Considerations on Rule Induction Methods by the Conventional Rough Set Theory from a View of STRIM

Tetsuro Saeki[1(✉)], Jiwei Fei[1], and Yuichi Kato[2]

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

Abstract. In this paper, the rule induction method STRIM, the classical Rough Sets (RS) theory and the notion of three-way decision rules are summarized and their performance is examined by applying them to a real-world dataset and a simulation dataset. From these experimental studies, the problems inherent in the rule induction method by the conventional RS theory based on the indiscernibility are pointed out and a comparison is made with STRIM. Specifically, the rule induction methods that are based on indiscernibility and do not consider the decision table which is only a sample of outcomes obtained by chance from a population of interest are highly dependent upon the samples in the decision table given. This paper states that such rule induction methods are thus problematic and need to be improved to create a more robust rule induction method.

1 Introduction

Extracting the properties and structures hidden in a large dataset is about discovering knowledge and/or information, and that is important for making good strategical decisions and acting consistently. For example, Rough Sets (RS) theory proposed by Pawlak [1] in 1982 is used for reducting a dataset, creating a decision table [2,3], and inducing if-then rules hidden in the decision table [4,5]. Here, the dataset is a set of objects each of which is featured by particular values: its condition attributes and its decision attribute. RS theory first focuses on an indiscernibility property of these objects and provides inclusion relationships of the target object set by defining lower and upper approximations. These approximate expressions provide two representative rules with necessity (accuracy = 1.0) and possibility (accuracy > 0.0) respectively. However, the necessity rule imposes a severe condition, i.e., accuracy = 1.0, on the rule induction. Therefore, Ziarko [6] proposed a variable precision rough set model (accuracy = $1.0 - \varepsilon$) with an admissible error ($\varepsilon \in [0.0, 0.5)$).

Yao [7–9] divided the target set into positive, negative, and boundary regions using the lower and upper approximations and proposed three-way decision rules

H. S. Nguyen et al. (Eds.): IJCRS 2018, LNAI 11103, pp. 202–214, 2018.
https://doi.org/10.1007/978-3-319-99368-3_16

corresponding to those regions. Yao also suggested that the boundary parameters (α, β) of the three-way decision rules should be determined by considering accuracy as a type of conditional probability representation and introducing a cost function from a Bayesian decision perspective. This consideration extends Pawlak's and Ziarko's rule induction methods and corresponds to them in some special cases. However, Yao does not propose a new reduction method or a new rule induction method for the decision table and the new related algorithms.

As an alternative to RS theory, the statistical test rule induction method (STRIM) which considers the decision table as a sample dataset obtained from a population has been proposed [10–17]. STRIM uses a statistical reduct method on the decision table [14] and a statistical rule induction method from the reducted table [16]. Note that STRIM was studied independently of the conventional RS methods and was not based on the approximation concept. Specifically, STRIM recognizes the condition attributes and decision attributes of the decision table as random variables and the decision table as their outcomes. Moreover STRIM proposes a data generation model of the decision table by a system which generates input sets of condition attribute values and transforms them into the corresponding output of the decision attribute value through prespecified if-then rules and hypotheses with regard to the decision attribute value based on causality. This system can also be used for confirming the validity of any rule induction method by applying the method to the dataset generated by the system and investigating whether the method can or cannot induce the pre-specified rules.

In this paper, we first summarize STRIM and give an example of testing its performance by applying it to a real-world dataset. We then state the basics of the if-then rule induction method by STRIM from the viewpoint of proof by contradiction in propositional logic. We then summarize the conventional RS theory based on indiscernibility, and point up the problem of its rule induction method based on indiscernibility in contrast to STRIM. We study this experimentally by applying the LEM2 algorithm, implementing the classical RS theory to the data generation model described above and comparing the results with those of the same experiment using STRIM. Lastly, the idea of three-way decision rules is summarized and we point out that the idea is fundamentally based on the concept of indiscernibility and will cause the same problems as does the classical RS theory. From three summarizations and studies of the conventional methods, this paper points out that the rule induction method based on the concept of indiscernibility of the given decision table needs to be improved as the decision table is merely a sample obtained from the population.

2 The Conventional STRIM

In RS theory, the decision table is expressed as: $S = (U, A = C \cup \{D\}, V, \rho)$. 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 $C(j)$, a condition attribute, is a member of C, and D is a decision attribute. V is a set of attribute values denoted $V = \bigcup_{a \in A} V_a$ and characterized by the information function ρ: $U \times A \to V$.

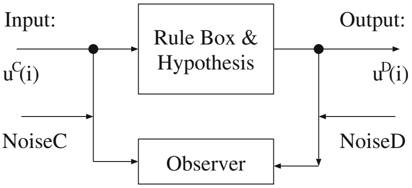

Fig. 1. Data generation model: The rule box contains if-then rules $R(d,k)$: if $CP(d,k)$ then $D = d$ $(d = 1, 2, ..., k = 1, 2, ...)$.

