Proposal of a New Classification Method Using Rule Table and Its Consideration

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

[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

Abstract. This paper proposed a new classification method using a rule table and demonstrated how to derive the rule table from the Rakuten Travel dataset that represents real-world datasets and how to use it for classification problems. The usefulness of the proposed method was shown using the classification rate referring to the random forest method. The proposed rule table concept showed the expansion of the if–then rules induced by the previous statistical test rule induction method including basic rules called trunk rules behind the dataset and the usefulness for various levels of rule description for real-world datasets.

Keywords: decision table · if-then rule · classification problem · random forest · principle of incompatibility

1 Introduction

With the growth of various network societies, numerous electric datasets are generated and stored for use under different policies and/or business strategies, and such datasets are often arranged in each suitable form for the application. This paper considered a Rakuten Travel dataset (R-dataset) [1,12] which is a real-world dataset (RWD) of questionnaire surveys with the accommodation (object) rating some feature items of each object and its overall category, and a typical decision table (DT) in the field of the Rough Sets (RSs) [2]. This paper proposed a new method for arranging the R-dataset into a rule table (RT), presenting the relationships between the feature items of each object and the overall category, and applying these relationships for classifying a new object into its belonging overall category. The classification results were evaluated using the random forest (RF) [3].

In addition, as mentioned in the principle of incompatibility [4], accurate arrangement and/or summarization against the dataset as well as the comprehensive expressions for decision making are necessary to support real-world activities. However, the arrangement using the RT was too complex for human beings, as also in the case of RF, and lost comprehensibility. After showing that the RT includes if-then rule candidates (RCs) with a proper statistical significance level

K. Honda et al. (Eds.): IUKM 2023, LNAI 14376, pp. 55–66, 2024.
https://doi.org/10.1007/978-3-031-46781-3_6

induced by the previous statistical test rule induction method (STRIM) [5–20] and RCs without it, the former RCs were also pointed out including the intuitively comprehensive basic rules. These three types of RCs in the dataset can organize the rule set, balancing accuracy and comprehensibility depending on the target matter. Meanwhile, RF does not provide such knowledge or information behind the dataset, although it remains a useful method for classification problems. In this way, the usefulness of the proposed method was confirmed.

2 Conventional RS and Its Rule Induction Method

The R-dataset is a typical DT in the field of the RSs [2] and the DT is formulated with an observation system S as follows: $S = (U, A = C \cup \{D\}, V, \rho)$, where $U = \{u(i)|i = 1, ..., N = |U|\}$ is a dataset, $u(i)$ denotes an object in a population, A denotes a set of given attributes of U, $C = \{C(j)|\, j = 1, ..., |C|\}$ denotes a set of the condition attribute $C(j)$, and D denotes a decision attribute. Meanwhile, V denotes a value set of the attribute, i.e., $V = \bigcup_{a \in A} V_a$, where V_a denotes the set of values for an attribute a and $\rho : U \times A \to V$ is called an information function. For example, let $a = C(j)$ $(j = 1, ..., |C|)$, then $V_a = \{1, 2, \ldots, M_{C(j)}\}$. If $a = D$, then $V_a = \{1, 2, \ldots, M_D\}$. Corresponding relationships with the R-dataset are given as follows: $|C| = 6$, $A = \{C(1) = Location$, $C(2) = Room$, $C(3) = Meal$, $C(4) = Bath\,(HotSpring)$, $C(5) = Service$, $C(6) = Amenity$, $D = Overall\}$, and $V_a = \{1 : Dissatisfied$, $2 : Slightly\,Dissatisfied$, $3 : Neither\,Dissatisfied\,nor\,Satisfied$, $4 : Slightly\,Satisfied$, $5 : Very\,Satisfied\}$, $a \in A$, i.e., $|V_{a=D}| = M_D = |V_{a=C(j)}| = M_{C(j)} = 5$.

