j_k)}$) as the condition part of the if-then rule, and derive the set $U(CP(k)) = \{u(i) | u^C(i)$ satisfies $CP(k)$, which is denoted by $u^{C=CP(k)}(i)$ hereafter $\}$. Also derive $U(m) = \{u(i) | u^{D=m}(i)\}$ $(m = 1, ..., M_D)$. Calculate the distribution $f = (n_1, n_2, ..., n_{M_D})$ of the decision attributes of $U(CP(k))$, where $n_m = |U(CP(k)) \cap U(m)|$ $(m = 1, ..., M_D)$. If the assumed $CP(k)$ does not satisfy the condition $U(Rd) \supseteq U(CP(k))$ (sufficient condition of specified rule Rd) or $U(CP(k)) \supseteq U(Rd)$ (necessary condition), $CP(k)$ only generates the indifferent data set based on Hypothesis 2 in Table 2, and the distribution f does not have partiality of decisions. Conversely, if $CP(k)$ satisfies either condition, f has partiality of the distribution, since $u^D(i)$ is determined by Hypothesis 1 or 3. Accordingly, whether f has partiality or not determines whether the assumed $CP(k)$ is a neither necessary nor sufficient condition. Whether f has partiality or not can be determined objectively by a statistical test of the following null hypothesis $H0$ and its alternative hypothesis $H1$:

$H0$: f does not have partiality. $H1$: f has partiality.

In order to illustrate this concept, Table 3 shows the number of examples of $CP(k)$, $(n_1, n_2, ..., n_{M_D})$ and an index of the partiality by z derived from Table 1 with $N = 10000$. For example, the first row means the following: 100000 denotes $CP(k = 1) = (C(1) = 1)$ (the rule length is $RL = 1$) and its corresponding $f = (496, 240, 275, 238, 224, 226)$ and $z = 12.52$, where,

$$z = \frac{n_d + 0.5 - np_d}{(np_d(1 - p_d))^{0.5}}, \tag{1}$$

```
int main(void) {
int rule[|C|]={0,...,0}; //initialize trying rules
int tail=-1; //initial vale set
input data; // set decision table
rule_check(tail,rule); // 1)-5) strategies
make Pyramid(l) (l=1,2,...) so that every r(k) belongs to
one Pyramid at least; // strategy 6)
make rePyramid(l) (l=1,2,...); // strategy 7)
reduce rePyramid; // strategy 8)
} // end of main

int rule_check(int tail,int rule[|C|]) {
  for (ci=tail+1; cj<|C|; ci++) {
    for (cj=1; cj<=|C[ci]|; cj++) {
      rule[ci]=cj; // a trying rule sets for test
      count frequency of the trying rule; // count n1 n2
      if (frequency>=N0) { //sufficient frequency ?
        if (|z|>3.0) { //sufficient evidence ?
          store necessary data such as rule, frequency of n1
          and n2, and z
        } // end of if |z|
      rule_check(ci,rule);
      } // end of if frequency
    } // end of for cj
    rule[ci]=0; // trying rules reset
  } // end of for ci
} // end of rule_check
```

Fig. 2. An algorithm for STRIM (Statistical Test Rule Induction Method)

$n_d = \max(n_1, n_2, ..., n_{M_D} = n_6)$, $(d \in \{1, 2, ..., M_{D=6}\})$, $p_d = P(D = d)$, $n = \sum_{m=1}^{M_D} n_m$. In principle, $(n_1, n_2, ..., n_{M_D})$ under $H0$ obeys a multinomial distribution which is sufficiently approximated by the standard normal distribution by use of n_d under the condition[11]: $p_d n \geq 5$ and $n(1 - p_d) \geq 5$. In the same way, the fifth row 110000 denotes $CP(k = 5) = (C(1) = 1 \wedge C(2) = 1)$ $(RL = 2)$, the 13-th row 110010 denotes $C(1) = 1 \wedge C(2) = 1 \wedge C(5) = 1$ $(RL = 3)$, and so on. Here, if we specify a standard of the significance level such as $z \geq z_\alpha = 3.0$ and reject $H0$, then the the assumed $CP(k)$ becomes a candidate for the rules in the rule box. For example, see $CP(1)$ having $z = 12.52 \geq z_\alpha = 3.0$ in Table 3 and confirm the partiality of f that n_1 is much greater than n_l $(l = 2, ..., 6)$.