Table 1. Hypotheses with regard to 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)

Generally, inducing if-then rules from a decision table implicitly assumes a causal relationship between the condition attributes and decision attributes. Therefore, in STRIM, we propose a model in which S is derived from the input/output relationships shown in Fig. 1. In other words, STRIM considers the decision table to be a sample dataset obtained from an input–output system that includes a rule box as shown in Fig. 1 and hypotheses regarding the decision attribute values, as shown in Table 1. A sample $u(i)$ consists of its condition attribute values $u^C(i)$ and decision attribute values $u^D(i)$. Here, $u^C(i)$ is an input to the rule box and is transformed to the output $u^D(i)$ using the rules (generally unknown) contained in the rule box and the hypotheses. The hypotheses consist of three cases corresponding to the nature of the input. The three cases are: uniquely determined, indifferent, and conflicted (see Table 1). In contrast, $u(i) = (u^C(i), u^D(i))$ is measured by an observer (Fig. 1). The existence of NoiseC and NoiseD causes missing values in $u^C(i)$ and changes $u^D(i)$ to create another $u^D(i)$ value. These noises bring the system closer to a real-world system. Differing from the conventional RS theory, STRIM includes the data generation model shown in Fig. 1. This data generation model suggests that the values $(u^C(i), u^D(i))$, i.e., a decision table is the outcome of the random variables $(C, D) = ((C(1), ..., C(|C|), D)$ observing the population. Therefore, in STRIM, $\rho(u(i), C(j))$ are the outcome of the random variables $C(j)$. Note that there is no concept of the information function in STRIM, i.e., $S = (U, A = C \cup \{D\}, V)$ is the decision table and V is the sample space in STRIM.

Table 2. STRIM rule induction results for Rakuten Travel dataset.

$CP(d,k)$	$C(1)C(2)$ ...$C(6)$	D	p-value (z)	Accuracy	Coverage	$f = (n_1, n_2, n_3, n_4, n_5)$
(5,1)	005050	5	0.0 (64.08)	0.876	0.629	(11, 12, 9, 146, 1258)
(5,2)	005005	5	0.0 (58.31)	0.915	0.486	(17, 6, 5, 62, 972)
(1,1)	000010	1	0.0 (57.78)	0.766	0.639	(1277, 346, 40, 4, 1)
(4,1)	040040	4	0.0 (40.37)	0.719	0.348	(16, 37, 90, 695, 129)
(3,1)	030030	3	0.0 (38.12)	0.633	0.392	(73, 203, 784, 170, 9)
(2,1)	020000	2	3.0E−168 (27.62)	0.494	0.348	(303, 695, 351, 51, 6)

Given a dataset created by the data generation model in Fig. 1, five processes are carried out: (1) STRIM extracts significant pairs of condition attributes and their values, e.g., $C(j) = v_{j_k}$, for rules of $D = d$ using the local reduct [14,16,17]; (2) STRIM constructs a trying condition part of the rules, e.g., $CP(d,k) = \wedge_j(C(j_k) = v_j)$, using the reduct results; (3) STRIM investigates whether $U(CP(d,k))$ has caused a bias at n_d in the frequency distribution of the decision attribute values $f = (n_1, n_2, ..., n_{M_D})$. Here, $n_m = |U(CP(d,k)) \cap U(m)|$ $(m = 1, ..., |V_D| = M_D)$, $U(CP(d,k)) = \{u(i)|u^{C=CP(d,k)}(i)$, i.e., $u^C(i)$ sastifies $CP(d,k)\}$, and $U(m) = \{u(i)|u^{D=m}(i)\}$ since the $u^C(i)$ coinciding with $CP(d,k)$ in the rule box is transformed to $u^D(i)$ based on hypothesis 1 or 3 (Table 1). In other words, $CP(d,k)$ coinciding with one of the rules in the rule box creates bias in $f = (n_1, n_2, ..., n_{M_D})$. Specifically, STRIM uses a statistical test method for the investigation of the bias specifying a null hypothesis $H0$: f does not have any bias, i.e., $CP(d,k)$ is not a rule; the alternative hypothesis is $H1$: f has a bias, i.e., $CP(d,k)$ is a rule and has a proper significance level. Here, $H0$ is tested using the sample dataset, i.e., the decision table and the proper test statistics; for example,

$$z = \frac{(n_d + 0.5 - np_d)}{(np_d(1-p_d))^{0.5}} \quad (d = 1, 2, ..., M_D), \tag{1}$$

where $p_d = P(D = d)$, $n = \sum_{j=1}^{5} n_j$, z obeys the standard normal distribution under a proper condition [18] and is considered an index of the bias of f; (4) If $H0$ is rejected, the assumed $CP(d,k)$ becomes a candidate for the rules in the rule box; (5) STRIM repeats processes (1–4) to obtain a set of rule candidates, then arranges the rule candidates and induces the final results [16,17].