The conventional RS theory finds the following subsets of U through C and D:

$$C_*(D_d) \subseteq D_d \subseteq C^*(D_d). \tag{1}$$

Here, $C_*(D_d) = \{u_i \in U|[u_i]_C \subseteq D_d\}$, $C^*(D_d) = \{u_i \in U|[u_i]_C \cap D_d \neq \emptyset\}$, $[u_i]_C = \{u(j) \in U|(u(j), u_i) \in I_C, u_i \in U\}$, and $I_C = \{(u(i), u(j)) \in U^2|\rho(u(i), a) = \rho(u(j), a), \forall a \in C\}$, where $[u_i]$ denotes the equivalence class with the representative u_i induced by the equivalence relation I_C , and $D_d = \{u(i)|(\rho(u(i), D) = d\}$. In equation (1), $C_*(D_d)$ and $C^*(D_d)$ are called the lower and upper approximation of D_d, respectively, and $(C_*(D_d)$,$C^*(D_d))$ is the rough set for D_d. Being found $C_*(D_d) = \{u(i)| \wedge_j (\rho(u(i), C(j)) = v_{j_k})\}$ by using the DT, the following if–then rule with necessity is obtained using the inclusion relation in equation (1): if CP then $D = d$, where the condition part (CP) is specifically $CP = \wedge_j (C(j) = v_{j_k})$. Similarly, the if–then rule with possibility is induced using $C^*(D_d)$. Thus, the conventional RS theory derives relationships between $C = (C(1), \ldots, C(6))$ and D. The specific algorithm and RSs can be respectively referred to in the literature [2,21].

However, in most cases, $u(i) = (u^C(i), u^D(i))$ is randomly collected from a population of interest so that the attribute values $u^C(i) = (v_{C(1)}(i), ..., v_{C(6)}(i))$ $(v_{C(j)}(i)\,(\in V_{C(j)}))$ or $u^D(i) = v_d(i)\,(\in V_D)$ follow random variations. The collection of the dataset from the same population indicates that U will variate

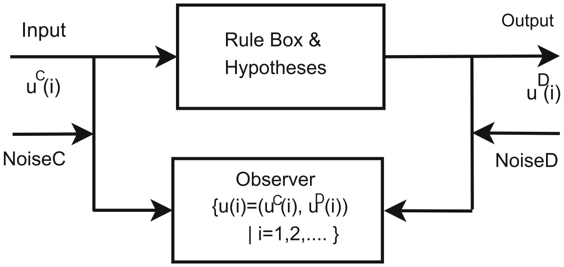

Fig. 1. Data generation model: Rule Box contains if-then rules and Hypotheses regulate how to apply rules for Input and transform Input into Output.

such that the corresponding induced rules also variate since the conventional RS theory directly uses the DT attribute values [5,8,13]. As a result, the rules induced by the conventional RSs do not fulfill the function, e.g., for the classification problem. In statistics, $C(j)$ and D are recognized as random variables, and $v_{C(j)}(i)\,(\in V_{C(j)})$ and $v_d(i)\,(\in V_D)$ are their respective outcomes. The conventional RS theory lacks these statistical views and does not have a model for collecting the DT values.

3 Outlines of Data Generation Model

The previous STRIM proposed a data generation model as shown in Fig. 1 in which Rule Box and Hypotheses transformed an input $u^C(i)$ into the output $u^D(i)$. Here, Rule Box contains pre-specified if-then rules and Hypotheses regulate how to apply those rules for the input and transform the input into the output. The model was used for generating a dataset on a simulation experiment as follows: (1) specifying some proper if-then rules in Rule Box, (2) randomly generating an input and transforming it into the output based on those rules and Hypotheses, (3) repeating (1) and (2) N times and forming $U = \{u(i) = (u^C(i), u^D(i))|i = 1, ..., N = |U|\}$. The generated U was used for investigating the rule induction abilities by applying it for any rule induction method (see details [5–20]). The previous STRIM [20] could induce the pre-specified rules while the rules induced by the method like RSs [15,21], CART [14,22] or association rule [19], [23] included a lot of meaningless rules and hardly corresponded with the pre-specified rules.

4 Introduction of RT and Its Application for Classification Problem

The validity and usefulness of the previous STRIM have been confirmed in simulation experiments which generate the dataset obeying pre-specified rules, that is, a well-behaved dataset. This paper newly expanded Rule Box and Hypotheses in Fig. 1 so as to adapt the R-dataset as one of RWD which includes an ill-behaved dataset caused by various raters with different rating standard, and investigated its rule induction abilities. Generally, the result of the rule induction from an RWD cannot be directly ascertained and increases the complexity of rules' description. This paper investigates the ability in the classification problem having the complimentary relation to the rule induction problem in which induced rules directly affect the classification rate.