The second stage process is that of arranging the set of rule candidates derived from the first process, and finally estimating the rules in the rule box, since some candidates may satisfy the relationship: $CP(ki) \supseteq CP(kj) \supseteq CP(kl)$ ..., for example, in the case 100000 $\supset$ 110000 $\supset$ 110010 (see Table 3). The basic notion is to represent the $CP(k)$ of the maximum z, that is, the maximum partiality. In the above example, STRIM selects the $CP(k)$ of 110000, which by chance coincides with the rule specified in advance. Figure 2 shows the STRIM algorithm[8].

Table 4 shows the estimated results for Table 1 with $N = 10000$. STRIM inducts all of twelve rules specified in advance, and also one extra rule. However, there are clear differences between them in the indexes of accuracy and coverage.

Table 4. Results of estimated rules for the decision table in Table 1 by STRIM

estimated rule $R(i)$	$C(1)$	$C(2)$	$C(3)$	$C(4)$	$C(5)$	$C(6)$	D	$f=(n_1,...,n_6)$	p-value (z)	accuracy	coverage
1	5	5	0	0	0	0	5	(7,8,5,7,271,4)	0(34.15)	0.897	0.162
2	0	0	1	1	0	0	1	(243,6,5,6,4,3)	0(32.68)	0.910	0.148
3	4	4	0	0	0	0	4	(10,2,8,252,7,6)	0(32.58)	0.884	0.150
4	0	0	5	5	0	0	5	(5,5,6,11,249,7)	0(32.27)	0.880	0.149
5	6	6	0	0	0	0	6	(10,12,4,7,6,253)	0(32.16)	0.866	0.154
6	3	3	0	0	0	0	3	(6,3,254,13,8,12)	0(31.00)	0.858	0.150
7	0	0	2	2	0	0	2	(4,243,2,8,5,14)	0(31.90)	0.880	0.146
8	0	0	3	3	0	0	3	(11,8,243,5,7,7)	0(31.48)	0.865	0.143
9	0	0	6	6	0	0	6	(7,2,8,10,9,240)	0(31.41)	0.870	0.146
10	0	0	4	4	0	0	4	(8,12,13,245,7,7)	0(30.91)	0.839	0.146
11	1	1	0	0	0	0	1	(235,10,12,2,7,8)	0(30.77)	0.858	0.143
12	2	2	0	0	0	0	2	(8,227,6,11,4,6)	0(30.47)	0.866	0.136
13	0	0	0	0	1	1	2	(39,61,44,44,35,31)	6.26e-4(3.23)	0.240	0.037

4 Studies of the Conventional Methods and Their Problems

The most basic strategy to induct the rules from a decision table is to use the inclusion relationship between the set derived by the condition attributes and the set by the decision attribute. Many methods of achieving this have been proposed [4–7]. Figure 3, for example, shows the well-known LEM2 algorithm[4]. In this B at LN (Line No.) $= 0$ is specified like $B = U(d) = \{u(i)|u^{D=d}(i)\}$ removing the conflicted data set. LEM2 with lower approximation derives $CP(k)$, satisfying $U(d) \supseteq U(CP(k))$. In the figure, t corresponds to $C(j_k) = v_{C(j_k)}$, $[t]$ to $U(t) = \{u(i)|u^{C=t}(i)\}$, T to $CP(k)$, and the final result τ to $\bigvee_k CP(k)$ is obtained by repeating from $LN = 3$ to $LN = 16$ until the condition $U(d) = U(\bigvee_k CP(k))$ is satisfied.