Figure 2 shows a STRIM algorithm that includes a reduct function. Here, line nos. (LN) 8 and 9 are the reduct part of process (1), process (2) is executed at LN 10, where the dimension rule[] is used as the rule candidate, process (3) is executed at LN 25 in the rule_check() function, process (4) is executed at LN 26, and process (5) is executed from LN 7 to LN 11 and LN 12.

A rule induction example obtained by applying STRIM to the Rakuten Travel dataset, which is maintained by the Rakuten Institute of Technology follows

```
Algorithm to induce if-then rules by STRIM with a reduct function
int main(void) {
int rdct_max[|CV|]={0,. . . ,0}; //initialize maximum value of C(j)
int rdct[|CV|]={0,. . . ,0}; //initialize reduct results by D=l
int rule[|C|]={0,...,0}; //initialize trying rules
int tail=-1; //initialize value set
input data; // set decision table
for (di=1; di<=|D|; di++) {// induce rule candidates every D=l
attribute_reduct(rdct_max)
set rdct[ck] ; // if (rdct_max[ck]==0) {rdct[ck]=0; }else {rdct[ck]=1; }
rule_check(rcdct, redct_max, tail, rule); // the first stage process
}// end di
arrange rule candidates // the second stage
}// end main
int attribute_reduct(int rdct_max[]) {
make contingency table for D=l vs. C(j)
Test H0(j,l);
if H0(j,l) is rejected then set rdct_max[j,l]=jmax else rdct_max[j,l]=0; // jmax:the attribute value of the maximum frequency
}// end of attribute_reduct
int rule_check(int rdct[], int rdct_max[], int tail,int rule[]) {// the first stage process
for (ci=tail+1; cj<|C|; ci++) {
for (cj=1; cj<=rdct[ci]; cj++) {
rule[ci]=rdct_max[cj]; // a trying rule set for test
count frequency of the trying rule; // count n1, n2, ...
if (frequency>=N0) {//sufficient frequency ?
if (|z|>3.0) {//sufficient evidence ?
add the trying rule as a rule candidate
}// end of if |z|
rule_check(ci,rule)
}// end if frequency
}// end cj
rule[ci]=0; // trying rules reset
}// end ci
}// end rule_check
```

Fig. 2. STRIM algorithm with reduct function.

[17] (for another example, see [16]). The dataset concerned contains approximately $6,200,000$ questionnaire surveys of ratings $A = \{\ C(1) =$ "Location," $C(2) =$ "Room," $C(3) =$ "Meal," $C(4) =$ "Bath (Hot Spring)," $C(5) =$ "Service," $C(6) =$ "Amenity," and $D =$ "Overall" $\}$ of approximately $130,000$ travel facilities by using a set of categorical values $V_a = \{$ "Dissatisfied ($DS(1)$)," "Somewhat dissatisfied ($SD(2)$)," "Neither satisfied nor dissatisfied ($NN(3)$)," "Satisfied ($ST(4)$)," and "Very Satisfied ($VS(5)$)" $\}$, where $\forall a \in A$, i.e., $|V_{a=D}| = |M_D| = |V_{a=C(j)}| = M_{C(j)} = 5$. We constructed a decision table of $N = 10,000$ questionnaire surveys by randomly selecting $2,000$ samples, each of $D = m$ $(m = 1, ..., 5)$, from approximately $400,000$ surveys from the 2013–2014 dataset, choosing these surveys because they contained heavy biases with respect to the frequency of $D = m$. We applied STRIM to this decision table and obtained Table 2, which represents the following:

(1) $CP(d = 5, k = 1)$ represents a rule stating that if $(C(3) = VS(5)) \bigwedge (C(5) = VS(5))$ then $D = VS(5)$, and its accuracy and coverage are 0.876 and 0.639, respectively.

Table 3. Examples of rules induced by LEM2 for the first simulation dataset (Case1).