Let there be a new relationship between $C(j)$ $(j = 1, ..., 6)$ and D by using the R-dataset. This section shows how to form the RT and apply the new relationship for a classification problem. The R-dataset of $N=$ 10,000 was first formed by randomly selecting 2,000 samples, each of $D = m$ $(m = 1, \ldots, 5)$ from about 400,000 surveys in the 2013–2014 dataset. The R-dataset was randomly divided into two groups. One is the R_L-dataset with N_L=5,000 for learning the R-dataset and the other is the R_C-dataset with N_C=5,000 for the classification experiment.

Regarding the learning process, let us consider a specific example of learning data: $(C(1), \ldots, C(6), D) = (1, 2, 3, 4, 5, 1, 3)$. This data can be derived by an if–then rule: if $CP = (C(1) = 1) \wedge (C(2) = 2) \wedge (C(3) = 3)$ (hereafter denoted with $CP = (123000)$) then $D = 3$. This rule is called the rule with rule length 3 $(RL = 3)$ as it involves three conditions. Assuming $RL = 3$, $CP = (023400)$ can be considered another RC. Thus, all the RCs with $RL = 3$ in this example can be ${}_6C_r|_{r=3} = 20$ different ways. Accordingly, all the possible RCs with $RL = 3$ is ${}_6C_r(5)^r|_{r=3} = 2,500$. The R_L-dataset was arranged in the RT with $RL = 3$ $(RT(r = 3))$ as shown in Table 1. The first row of the table (1,2,3) = (1,1,1) represents the CP: $(C(1) = 1) \wedge (C(2) = 1) \wedge (C(3) = 1)$. Meanwhile, (79,0,0,0,0) is the frequency distribution of D satisfying the condition in the R_L-dataset. In other words, $D = 1$ represent the maximum frequency (if there are the same frequencies, D is randomly selected between them). Most of the distributions of D in Table 1 widely fluctuate corresponding to the same CP, which was caused by different raters with varying standards. If an RC has a frequency of (0,0,0,0,0), it is called the empty RC. In Table 1, each RC is called a sub-rule of the RT. In addition, by using the $RT(r)$, the rule set $\cup_{r=1}^{|C|=6} RT(r)$ includes all the rules behind the R_L-dataset. This RT is the newly expanded Rule Box in Fig. 1.

With respect to the classification process of transforming an input $u^C(i)$ into the output$u^D(i)$, sub-rules of ${}_6C_r|_{r=3} = 20$ that match the input pattern should be considered to adapt to different rating standards. Therefore, this paper adopted the vote of each sub-rule's output. Figure 2 provides a specific example where the input $u^C(i) = (4, 5, 1, 4, 4, 3)$ $(u^D(i) = 4)$ is classified by 20 sub-rules with $RL = 3$ using the RT in Table 1. In the first row, the input values $(C(1) = 4, C(2) = 5, C(3) = 1)$ correspond to the CP of the sub-rule: (1,2,3) = (4,5,1) in the RT. It is classified as $D = 1$ by selecting the maximum frequency. Similarly, in the 19th row, $D = 2$ or 3 is randomly selected. By arranging non-empty cases (i.e., deleting empty sub-rules) and by counting the votes, the maximum frequency is 9 at $\hat{D} = 4$, which is the final result that happens to coincide with $u^D(i)$. In the case of same values, one of them is randomly selected. These processes are the new Hypotheses in Fig. 1.

Table 1. Example of RT with $RL = 3$ of the R_L-dataset.

Condition Part $(C(j1), C(j2), C(j3))(j1 < j2 < j3)$ $(j1, j2, j3) = (k1, k2.k3)$	Frequency of $D(n_1, n_2, ..., n_5)$	Decision Part D
(1,2,3) = (1,1,1)	(79, 0, 0, 0, 0)	1
(1,2,3) = (1,1,2)	(9, 2, 0, 0, 0)	1
(1,2,3) = (1,1,3)	(9, 0, 0, 0, 0)	1
...	...	...
(1,2,3) = (5,5,5)	(8, 6, 4, 28, 445)	5
(1,2,4) = (1,1,1)	(82, 1, 0, 0, 0)	1
...	...	...
(1,2,4) = (5,5,5)	(10, 9, 5, 19, 393)	5
...	...	...
(4,5,6) = (1,1,1)	(188, 17, 0, 0, 0)	1
...	...	...
(4,5,6) = (5,5,5)	(4, 5, 5, 23, 409)	5

input: $(C(1), ..., C(6), D)$	(4, 5, 1, 4, 4, 3, $D = 4$)