However, as previously shown [8], LEM2 is likely to induct many sub-rules of their true rules with longer rule length, since it executes the algorithm until the condition $U(d) = U(\bigvee_k CP(k))$ is satisfied. In 1993 Ziarko [5] introduced the variable precision rough set, which inducts $CP(k)$ satisfying the following conditions:

$$C_\varepsilon(U(d)) = \{u(i)|acc \geq acc0, \\ acc = |U(d) \cap U(CP(k)|/|U(CP(k)| = n_d/n\}, \quad (2)$$

where acc is accuracy of the rule and acc0 is a constant depending on ε. Ziarko further defined (2) in two cases as follows:

$$\underline{C}_\varepsilon(U(d)) = \{u(i)|acc \geq 1-\varepsilon\}, \quad (2a)$$
$$\overline{C}_\varepsilon(U(d)) = \{u(i)|acc \geq \varepsilon\}, \quad (2b)$$

where $\varepsilon \in [0, 0.5)$ is an admissible classification error. $\underline{C}_\varepsilon(U(d))$ and $\overline{C}_\varepsilon(U(d))$ are respectively called a ε-lower and ε-upper approximation of VPRS, and coincides with the ordinary lower and upper approximation by $\varepsilon = 0$. Their difference

```
Line Procedure LEM2
 No.
  0 (input: a set B
    output: a single local covering τ of set B);
  1 begin
  2   G := B;
      τ := φ;
  3   while G ≠ φ
  4   begin
  5     T := φ;
        T(G) := {t|[t] ∩ G ≠ φ};
  6     while T = φ or not([T] ⊆ B)
  7     begin
  8       select a pair t ∈ T(G)
          such that |[t] ∩ G| is maximum;
          if a tie occurs,
          select a pair t ∈ T(G)
          with the smallest cardinality of [t];
          if another tie occurs,
          select first pair;
  9       T := T ∪ {t};
 10       G := [t] ∩ G;
 11       T(G) := {t|[t] ∩ G ≠ φ};
 12       T(G) := T(G) − T;
 13     end {while}
 14     for each t ∈ T do
          if [T − {t}] ⊆ B then T := T − {t};
 15     τ := τ ∪ {T};
        G := B − ∪T∈τ [T];
 16   end {while};
 17   for each T ∈ τ do
        if ∪S∈τ−{T} [S] = B then τ := τ − {T};
 18 end {procedure}.
```

Fig. 3. An algorithm for LEM2

is in the range of their accuracy; that is the accuracy of $\underline{C}_\varepsilon(U(d)) \in (0.5, 1.0]$ and that of $\overline{C}_\varepsilon(U(d)) \in (0.0, 0.5]$. VPRS adopts the rules with the high index of coverage defined by $cov = |U(d) \cap U(CP(k))|$ / $|U(d)|$; this can squeeze the above sub-rules. VPRS has been widely used for solving real-world problems, and a variety of modified VPRSs have been proposed [12–14]. However, the standard of adopting rules is not so clear. For example, (acc, cov) of $CP(k = 1)$, $CP(k = 5)$ and $CP(13)$ in Table 3 are $(0.281, 0.285)$, $(0.857, 0.143)$ and $(0.864, 0.0231)$ respectively. $CP(k = 13)$ should be adopted as the most accurate, $CP(k = 1)$ as the widest coverage, and $CP(k = 5)$ as the moderate index of both; this requires a decision by the analyst studying the matter. This unclearness and lack of the standard lead the making of an algorithm such as LEM2 to difficulty.

Jaworski [10] further pointed out the problem in VPRS that the standard of adopting rules by using (acc, cov) is highly dependent on the decision table; that is the sample set, as (acc, cov) will change in each sample set. He then extended the decision table in the sample set to that in the population, and proposed a type of confidence interval for each index. For example, for the index of accuracy,

$$P(acc|population) \geq acc|_{sample} - \sqrt{\frac{\ln(1-\gamma_n)}{-2n}}, \tag{3}$$

where, γ_n is a degree of confidence. In $R(i=1)$ of Table 4, $acc|_{sample} = 0.897$, $n = 302$ and let specify $\gamma_n = 0.85$, then $P(acc|population) \geq 0.897 - 0.037 = 0.860$. This means that the accuracy in the population is greater than 0.860, with the degree of confidence of 0.85. However, two-story uncertainty $acc \in [0.0, 1.0]^{[0.0,1.0]}$ is very complicated, and hard to understand.