Rule no.	Rule	(accuracy, coverage)	$f = (n_1, n_2, n_3, n_4, n_5, n_6)$
1	(C1 = 1) & (C2 = 1) & (C4 = 6) = >(D = 1)	(1.0, 0.0296)	(15,0,0,0,0,0)
2	(C1 = 4) & (C3 = 1) & (C4 = 1) = >(D = 1)	(1.0, 0.0355)	(18,0,0,0,0,0)
3	(C1 = 1) & (C2 = 1) & (C3 = 1) = >(D = 1)	(1.0, 0.0197)	(10,0,0,0,0,0)
4	(C1 = 5) & (C2 = 6) & (C3 = 1) & (C4 = 1) = >(D = 1)	(1.0, 0.0138)	(7,0,0,0,0,0)
...	...	...	...
8	(C1 = 1) & (C2 = 1) & (C3 = 5) & (C5 = 6) = >(D = 1)	(1.0, 0.0099)	(5,0,0,0,0,0)
...	...	...	...
24	(C1 = 5) & (C2 = 6) & (C3 = 5) & (C4 = 4) & (C5 = 2) = >(D = 1)	(1.0, 0.002)	(1,0,0,0,0,0)
...	...	...	...
27	(C1 = 2) & (C3 = 2) & (C4 = 5) & (C5 = 6) & (C6 = 5) = >(D = 1)	(1.0, 0.002)	(1,0,0,0,0,0)
...	...	...	...

(2) This rule implies the frequency $f = (11, 12, 9, 146, 1258)$ of the decision attribute values, and the bias at $D = 5$ is $z = 64.08$ as calculated by Eq. (1) corresponding to the p-value= 0.0.

(3) STRIM suggests that $C(1)$ = "Location" and $C(4)$ = "Bath (Hot Spring)" can be reducted because no rules use those attributes.

3 Considerations on a Rule Induction Method by STRIM from the Viewpoint of Proof by Contradiction

In propositional logic, a logical expression Q is often derived from several logical expressions $P_1, P_2, ..., P_n$. It can be proved that Q is also true (T) from the interpretation that all P_j $(j = 1, ..., n)$ is T. Simultaneously, if $P_1 \wedge P_2 \wedge ... \wedge P_n = P$, $P \rightarrow Q$ is valid. Here, Q is referred to as a logical consequence from P. If $P \rightarrow Q$ is shown to be true, a reasoning result $Q^{'}$ for arbitrary $P^{'}$ can be obtained using reasoning rules by modus ponens. In propositional logic, to demonstrate that $P \rightarrow Q$ is true, the proof by contradiction is often used to indicate that $P \wedge \sim Q =$ false (F) because $P \rightarrow Q = \sim P \vee Q = \sim (P \wedge \sim Q) =$ T.

As described in Sect. 2, rules hidden in the decision table are derived by evaluating the condition part $CP(d,k) = \wedge_j \ (C(j_k) = v_j)$ of the if-then rule for $D = d$ by a hypothesis test. We propose an algorithm to estimate rule candidates by rejecting $H0$: f does not have any bias and $CP(d,k)$ is not a rule. Now, let $P_j = \mathrm{T}$ when $C(j_k) = v_k$ and let $P_j = \mathrm{F}$ when $C(j_k) \neq v_k$. In addition, let $Q = \mathrm{T}$ when $D = d$ and $Q = \mathrm{F}$ when $D \neq d$. For example, in $CP(d = 5, k = 1)$ in Table 2, the number of samples of U where $P = \mathrm{T}$ is 11 + 12 + 9 + 146 + 1,258 = 1,436, and among them the number of samples where $D \neq 5$ ($Q = \mathrm{F}$, i.e., $\backsim Q = \mathrm{T}$) is 11 + 12 + 9 + 146 = 178. Therefore, under $H0$, the number of samples for $P \wedge \sim Q = \mathrm{T}$ is 178. Note that $(C, D) = ((C(1), ..., C(|C|)), D)$ are random variables. Under $P(D = 5) = 1/5$ and the judgment model in Table 1, the occurrence probability of such a distribution shows that the p-value is equal to or less than 0.0. Thus, $H0$ is rejected in this case, i.e., it is determined statistically that $P \wedge \sim Q = \mathrm{F}$. Therefore, it can be seen that $P \rightarrow Q = \mathrm{T}$ is shown with critical p-value = 0.0. Here, since (C, D) are random variables it is necessary to consider the problem that the if-then rule induction method (Sect. 2) is rooted in the fact that the propositional logic $P \rightarrow Q$ is judged to be statistically true or false using proof by contradiction.

4 Considerations on Conventional RS Theory and Its Application to a Rule Induction Problem

Conventional RS theory focuses on the following equivalence relation and the equivalence set of indiscernibility within the decision table S of interest:

$$I_B = \{(u(i), u(j)) \in U^2 | \rho(u(i), a) = \rho(u(j), a), \forall a \in B \subseteq C\}.$$

Here, I_B is an equivalence relation in U and derives the quotient set, $U/I_B = \{[u_i]_B | i = 1, 2, ..., |U| = N\}$, and $[u_i]_B = \{u(i) \in U | (u(j), u_i) \in I_B, u_i \in U\}$. $[u_i]_B$ is an equivalence set with the representative element u_i. Let it be that $\forall X \subseteq U$, then X can be approximated as $B_*(X) \subseteq X \subseteq B^*(X)$ using the equivalence set:

$$B_*(X) = \{u_i \in U | [u_i]_B \subseteq X\}, \tag{2}$$

$$B^*(X) = \{u_i \in U | [u_i]_B \cap X \neq \phi\}. \tag{3}$$

$B_*(X)$ and $B^*(X)$ are the lower and upper approximations respectively of X by B. Note that the pair $(B_*(X), B^*(X))$ is typically referred to as a rough set of X by B.