↓

$(j1, j2, j3)$	$(k1, k2, k3)$	Distribution of D	D
1: (1,2,3)	(4, 5, 1)	(6, 0, 1, 2, 0)	1
2: (1,2,4)	(4, 5, 4)	(6, 1, 2, 43, 48)	5
3: (1,2,5)	(4, 5, 4)	(1, 0, 1, 37, 27)	4
4: (1,2,6)	(4, 5, 3)	(4, 1, 3, 14, 10)	4
5: (1,3,4)	(4, 1, 4)	(19, 14, 3, 4, 1)	1
...	...	...	...
19: (3,5,6)	(1, 4, 3)	(4, 5, 5, 2, 0)	2 or 3
20: (4,5,6)	(4, 4, 3)	(3, 4, 28, 54, 9)	4

↓

arrange 20 cases output of $\hat{D}$	(1, 5, 4, 4, 1, 2, 2, 4, 4, 4, 1, 1, 2, 4, 4, 4, 1, 1, 2, 4) → (6,4,0,9,1) → 4

Fig. 2. Example of classification process by RT with $RL = 3$

5 Classification Experiments on R-Dataset Using RT and Comparison With RF

The RT accompanied with the classification method proposed in Sect. 4 was applied for the R-dataset to investigate its ability and confirm the usefulness. The experiment was executed for every $RT(r)$ $(r = 1, \ldots, 6)$ as follows: 1) composing $RT(r)$ using the R_L-dataset, 2) forming the classification dataset by randomly sampling $N_b = 500$ from the R_C-dataset, 3) classifying the N_b dataset according to the classification process (see Fig. 2), and 4) repeating the previous three procedures by $N_r = 50$ times. Table 2 summarizes one of the results classified by $RT(r)$ $(r = 1, \ldots, 6)$ with $(m_{RT(r)}, SD_{RT(r)})$ [%]. Here $m_{RT(r)}$ and $SD_{RT(r)}$ denote the N_r times mean of the classification rate and its standard deviation, respectively. The following represents the comparisons between the results by $RT(RL = r)$ ($r = 1, \ldots, 6$):

(1) When RL is small, sub-rule accuracy tend to be low, while their coverage tends to be high. Consequently, there are rarely any empty sub-rules, and the frequency distribution bias of D is low, leading to a lower classification ability. Here, $accuracy = |U(d) \bigcap U(CP)|/|U(CP)| = P(D = d|CP)$, $coverage = |U(d) \bigcap U(CP)|/|U(d)| = P(CP|D = d)$, $U(d) = \{u(i)|u^{D=d}(i)\}$, $U(CP) = \{u(i)|u^C(i)\, satisfies\, CP\}$.

(2) When RL becomes too high, the sub-rule accuracy tend to increase, while the coverage decreases. Consequently, there are many empty sub-rules, leading to a decrease in the classification ability.

(3) Suppose that $X_{r1} \sim N(\mu_{r1}, \sigma_{r1}^2)$ and $X_{r2} \sim N(\mu_{r2}, \sigma_{r2}^2)$. Here, X_{r1} and X_{r2} denote random variables of the classification rate by $RT(r1)$ and $RT(r2)$, respectively. The N_r times mean $\overline{X_{r1}}$and $\overline{X_{r2}}$, and their normalized difference $\overline{X_{r1}} - \overline{X_{r2}}$ is given as follows

$$Z = \frac{\overline{X_{r1}} - \overline{X_{r2}} - (\mu_{r1} - \mu_{r2})}{\left(\frac{\sigma_{r1}^2}{N_r} + \frac{\sigma_{r2}^2}{N_r}\right)^{0.5}}. \tag{2}$$

Under null hypothesis $H0$: $\mu_{r1} = \mu_{r2}$, $Z = \frac{\overline{X_{r1}} - \overline{X_{r2}}}{\left(\frac{\sigma_{r1}^2}{N_r} + \frac{\sigma_{r2}^2}{N_r}\right)^{0.5}} \sim N(0, 1)$.