5 Studies of the Relationship between VPRS and STRIM

Let us consider a $CP(k)$ satisfying (2a). The greater part of the decision attribute value of the $U(CP(k))$ is now included in $U(d)$, since the $CP(k)$ satisfies (2a). Accordingly, the distribution of the decision attribute value of the $U(CP(k))$ has partiality in $D = d$, which coincides with the basic concept of STRIM. As shown in 3, STRIM statistically tests whether $(n_1, n_2, ..., n_{M_D})$ is partial or not, and gives the decision with a significance level z_α. In (2) n_d satisfies the event $n_d \geq n \cdot acc0$, and the probability of the event is evaluated thus:

$$P(n_d \geq n - n \cdot \varepsilon) = P\left(\frac{n_d + 0.5 - np_d}{(np_d(1-p_d))^{0.5}} \geq \frac{n + 0.5 - n \cdot \varepsilon - np_d}{(np_d(1-p_d))^{0.5}}\right). \quad (4)$$

Accordingly, $z_\alpha = \dfrac{n + 0.5 - n \cdot \varepsilon - np_d}{(np_d(1-p_d))^{0.5}}$ is obtained comparing (4) with (1) and then the following relationship between $acc0$ and z_α holds:

$$acc0 = 1 - p_d + 0.5/n - z_\alpha \left(\frac{p_d}{n}(1-p_d)\right)^{0.5} \quad (5)$$

In $R(i=1)$ in Table 4, substituting $n = 302$, $p_d = 1/6$ for (5), $acc0 = 0.229 \in (0.0, 0.5]$ and then $\varepsilon = 0.229$. Let specify $z_\alpha = 30.0$ since $z = 34.5$ in $R(i=1)$, then $acc0 = 0.808 \in (0.5, 1.0]$ and $\varepsilon = 1 - acc0 = 0.192$.

The covering index $cov = n_d/|U(d)|$ in VPRS is considered to reflect the degree of support and/or sufficient evidence of the inducted rule. However, the standard of the degree of support has not yet been clearly expressed. On the other hand, STRIM requested the number of data as the evidence satisfying the conditions [11]: $p_d n \geq 5$ and $n(1-p_d) \geq 5$, which are needed for testing $H0$. Accordingly, the covering index corresponds to the testing condition in STRIM and the condition is clearly given by statistics, whether the covering index is somewhat higher or lower (See Table 4).

As the relationships considered above between VPRS and STRIM, STRIM can yield the validity of inducting rules; that is, the clear meaning and standard of the index of accuracy and coverage for the inducted rules from statistical viewpoints.

6 Studies of Arrangement of Rule Candidates

There may be rule candidates satisfying relationships $CP(ki) \supset CP(kj) \supset CP(kl)$ ... after the first stage process. STRIM selects the candidate with the