Specifically, we let $X = \{u(i) | \rho(u(i), D) = d\} = U(d) = \{u(i) | u^{D=d}(i)\}$, and define a set of $u(i)$ as $U(CP) = \{u(i) | u^{C=CP}(i)\}$. If $U(CP) \subseteq U(d)$, then, with necessity, CP can be used as the condition part of the if-then rule of $D = d$. In other words, the following expression of if-then rules with necessity is obtained:

$$Rule(d, k): \text{ if } CP = \wedge_j (C(j_k) = v_{j_k}) \text{ then } D = d. \tag{4}$$

Similarly, with possibility, $C^*(X)$ derives the condition part CP of the if-then rule of $D = d$. However, the approximations $B_*(X) \subseteq X \subseteq B^*(X)$ of $U(d)$ by lower/upper approximation are too severe or too loose, respectively, and, in many cases, it is impossible to induce effective rules due to the inclusion relationship. Ziarko then expanded the original RS by introducing an admissible error in two ways:

$$\underline{B}_{\epsilon}(U(d)) = \{u(i) | acc \geq 1 - \varepsilon\}, \tag{5}$$

Table 4. Examples of rules induced by STRIM for the first simulation dataset (Case1).

$CP(d,k)$	$C(1)$ $...C(6)$	D	p-value(z)	Accuracy	Coverage	$f = (n_1, n_2, n_3, n_4, n_5, n_6)$
(6,1)	660000	6	5.91E−98(20.97)	0.938	0.1883	(1, 2, 1, 2, 0, 90)
(3,1)	330000	3	1.94E−97(20.92)	0.978	0.1778	(0, 0, 88, 1, 1, 0)
(2,1)	002200	2	2.70E−89(20.00)	0.942	0.1698	(,1 81, 1, 1, 1, 1)
(5,1)	550000	5	1.71E−81(19.08)	0.987	0.1477	(0, 0, 0, 0, 78, 1)
(6,2)	006600	6	2.99E−81(19.05)	0.889	0.1674	(6, 1, 1, 0, 2, 80)
(5,2)	005500	5	9.91E−81(18.99)	0.964	0.1515	(0, 1, 1, 1, 80, 0)
(1,1)	001100	1	2.42E−79(18.82)	0.920	0.1578	(80, 1, 2, 0, 3, 1)
(3,1)	003300	3	8.65E−77(18.50)	0.888	0.1596	(3, 2, 79, 2, 2, 1)
(4,1)	004400	4	1.50E−76(18.48)	0.949	0.1456	(1, 0, 1, 75, 1, 1)
(1,2)	110000	1	4.86E−74(18.17)	0.959	0.1381	(70, 1, 1, 0, 1, 0)
(2,2)	220000	2	9.07E−68(17.35)	0.938	0.1279	(0, 61, 1, 0, 2, 1)
(4,2)	440000	4	1.45E−65(17.06)	0.918	0.1301	(1, 1, 0, 67, 2, 2)
(6,3)	600600	6	6.82E−24(10.01)	0.532	0.1046	(8, 9, 11, 6, 10, 5)
(5,3)	500500	5	7.14E−08(7.08)	0.464	0.0739	(10, 10, 11, 5, 39, 9)
(3,3)	030300	3	2.33E−08(5.46)	0.390	0.0606	(11, 6, 30, 12, 10, 8)

Table 5. Comparison of the number of induced rules by rule length derived by using LEM2 and STRIM.

Case no.	Method	Number of rules by rule length						
		1	2	3	4	5	6	Total
Case1	LEM2	0	0	82	1073	623	0	1778
	STRIM	0	15	0	0	0		15
Case2	LEM2	0	0	72	1108	556	0	1736
	STRIM	0	14	0	0	0	0	14
Case3	LEM2	0	0	74	1106	616	0	1796
	STRIM	0	13	0	0	0	0	13

$$\overline{B}_{\varepsilon}(U(d)) = \{u(i) | acc > \varepsilon\}, \tag{6}$$

where $acc = |U(d) \cap U(CP(k))|/|U(CP(k))| = n_d/n$, $\varepsilon \in [0, 0.5)$. The pair $(\underline{B}_{\epsilon}(U(d)), \overline{B}_{\varepsilon}(U(d)))$ is called an ε-lower and ε-upper approximation that satisfies the properties $B_*(U(d)) \subseteq \underline{B}_{\varepsilon}(U(d)) \subseteq \overline{B}_{\varepsilon}(U(d)) \subseteq B^*(U(d))$, $\underline{B}_{\varepsilon=0}(U(d)) = B_*(U(d))$, and $\overline{B}_{\varepsilon=0}(U(d)) = B^*(U(d))$. The ε-lower and/or ε-upper approximations induce if-then rules with admissible errors in the same manner as the lower and/or upper approximations.