For example, placing $\overline{X_{r1}} = m_{RT(4)}$, $\sigma_{r1} = SD_{RT(4)}$, $\overline{X_{r2}} = m_{RT(3)}$, $\sigma_{RT(3)} = SD_{RT(3)}$, z =2.83 with p-value = 2.31E-3, resulting in the rejection of $H0$, i.e., statistically $\mu_{RT(4)} > \mu_{RT(3)}$.

Accordingly, $RT(r = 4)$ was used for the classification experiment in the R_C-dataset due to the highest classification rate. Table 3 presents one of the results classified by $RT(r = 4)$, arranged as the confusion matrix of all the datasets (500×50). For example, the first row shows the total data number of $D = 1$ is (3,827+1,032+95+66+52) = 5,072, the rate classified $D = 1$ is 3,827/5,072 = 0.755, $D = 2$ is 1032/5,072 = 0.203, and the class error is 0.245.

The same experiment was conducted for the same R_L-dataset and R_C-dataset by RF for comparison with the classification results by the RT method. Specifically, the same R_L-dataset was first used for the learning RF model as shown in Fig. 3: classmodel = randomForest(x, y, $mtry = 2$). Here, the function: randomForest is in the R-language library [24], $\{u^C(i) \in R_L$-datase$\}$ was set to x, and $\{u^D(i) \in R_L$-datase$\}$ was set to y after changing their data classes appropriately. The parameter $mtry = 2$ was found to be the least error rate of out of bag (OOB) in the preliminary experiment. Figure 3 shows one of the outputs of

Table 2. Summary of classification experiment for R-dataset by $RT(r)$, RF and tr-STRIM.

r	$RT(1)$	$RT(2)$	$RT(3)$	$RT(4)$	$RT(5)$	$RT(6)$	RF	$tr-STRIM$
$(m_r,$	(56.0,	(63.0,	(67.4,	(68.6,	(65.2,	(56.1,	(68.6	(60.5,
$SD_r)$	2.08)	2.07)	2.09)	2.25)	1.95)	2.16)	1.70)	2.12)

Table 3. Results of R-dataset classification experiment by $RT(r = 4)$.

D	1	2	3	4	5	class error
1	3827(0.755)	1032(0.203)	95(0.019)	66(0.013)	52(0.010)	0.245
2	1336(0.266)	2718(0.542)	770(0.153)	155(0.031)	38(0.008)	0.458
3	169(0.034)	994(0.203)	3099(0.632)	575(0.117)	67(0.014)	0.368
4	30(0.006)	86(0.017)	750(0.147)	3412(0.670)	813(0.160)	0.330
5	8(0.002)	6(0.001)	61(0.012)	739(0.150)	4102(0.834)	0.166

"classmodel" is $(729 + 201 + 32 + 15 + 11) = 988$ for the dataset of $D = 1$ and class error $= (201+32+15+11)/988 = 0.262$. Table 4 corresponding to Table 3 shows one of the results of the classification experiment by the RF implemented in Fig. 3 and each class error is similar as that in Fig. 3. The N_r times mean rate of the classification rate and its standard deviation was $(m_{RF}, SD_{RF}) =$ (68.6,1.70) [%], as shown in Table 2.

The comparison of the classification results of the R-dataset between $RT(r = 4)$ and RF revealed no significant difference between $\mu_{RT(4)} > \mu_{RF}$ and $\mu_{RT(4)} < \mu_{RF}$ using equation (2) ($z = 0.075$), indicating that $\mu_{RT(4)} = \mu_{RF}$ for time being. Figure 4 also shows the comparison of class errors with D between Table 3 (by $RT(r = 4)$) and Table 4 (by RF) which shows the same tendency and both appears to execute the classification based on the close rules each other. RF is also one of the classification methods which uses the voting result through a large number of decision trees with randomly selected variables to decrease the correlation among those decision trees, improving the CART method by constructing a tree structure [3,22]. However, the difference between them is that the RT explicitly induces rules whereas RF cannot.

