Application of STRIM to Datasets Generated by Partial Correspondence Hypothesis

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

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

Abstract. STRIM (Statistical Test Rule Induction Method) has been proposed for an if-then rule induction method from the decision table independently of Rough Sets theory, not utilizing the notion of the approximation and the validity of the method has also been confirmed by a simulation model for data generation and verification of induced rules. However, the previous STRIM used a plain hypothesis of the complete correspondence with rules while a real-world dataset judged by human beings often seems to obey a partial correspondence hypothesis (PCH). This paper studies STRIM incorporating the PCH and improves the previous STRIM into a new version, STRIM2, of which performance and caution for use is examined by the above simulation model incorporating PCH. STRIM2 is also applied to the real-world dataset and draws results showing interesting suggestions.

Keywords: Rough sets · Statistical method · If-then rules

1 Introduction

Nowadays, a large number of electronic datasets are being generated with the growth of a network society. Among such datasets, those generated in the e-commerce area are used for various business strategies and such trials have recently proliferated quickly. The e-commerce takes in the various datasets including their attributes with regard to items for sale as well as their customers so that their relationships, structures and features are easily analyzed and used for strategies of providing it with new items and/or services for sale as well as acquiring new customers. In those processes, the conventional data mining or analyzing methods are used, or new methods are needed and developed for improving their precision and adaptation of new aims. Demands from such a network society generate research and development in those data science areas.

A statistical test rule induction method (STRIM) [1–8] also has been proposed for improving rule induction methods by the conventional Rough Sets

D. Fagan et al. (Eds.): TPNC 2018, LNCS 11324, pp. 74–86, 2018.
https://doi.org/10.1007/978-3-030-04070-3_6

methods [9–12] which are used for inducing if-then rules from a dataset called the decision table. Specifically, STRIM recognized the if-then rules as an input-output system and proposed a data generation model for the decision table in order to clarify the relationship between if-then rules and the decision table, the stochastic uncertainty included in the table and what is a rule hidden in the table. The data generation model made up for faults of the conventional Rough Sets lacking statistical views. An algorithm for the rule induction by STRIM also has been proposed and the validity and the usefulness have been confirmed by applying it to real-world datasets after simulation experiments.

However, the plain hypotheses were used in the process of transforming the input into the output in order to simply study the data generation process. Specifically, the previous data generation process used a complete correspondence hypothesis (CCH) that the input was transformed by the pre-specified rules only when it completely corresponded with them. In the real-world, human beings often use their rules even when the input partially corresponds with them and they decide to compromise with the second best. This paper experimentally studies an if-then rule induction problems from the dataset generated based on a partial correspondence hypothesis (PCH) in order to better match the previous STRIM to the real-world dataset judged in the processes such as human decision-making. Specifically, the previous STRIM is first applied to the PCH dataset in a simulation experiment. The experimental consideration suggests that the interim results by the previous STRIM can be used for inferring the original rules by use of a Hamming distance and a technique of a one-strike sketch. STRIM2 named after the revised STRIM is applied for the real-world dataset, Rakuten Travel dataset and draws results showing interesting suggestions.

2 Introduction of Decision-Making Processes

In statistics, a dataset $U = \{u(i)|i = 1, ..., N = |U|\}$ is collected from a population of interest to estimate and/or infer properties and features of the population. Here, $u(i)$ is an object with several attributes, whose properties and features contribute to the estimation and inference of the population. Let us denote an observation system by $S = (U, A, V)$. Here, A is the set of an attribute and V is the set of the attribute's values; that is, $V = \bigcup_{a \in A} V_a$ and V_a is the set of the value of attribute a. When randomly sampling $u(i)$ from the population, each attribute becomes a random variable with the respective attribute value as its outcome.

Here, there are two main types of datasets, with a division between the response and explanatory variables and those without it. In the former case, the set of attributes A is denoted $A = C \cup \{D\}$ to distinguish from the latter case. Here, D is a decision attribute and the response variable, and $C = \{C(j)|j = 1, ..., |C|\}$ is the set of condition attribute $C(j)$ and $C(j)$ is also an explanatory variable for the response variable. If D and $C(j)$ are qualitative variables, D represents the random variable of the class containing $u(i)$ and is affected by the set C of the random variable $C(j)$. This paper studies the former case dealing with qualitative variables based on the system $S = (U, A = C \cup D, V)$ called the decision table in the Rough Sets theory.