As described above, in conventional RS theory, an equivalence relation I_B at a given U is first focused on. Then, based on this relation, an equivalence set at a given U is derived, and the target set is approximated by the equivalence set. Using these approximated sets, if-then rules are induced respectively, as described above. However, the outcome $\rho(u(i), C(k))$ of the random variable $C(k)$ is used for the equivalence relation $I_B = \{(u(i), u(j)) \in U^2 | \rho(u(i), a) = \rho(u(j), a), \forall a = \forall C(k) \in B \subseteq C\}$. Therefore, the equivalence event I_B is a probability event controlled by the conditional joint probability $P((C(k) = \rho(u(i), C(k)), C(k) = \rho(u(j), C(k))) | \rho(u(i), C(k)) = \rho(u(j), C(k)), \forall C(k) \in B \subseteq C)$.

Here, we confirm the rule induction performance using the conventional RS theory in a simulation experiment. First, we set the following rule in the Rule Box in Fig. 1:

$$R(d): \text{ if } R_d \text{ then } D = d, \quad (d = 1, ..., M_D = 6) \tag{7}$$

$$R_d = (C(1) = d) \wedge (C(2) = d) \vee (C(3) = d) \wedge (C(4) = d).$$

Assume that random variables $C(j)$ $(j = 1, ..., |C| = 6)$ are distributed uniformly and generate inputs $u^C(i) = (v_{C(1)}(i), ..., v_{C(6)}(i))$ $(i = 1, ..., N = 10000)$. Then, using the pre-specified rule (7) and the hypothesis in Table 1, the output $u^D(i)$ is generated to create a decision table. We randomly selected samples by $N_B = 3,000$ from the decision table and formed a new decision table. Table 3 shows some of the 1,778 rules obtained by applying the LEM2 algorithm implementing the lower approximation in ROSE2 [18] to this decision table. In Table 3, by focusing on the rule for $D = 1$ as an example, two or three rules are shown for rule lengths 3 4, and 5. Table 4 shows the results of analyzing the same decision table by STRIM. This simulation experiment was repeated three times, and the numbers of rules induced by each method were arranged and compared according to the rule length in Table 5. We observe the following from these tables.

(1) LEM2 induced all rules for accuracy = 1. Some of the induced rules with rule length 3 or 4 shown in Table 3 are sub-rules of the pre-specified rules. If specifying admissible error ε for accuracy and estimating rules by use of VPRS, it is possible to induce the pre-specified rules shown in Table 4. However, in VPRS neither an induction algorithm nor a specifying method for ε has been proposed.

(2) As shown in Table 4, STRIM induced all 12 pre-specified rules and three extra rules. Statistical evidence (p-value or z-value) is shown in these rules. Although it seems that the pre-specified rules can be estimated using appropriate ε and VPRS, the main component of the induction in STRIM is the statistical test The induced rules are based on evidence, i.e., a sufficient number of data that can be used by the statistical test. On the other hand, the coverages of the rules induced in LEM2 are only small percentages, i.e., they include rules of length 5, and by any criterion that is not sufficiently restrictive to be accepted as a rule.

(3) The decision table can be considered a collection of many unarranged if-then rules. LEM2 and STRIM summarize those rules so that human beings can grasp and use the structure and/or features of the rules. From conducting the rule induction experiment three times by LEM2 and STRIM (Table 5), we see that LEM2 summarizes 3,000 rules in somewhat more than 1,700 rules; however, it is clear that LEM2 cannot adequately deal with the given decision table. On the other hand, STRIM induces all pre-specified rules (generally unknown). Note that STRIM induces several additional rules; however, the difference between STRIM and LEM2 can be clearly observed from the accuracy coverage and z-value (Table 4). The validity of the analyzed result by STRIM for the real-world dataset in Table 2 can be inferred to some extent from this simulation result. In any case, we can infer that the rule induction method by the conventional RS based on stochastically varying equivalence relations derives different rules for each decision table, and that the lower approximation rule based on such an equivalence relation cannot fully summarize the decision table.