6 Proposal of Trunk Rules and Its Consideration

The classification experiments on the R-dataset by the RT method accompanied with classification procedures showed the equivalent ability to that by RF. However, the RT method arranged the R-dataset into numerous sub-rules and used the RT for voting to obtain the classified result, although the method was adaptive to the ill-behaved RWD, resulting in RT losing comprehensibility for human beings. Meanwhile, the previous STRIM [5–19] used the following principle for exploring the CP of if–then rules: $P(D = d|CP) \neq P(D = d)$, setting the null hypothesis $H0$: the CP is not a rule candidate $(P(D = d|CP) = P(D = d))$

```
Call:
randomForest(x = x, y = y, mtry = 2)
Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 2
OOB estimate of error rate: 30.9%
Confusion matrix:
       1      2      3      4      5 class.error
1    729    201     32     15     11       0.262
2    239    543    165     22      4       0.442
3     28    192    667    132      6       0.349
4      6     12    169    652    148       0.339
5      1      1      8    152    865       0.158
```

Fig. 3. Results of error rate of OOB and confusion matrix in R_L-dataset.

Table 4. Results of classification experiment for R_C-dataset by RF

D	1	2	3	4	5	class error
1	3785(0.744)	1066(0.209)	149(0.029)	56(0.011)	33(0.006)	0.256
2	1316(0.261)	2821(0.559)	752(0.149)	135(0.027)	22(0.004)	0.441
3	156(0.032)	936(0.190)	3200(0.650)	555(0.113)	77(0.016)	0.350
4	7(0.001)	98(0.019)	821(0.162)	3283(0.649)	848(0.168)	0.351
5	5(0.001)	8(0.002)	51(0.010)	760(0.156)	4060(0.831)	0.169

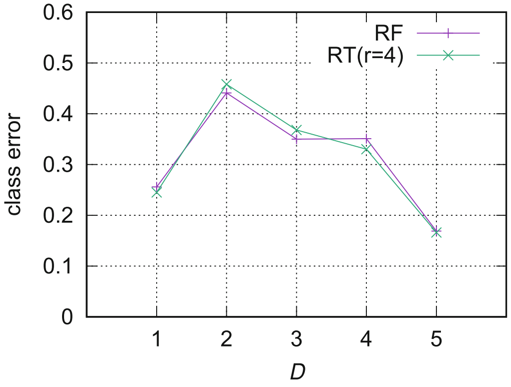

Fig. 4. Class error tendency corresponding to Tables 3 and 4

and executing the statistical test using the R_L-dataset. The frequency distribution of D, e.g., (79,0,0,0,0) at the first row in Table 1, rejects the null hypothesis, i.e., it is a rule candidate. However, $H0$ cannot be rejected by the second (9,2,0,0,0) and the third (9,0,0,0,0) as they do not satisfy the necessary test sample size n (in this specification, approximately $n \geq 25$ and the sample size of the second is $9 + 2 = 11$ (see [13])). Thus, the RT can be divided into two types of sub-rules: rejecting $H0$ and not rejecting. Table 5 shows the number of induced RCs from the whole of RT: $\cup_{r=1}^{|C|=6} RT(r)$, which satisfies the principle with $p-value < 1.0E-5$. It arranges the induced RCs by every $RL = r$, which coincides with the result by the previous STRIM. That is, the RT expands the range of RCs by the previous STRIM, and the RT with $RL = r$ method was labeled as expanded STRIM ($ex - STRIM|RL = r$).

Although the details were omitted due to space limitations, all the induced RCs with $RL = 1$ in Table 5 are of the following form:

$$if\, C(j) = d\, then\, D = d\, (j = 1, \ldots, 6,\, d = 1, \ldots, 5). \tag{3}$$

Table 6 shows the RCs with $RL \geq 2$, having only $C(j) = d$, extracted from Table 5 and arranged in the same manner as Table 5. The number of all RCs with $RL = r$ constructed by only $C(j) = d$ is given by ${}_6C_r$ and presented in Table 6, except for the cases of $D = 2$ with $RL = 4$, 5, and 6. In addition, eight RCs of $D = 2$ with $RL = 4$ were discovered: CP= (222002), (220202), (220022), (022220),(022202), (022022), (020222), and (002222). The values in both the condition and decision parts of the if–then rule coincide with each other, considering that the rating scale of $C(j)$ ($j = 1, \ldots, 6$) is the same ordinal scale including D. Consequently, the RT or the previous STRIM induced such understandable RCs, which were labeled trunk rules, and an inducing method trunk STRIM (tr-STRIM). As shown in Table 2, a classification result by the tr-STRIM was (m_{tr-ST}, SD_{tr-ST}) = (60.5,2.12) [%], which is positioned in the middle of $RT(r)$ with r= 1 and r= 2.