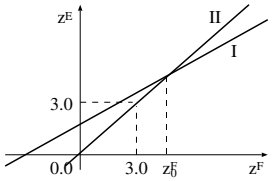

Fig. 4. Relationship between z^F and z^E

maximum partiality in the relationship. Let us consider whether STRIM is assured of selecting the true rule in the rule box. Hereafter, $CP(F)$ denotes the true rule in the rule box, and we assume that it satisfies the relationship $CP(E) \supset CP(F) \supset CP(G)$ and has the maximum n_d at $d = 1$ like $CP(E) =$ "100000", $CP(F) =$ "110000" and $CP(G) =$ "110010" in Table 3 and the distribution of $f = (n_1^F, n_2^F, ..., n_{M_D}^F)$. Then: $z^F = \dfrac{n_1^F + 0.5 - n^F p_1}{(n^F p_1(1-p_1))^{0.5}} \simeq \dfrac{n^F(a^F - p_1)}{\sigma^F}$. Here, 0.5 $<< n^F p_1$, $n_1^F = a^F n^F$ $(0 < a^F \leq 1)$, $n^F = \sum_{j=1}^{M_D} n_j^F$ and $\sigma^F = (n^F p_1(1-p_1))^{0.5}$. In the same way with regard to G, $z^G = \dfrac{n_1^G + 0.5 - n^G p_1}{(n^G p_1(1-p_1))^{0.5}} \simeq \dfrac{n^G(a^G - p_1)}{\sigma^G}$. Here $U(F) \supset U(G)$, $a^G \simeq a^F$ and $n^F > n^G$ lead to $n^G = rn^F$ $(0 < r < 1)$ Accordingly, $z^G \simeq \dfrac{n^G(a^G - p_1)}{\sigma^G} = \dfrac{r^{0.5} n^F(a^F - p_1)}{\sigma^F} = r^{0.5} z^F$, which means $z^G < z^F$ and STRIM selects not $CP(G)$ but $CP(F)$. In the same way with regard to E, $z^E = \dfrac{n_1^E + 0.5 - n^E p_1}{(n^E p_1(1-p_1))^{0.5}}$. Here, $n_1^E = n_1^F + n_1^{EF} = a^E n^E + a^{EF} n^{EF}$, $n^{EF} = |U(EF)|$ and $U(EF) = U(E) - U(F)$. $U(EF)$ is an indifferent data set (See Hypothesis 2). Taking their relationships into consideration, the following equation holds: $z^E \simeq \dfrac{n^F(a^F - p_1)}{\sigma^E} + \dfrac{n^{EF}(a^{EF} - p_1)}{\sigma^E} = \dfrac{\sigma^F}{\sigma^E} z^F + \dfrac{\sigma^{EF}}{\sigma^E} z^{EF} = s_E^F z^F + s_E^{EF} z^{EF}$. Here $n^F < n^E$ and $n^{EF} < n^E$ lead to $\dfrac{\sigma^F}{\sigma^E} = s_E^F < 1$ and $\dfrac{\sigma^{EF}}{\sigma^E} = s_E^{EF} < 1$. Figure 4 shows the relationship between z^F and z^E by use of the following two lines: $z^E = s_E^F z^F + s_E^{EF} z^{EF}$ (I), $z^E = z^F$ (II). The cross point of the two lines is $\left(z_0^F = \dfrac{s_E^{EF}}{1 - s_E^F} z^{EF}, z_0^F \right)$. Accordingly, STRIM always selects $CP(F)$ if the inequality $z^F \geq z_0^F$ holds. As conclusion in this section, STRIM necessarily selects $CP(F)$ of the true rule only if $z^F \geq z_0^F$ holds. In Table 1, $s_E^F = \left(\dfrac{1}{M_C}\right)^{0.5}$, $s_E^{EF} = \left(1 - \dfrac{1}{M_C}\right)^{0.5}$ and $M_C = 6$, and $|z^{EF}| < z_\alpha = 3.0$ holds with less than 1 [%] error. Accordingly, if $z^F \geq z_0^F = \dfrac{s_E^{EF}}{1 - s_E^F} \simeq 5$ holds, then STRIM

selects not $CP(E)$ but $CP(F)$. We can confirm the validity of the consideration of this section in Table 3 and 4. Especially, note that $R(i = 13)$ in Table 4 does not satisfy the condition $z \geq 5.0$, and doubt of the inducted result thus arises.

7 Conclusions

This paper summarized the basic concept of the rule induction method by STRIM [8, 9], and the conventional methods, especially VPRS [5] and their problems, using a simulation experiment. We illustrated the following features and relationships between results from the STRIM model and those from conventional methods, especially VPRS[5]:

1) VPRS uses the indexes of accuracy and coverage with an admissible error when it selects the rule candidates. The accuracy can be recognized as the index of the partiality of the distribution of the decision attribute values for the trying rule, which coincides with the idea of STRIM. The corresponds to the significance level by z_α in STRIM. The coverage corresponds to the applicable condition for a statistical test by STRIM. However, VPRS does not have the objective standard of both indexes to select rule candidates, since to date the conventional methods do not view the decision table as a sample data set obtained from its population.
2) STRIM provides assurance for an analyst searching for the true rules under the proper conditions studied in 6, as the results show whether those rules inducted are true or not. In contrast, VPRS provides no such assurance, since it is not a method based on a data generating model, as is STRIM.