5 Three-Way Decision Rules and Their Application to the Classification Problem

Yao proposed the concept of three-way decision rules as a new rule induction and decision-making method based on a new interpretation of the classical RS theory [7–9]. Specifically, using a classical RS, Yao proposed to divide U into three regions of X, i.e., the positive region $POS(X)$, the boundary region $BND(X)$, and the negative region $NEG(X)$:

$$POS(X) = B_*(X), \tag{8}$$

$$BND(X) = B^*(X) - B_*(X), \tag{9}$$

$$NEG(X) = U - POS(X) \cup BND(X) = U - B^*(X) = (B^*(X))^C. \tag{10}$$

Any element $x \in POS(X)$ certainly belongs to X, and any element $x \in NEG(X)$ does not belong to X. One cannot decide with certainty whether or not an element $x \in BND(X)$ belongs to X. Similar to the conventional RS theory, we let $X = U(d)$ and can obtain the following decision rules corresponding to (8), (9), and (10):

$$Des([x]) \rightarrow_P Des(U(d)), \text{ for } [x] \subseteq POS(U(d)), \tag{11}$$

$$Des([x]) \rightarrow_B Des(U(d)), \text{ for } [x] \subseteq BND(U(d)), \tag{12}$$

$$Des([x]) \rightarrow_N Des(U(d)), \text{ for } [x] \subseteq NEG(U(d)). \tag{13}$$

Here, $Des([x])$ denotes the logic formula defining the equivalence class $[x]$. For example, $[x]$ is defined by $\wedge_j\ (C(j_k) = v_{j_k})$.

Yao links (11), (12), and (13) to the rule accuracy (or confidence) based on the probability measure as follows:

$$acc(Des([x]) \rightarrow_{\Lambda} Des(U(d))) = Pr(U(d)|[x]) = \frac{|[x] \cap U(d)|}{|[x]|}. \tag{14}$$

Here, $Pr(U(d)|[x])$ is the conditional probability of $U(d)$ given $[x]$. In other words, the probability that the element of $[x]$ exists in $U(d)$ is estimated by the cardinal number. According to accuracy, the positive, boundary, and negative rules are defined by the conditions: $acc = 1$, $0 < acc < 1$, and $acc = 0$, respectively. However, like the idea of VPRS, such approximation based on acc is impractical because the condition is too severe to handle real-world datasets. Therefore, Yao introduced tolerance, similar to VPRS, and proposed rules for the classification problem as follows:

(P1) If $Pr(U(d)|[x]) \geq \alpha$, decide $[x] \subseteq POS(U(d))$,
(B1) If $\beta < Pr(U(d)|[x]) < \alpha$, decide $[x] \subseteq BND(U(d))$,
(N1) If $Pr(U(d)|[x]) \leq \beta$, decide $[x] \subseteq NEG(U(d))$.

Here, $0 \leq \beta < \alpha \leq 1$. As described above, Yao associated the accuracy of the induced rule with the conditional probability. Furthermore, when applying this induced rule to the classification problem, Yao proposed determining boundary parameters (α, β) in accordance with a criterion that minimizes the costs and/or losses by errors based on Bayesian statistics [19]. A detailed discussion is given in the literature [8].

Ziarko did not report a method to specify a reasonable admissible error ε. Yao specified error ε based on Bayesian statistics and included previous studies as a special case. For example, Eqs. (5) and (6) correspond to $\alpha = 1 - \varepsilon$ and $\beta = \varepsilon$, respectively. However, Yao did not propose a specific rule induction method and/or algorithm, such as the decision matrix method [4] or LEM2 [5]. In addition, the three-way decision rules constructing three regions, i.e., the positive, boundary, and negative regions are based on the equivalence relation, which depends on the given decision table and will induce different rules for each sample dataset obtained from the same population similar to the results in classical RS theory.

6 Conclusion

This paper has summarized the concept and validity of a STRIM algorithm that induces rules without using RS theory but by using a statistical test. Furthermore, the rule induction performance of STRIM has been demonstrated through a real-world dataset analysis and a simulation experiment. STRIM has the following features.

(1) There is a data generation model in which the roles of input, output, input/output converting mechanism, observation, and noise generation are clear.

(2) The condition attributes (input) and the decision attribute (output) are considered random variables. Therefore, for example, $\rho(u(i), C(k))$ in the decision table are the outcomes of the random variables $C(k)$. In other words, the decision table is the set of outcomes randomly obtained from the population with condition attributes and decision attribute.

(3) The if-then rule is an input/output converting mechanism that causes bias in the output distribution under the decision attribute value hypothesis (Table 1).

(4) The judgment of bias in the output distribution is determined by a statistical test using a given decision table. Therefore, although STRIM uses a sample dataset, it has an objective criterion that satisfies the criteria for statistical testing with a significance level.

(5) The statistical test is rooted in the proof by contradiction, which is often used when demonstrating the logical consequences of propositional logic.

We have also summarized the conventional RS theory and the associated rule induction method, and pointed out problems there with shown by the results of the simulation experiment. Corresponding to points (1) to (4) above, the conventional RS theory and the rule inducing method are described as follows.

(i) There is no data generation model. Thus, there is no alternative to studying the given decision table at the starting point.