The inclusion relationship of RCs induced by the three types of STRIM is arranged as follows:

(1) $Rset(ex - STRIM|RL = r) \supset Rset(STRIM|RL = r) \supset Rset(tr - STRIM|RL = r)$,

(2) $\cup_{r=1}^{|C|=6} Rset(ex - STRIM|RL = r) \supset Rset(STRIM) \supset Rset(tr - STRIM)$,

(3) $Rset(ex - STRIM|r = 1) \supset Rset(ex - STRIM|r = 2) \supset, \ldots, Rset(ex - STRIM|r = 6)$.

In this context, $Rset(method)$ refers to the rule set induced by a particular method. The average classified results can be summarized as follows, including no-show relationships due to space limitations: $C(Rset(ex - STRIM|RL = r), data) > C(Rset(STRIM|RL = r), data) > C(Rset(tr - STRIM|RL = r), data)$, where $C(Rset(method), data)$ represents the average classification result against a dataset by $Rset(method)$. To express the classification results qualitatively, the trunk rules were improved by adding the branch rules with statistical significance level and further by the leaf rules without it. These studies can be beneficial in considering a level of "the principle of incompatibility" depending on the targeted matter. For instance, these studies can be useful in various policies or business strategies where RF cannot explicitly induce rules, indicating that the contents of the dataset and the level need to be considered.

Table 5. Number of induced rules with statistical significance level from whole of RT.

	$RL = 1$	2	3	4	5	6	Total
$D = 1$	6	81	226	78	13	2	406
2	6	61	158	49	1	0	275
3	6	43	124	134	33	3	343
4	6	45	149	141	22	2	365
5	6	50	142	151	53	5	407
Total	30	280	799	553	122	12	1796

Table 6. Number of extracted trunk rules for each RL from Table 5

	$RL = 1$	2	3	4	5	6	Total
$D = 1$	6	15	20	15	6	1	63
2	6	15	20	8	0	0	49
3	6	15	20	15	6	1	63
4	6	15	20	15	6	1	63
5	6	15	20	15	6	1	63
Total	30	75	100	68	24	4	301

7 Conclusion

The validity and usefulness of the previous rule induction method, STRIM have been confirmed using a simulation experiment. However, the examination of its usefulness in an RWD was left for future study [8,13,20]. This study specifically used the R-dataset with DT in the field of RS for experimentally addressing the issue by newly proposing the RT method for adaption to the RWD. The validity or usefulness of the method was confirmed by applying it to the classification problem, as their confirmation of the rule induction from the RWD cannot be generally ascertained, even though the result directly reflects the classification result. The usefulness of the method was confirmed using the classification result by RF. The following aspects were specifically considered:

(1) The newly proposed RT method demonstrated how to arrange the dataset into $RT(r)$ which is a set of sub-rules and all the possible RCs with $RL = r$, and how to use $RT(r)$ to classify a new object.

(2) The validity and usefulness of the RT method were experimentally examined by applying it to the classification problem after selecting a proper $RT(r)$. The result of the classification rate showed equivalence to that by RF, i.e., $\mu_{RT(r=4)} = \mu_{RF}$.

(3) However, the induced $RT(r)$ and the decision trees generated by RF are increasingly difficult to understand. Both methods offer limited insights and information regarding the dataset in the form of if–then rules. Thus, the whole of RT: $\cup_{r=1}^{|C|=6} RT(r)$ was subsequently reviewed and then organized into trunk rules using the trunk-STRIM method. The relationships between these rules were shown as $Rset(ex - STRIM|RL = r) \supset Rset(STRIM|RL = r) \supset Rset(tr - STRIM|RL = r)$ and were useful in providing "the principle of incompatibility" depending on the targeted matter. On the other hand, RF does not provide such knowledge, although it remains a useful method for classification problems.

The investigation of the following points is recommended for future studies:

(1) To validate the findings of this study, future research should focus on applying the three types of STRIM, namely the previous, ex-STRIM, and tr-STRIM, to various other RWDs and replicate the findings to validate the findings of this study.