Focus for future studies:

1) To consider relationship to Variable Consistency Rough Sets Approaches (VC-IRSA and VC-DRSA) [15].
2) To see how good are rules found by STRIM in a accuracy cross-validation experiment comparing them to the ones found by LEM2 [4] and/or other classifiers.
3) To consider relationship to rule quality measures which seek to find a trade-off between the rule precision (rule accuracy) and coverage [16].

References

1. Pawlak, Z.: Rough sets. Internat. J. Inform. Comput. Sci. 11(5), 341–356 (1982)
2. Skowron, A., Rauszer, C.M.: The Discernibility Matrix and Functions in Information Systems. In: Slowinski, R. (ed.) Intelligent Decision Support, Handbook of Application and Advances of Rough Set Theory, pp. 331–362. Kluwer Academic Publishers (1992)
3. Bao, Y.G., Du, X.Y., Deng, M.G., Ishii, N.: An Efficient Method for Computing All Reducts. Transaction of the Japanese Society for Artificial Intelligence 19(3), 166–173 (2004)

4. Grzymala-Busse, J.W.: LERS- A system for learning from examples based on rough sets. In: Słowński, R. (ed.) Intelligent Decision Support. Handbook of Applications and Advances of the Rough Sets Theory, pp. 3–18. Kluwer Academic Publishers (1992)
5. Ziarko, W.: Variable precision rough set model. Journal of Computer and System Science 46, 39–59 (1993)
6. Shan, N., Ziarko, W.: Data-based acquisition and incremental modification of classification rules. Computational Intelligence 11(2), 357–370 (1995)
7. Nishimura, T., Kato, Y., Saeki, T.: Studies on an Effective Algorithm to Reduce the Decision Matrix. In: Kuznetsov, S.O., Ślęzak, D., Hepting, D.H., Mirkin, B.G. (eds.) RSFDGrC 2011. LNCS (LNAI), vol. 6743, pp. 240–243. Springer, Heidelberg (2011)
8. Matsubayashi, T., Kato, Y., Saeki, T.: A new rule induction method from a decision table using a statistical test. In: Li, T., Nguyen, H.S., Wang, G., Grzymala-Busse, J., Janicki, R., Hassanien, A.E., Yu, H. (eds.) RSKT 2012. LNCS (LNAI), vol. 7414, pp. 81–90. Springer, Heidelberg (2012)
9. Kato, Y., Saeki, T., Mizuno, S.: Studies on the Necessary Data Size for Rule Induction by STRIM. In: Lingras, P., Wolski, M., Cornelis, C., Mitra, S., Wasilewski, P. (eds.) RSKT 2013. LNCS (LNAI), vol. 8171, pp. 213–220. Springer, Heidelberg (2013)
10. Jaworski, W.: Rule Induction: Combining Rough Set and Statistical Approaches. In: Chan, C.-C., Grzymala-Busse, J.W., Ziarko, W.P. (eds.) RSCTC 2008. LNCS (LNAI), vol. 5306, pp. 170–180. Springer, Heidelberg (2008)
11. Walpole, R.E., Myers, R.H., Myers, S.L., Ye, K.: Probability & Statistics for Engineers & Scientists, 8th edn., pp. 187–194. Pearson Prentice Hall (2007)
12. Xiw, G., Zhang, J., Lai, K.K., Yu, L.: Variable precision rough set group decision-making, An application. International Journal of Approximate Reasoning 49, 331–343 (2008)
13. Inuiguchi, M., Yoshioka, Y., Kusunoki, Y.: Variable-precision dominance-based rough set approach and attribute reduction. International Journal of Approximate Reasoning 50, 1199–1214 (2009)
14. Huang, K.Y., Chang, T.-H., Chang, T.-C.: Determination of the threshold β of variable precision rough set by fuzzy algorithms. International Journal of Approximate Reasoning 52, 1056–1072 (2011)
15. Greco, S., Matarazzo, B., Słowiński, R., Stefanowski, J.: Variable Consistency Model of Dominance-Based Rough Sets Approach. In: Ziarko, W., Yao, Y. (eds.) RSCTC 2000. LNCS (LNAI), vol. 2005, pp. 170–181. Springer, Heidelberg (2001)
16. Janssen, F., Fürnkranz, J.: On the quest for optimal rule learning heuristics. Machine Learning 78, 343–379 (2010)