(ii) As there is no data generation model, such as the information function $\rho(u(i), C(k))$, $\rho(u(i), D)$ is needed for convenience. The information function is such that the function value is different for each sample for the same attribute $C(k)$.

(iii) The criterion for adopting a rule is accuracy, and the adoption criteria are not clear (coverage is very small e.g. only one sample satisfies the rule).

(iv) The induced rules are established using only the given decision table, and different rules are derived from different decision tables obtained from the same population because the equivalence class and lower and upper approximation sets differ for each decision table.

From the above, it is considered that the indiscernibility based on the equivalence class is not the essence of a good rule induction method and an improved rule induction method is needed.

References

1. Pawlak, Z.: Rough sets. Int. J. Inform. Comput. Sci. **11**(5), 341–356 (1982)
2. Skowron, A., Rauser, C.M.: The discernibility matrix and functions in information systems. In: Słowiński, R. (ed.) Intelegent Decision Support, Handbook of Application and Advances of Rough Set Theory, pp. 331–362. Kluwer Academic Publishers, Boston (1992)
3. Thangavel, K., Pethalakshmi, A.: Dimensional reduction based on rough set theory. Rev. Appl. Soft Comput. **9**, 1–2 (2009)
4. Shan, N., Ziarko, W.: Data-based acquisition and incremental modification of classification rules. Comput. Intell. **11**(2), 357–370 (1995)

5. Grzymala-Busse, J.W.: LERS – a system for learning from examples based on rough sets. In: Słowiński, R. (ed.) Intelligent Decision Support, Handbook of Applications and Advances of the Rough Sets Theory, pp. 3–18. Kluwer Academic Publishers, Boston (1992)
6. Ziarko, W.: Variable precision rough set model. J. Comput. Syst. Sci. **46**, 39–59 (1993)
7. Yao, Y.: Three-way decision: an interpretation of rules in rough set theory. In: Wen, P., Li, Y., Polkowski, L., Yao, Y., Tsumoto, S., Wang, G. (eds.) RSKT 2009. LNCS, vol. 5589, pp. 642–649. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02962-2_81
8. Yao, Y.: Three-way decision with probabilistic rough sets. Inf. Sci. **180**, 341–353 (2010)
9. Yao, Y.: Rough sets and three-way decisions. In: Ciucci, D., Wang, G., Mitra, S., Wu, W.-Z. (eds.) RSKT 2015. LNCS (LNAI), vol. 9436, pp. 62–73. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25754-9_6
10. Matsubayashi, T., Kato, Y., Saeki, T.: A new rule induction method from a decision table using a statistical test. In: Li, T., et al. (eds.) RSKT 2012. LNCS (LNAI), vol. 7414, pp. 81–90. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31900-6_11
11. 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, vol. 8171, pp. 213–220. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41299-8_20
12. Kato, Y., Saeki, T., Mizuno, S.: Considerations on rule induction procedures by STRIM and their relationship to VPRS. In: Kryszkiewicz, M., Cornelis, C., Ciucci, D., Medina-Moreno, J., Motoda, H., Raś, Z.W. (eds.) RSEISP 2014. LNCS (LNAI), vol. 8537, pp. 198–208. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-08729-0_19
13. Kato, Y., Saeki, T., Mizuno, S.: Proposal of a statistical test rule induction method by use of the decision table. Appl. Soft Comput. **28**, 160–166 (2015)
14. Kato, Y., Saeki, T., Mizuno, S.: Proposal for a statistical reduct method for decision tables. In: Ciucci, D., Wang, G., Mitra, S., Wu, W.-Z. (eds.) RSKT 2015. LNCS (LNAI), vol. 9436, pp. 140–152. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25754-9_13
15. Kitazaki, Y., Saeki, T., Kato, Y.: Performance comparison to a classification problem by the second method of quantification and STRIM. In: Flores, V., et al. (eds.) IJCRS 2016. LNCS (LNAI), vol. 9920, pp. 406–415. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-47160-0_37
16. Fei, J., Saeki, T., Kato, Y.: Proposal for a new reduct method for decision tables and an improved STRIM. In: Tan, Y., Takagi, H., Shi, Y. (eds.) DMBD 2017. LNCS, vol. 10387, pp. 366–378. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61845-6_37
17. Kato, Y., Itsuno, T., Saeki, T.: Proposal of dominance-based rough set approach by STRIM and its applied example. IJCRS 2017, Part I. LNCS (LNAI), vol. 10313, pp. 418–431. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-60837-2_35
18. Walpole, R.E., Myers, R.H., Myers, S.L., Ye, K.: Probability and Statistics for Engineers and Scientists, 8th edn, pp. 187–191. Pearson Prentice Hall, Upper Saddle River (2007)
19. Dud, R., Hart, P.E.: Pattern Classification and Scene Analysis. Wiley, New York (1973)