(2) When using ex-STRIM with $RL = r$ as a classification method, whether $\mu_{(ex-ST|r)} = \mu_{RF}$ is always equivalent or not should be studied, considering factors, such as the size of the learning dataset and changes in RWD.

Acknowledgments. This work was supported by JSPS KAKENHI Grant Number JP20K11939. We truly thank Rakuten Inc. for presenting Rakuten Travel dataset [1].

References

1. https://www.nii.ac.jp/dsc/idr/rakuten/
2. Pawlak, Z.: Rough sets. Int. J. Comput. Inf. Sci. **11**(5), 341–356 (1982)
3. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
4. Zadeh, L.A.: Outline of a new approach to the analysis of complex systems and decision processes. IEEE Trans. Syst. Man Cybern. **3**, 28–44 (1973)
5. 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). https://doi.org/10.1007/978-3-642-31900-6_11
6. 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). https://doi.org/10.1007/978-3-642-41299-8_20
7. 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
8. 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)
9. 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

10. 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
11. 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
12. Kato, Y., Itsuno, T., Saeki, T.: Proposal of dominance-based rough set approach by STRIM and its applied example. In: Polkowski, L., et al. (eds.) IJCRS 2017. LNCS (LNAI), vol. 10313, pp. 418–431. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-60837-2_35
13. Kato, Y., Saeki, T., Mizuno, S.: Considerations on the principle of rule induction by STRIM and its relationship to the conventional rough sets methods. Appl. Soft Comput. **73**, 933–942 (2018)
14. Kato, Y., Kawaguchi, S., Saeki, T.: Studies on CART's performance in rule induction and comparisons by STRIM. In: Nguyen, H.S., Ha, Q.-T., Li, T., Przybyła-Kasperek, M. (eds.) IJCRS 2018. LNCS (LNAI), vol. 11103, pp. 148–161. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99368-3_12
15. Saeki, T., Fei, J., Kato, Y.: Considerations on rule induction methods by the conventional rough set theory from a view of STRIM. In: Nguyen, H.S., Ha, Q.-T., Li, T., Przybyła-Kasperek, M. (eds.) IJCRS 2018. LNCS (LNAI), vol. 11103, pp. 202–214. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99368-3_16
16. Kato, Y., Saeki, T., Mizuno, S.: Considerations on the Principle of Rule Induction by STRIM and Its Relationship to the Conventional Rough Sets Methods; Applied Soft Computing, 73 pp, pp. 933–942, Elsevier (2018)
17. Kato, Y., Saeki, T., Fei, J.: Application of STRIM to datasets generated by partial correspondence hypothesis. In: Fagan, D., Martín-Vide, C., O'Neill, M., Vega-Rodríguez, M.A. (eds.) TPNC 2018. LNCS, vol. 11324, pp. 74–86. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04070-3_6
18. Kato, Y., Saeki, T.: Studies on reducing the necessary data size for rule induction from the decision table by STRIM. In: Mihálydeák, T., Min, F., Wang, G., Banerjee, M., Düntsch, I., Suraj, Z., Ciucci, D. (eds.) IJCRS 2019. LNCS (LNAI), vol. 11499, pp. 130–143. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22815-6_11
19. Kato, Y., Saeki, T.: New rule induction method by use of a co-occurrence set from the decision table. In: Gutiérrez-Basulto, V., Kliegr, T., Soylu, A., Giese, M., Roman, D. (eds.) RuleML+RR 2020. LNCS, vol. 12173, pp. 54–69. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-57977-7_4
20. Kato, Y., Saeki, T.: Application of Bayesian STRIM to datasets generated via partial correspondence hypothesis. In: 18th International Conference on Machine Learning and Data Mining, MLDM 2022, pp. 1–15 (2022)
21. Grzymala-Busse, J.W. : LERS- A System for Learning from Examples Based on Rough Sets In: Słowiński, R. (ed.) Kluwer Academic Publishers: Intelligent Decision Support; Handbook of Applications and Advances of the Rough Sets Theory, pp. 3–18 (1992)
22. Decision tree learning - Wikipedia
23. Association rule learning - Wikipedia
24. randomForest: Breiman and Cutler's Random Forests for Classification and Regression r-project.org