Figure 1 outlines the data generation process. Randomly sampling $u(i)$ from the population, the outcome of $C = (C(1), ..., C(|C|))$; that is, $u^C(i) = (v_{C(1)}(i), ..., v_{C(|C|)}(i))$ is obtained and becomes the input into the rule box. The rule box transforms $u^C(i)$ into the output $u^D(i)$ using the rule box's pre-specified rules $R(d,k)$: if $CP(d,k)$ then $D = d$ $(d = 1, 2, ..., k = 1, 2, ...)$ and the following partial correspondence hypothesis with the input modifying CCH shown in Table 1.

Partial correspondence hypothesis (PCH): The degree Dgr of $u^C(i)$ for correspondence with the box's pre-specified rules is estimated and the rule of the highest Dgr is applied for transforming $u^C(i)$ into $u^D(i)$. If there are several rules of ties, one of them is randomly determined in the same way as Hypothesis 3 in Table 1. PCH expands and generalizes three cases for $u^C(i)$ in Table 1 for CCH, taking human decision-making into account. The observer in Fig. 1 records $u(i) = (u^C(i), u^D(i))$. NoiseC and NoiseD are introduced to adapt the model for the real-world dataset. NoiseC adjusts the value of $u^C(i) = (v_{C(1)}(i), ..., v_{C(|C|)}(i))$ or makes $v_{C(j)}(i)$ a missing value, and NoiseD adjusts the value of $u^D(i)$.

Generating $u^C(i) = (v_{C(1)}(i), ..., v_{C(|C|)}(i))$ using random numbers and transforming it into $u^D(i)$ using the model shown in Fig. 1, including PCH, $U = \{u(i) = (u^C(i), u^D(i)) | i = 1, ..., N = |U|\}$ can be obtained and applied to any rule induction method to investigate the extent to which the method applied induces the pre-specified rules.

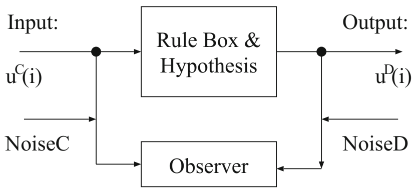

Fig. 1. A simulation model for data generation and verification of induced rules. The rule box contains if-then rules $R(d,k)$: if $CP(d,k)$ then $D = d$ $(d = 1, 2, ..., k = 1, 2, ...)$.

3 Simulation Experiment by the Previous STRIM

We implemented the data generation process with PCH and the verification process applying the previous STRIM as follows: (1) Specified rules, for example, shown in Table 2 in the rule box in Fig. 1, where $|C| = 6$, $V_a = \{1, 2, ..., 6\}$ $(a = C(j)(j = 1, ..., |C|), a = D)$, and $CP(1,1) = 110010$ denoted $CP(1,1) = (C(1) = 1) \bigwedge (C(2) = 1) \bigwedge (C(5) = 1)$ and was called a rule of the rule length 3 $(RL = 3)$ having three conditions. (2) Generated $v_{C(j)}(i)$ $(j = 1, ..., |C| = 6)$ with a uniform distribution and formed $u^C(i) = (v_{C(1)}(i), ..., v_{C(6)}(i))$ $(i =$

Table 1. Complete correspondence hypothesis with regard to the input.

Hypothesis 1	$u^C(i)$ coincides with $R(d,k)$, and $u^D(i)$ is uniquely determined as $D = d$ (uniquely determined data)
Hypothesis 2	$u^C(i)$ does not coincide with any $R(d,k)$, and $u^D(i)$ can only be determined randomly (indifferent data)
Hypothesis 3	$u^C(i)$ coincides with several $R(d,k)$ $(d = d1, d2, ...)$, 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)

$1, ..., N = 10,000$). (3) Transformed $u^C(i)$ into $u^D(i)$ using the pre-specified rules in Table 2 and PCH, without generating NoiseC and NoiseD for a simple experiment. Here, Dgr was simply estimated by the sum of the number of the conditions satisfied for each rule. For example, if $u^C(i) = 112251$ then $Dgr = 2$ at $R(1,1)$, $Dgr = 1$ at $R(1,2)$, $Dgr = 0$ at $R(2,1)$, and so on. Accordingly, $R(1,1)$ or $R(2,2)$ having the highest $Dgr = 2$ were randomly selected. We will refer to the dataset generated based on the above procedures as the PCH dataset. We randomly sampled $N_B = 5,000$ data and formed a new dataset as the decision table.

We applied the previous STRIM [1–8] to the PCH dataset. Figure 2 shows an outline of the algorithm implementing the STRIM written in C-language style (details in [7,8]). At $LN = 8-9$, for each decision attribute value di, the statistically independent condition attributes against di are reducted. At $LN = 10$, the function rule_check() (the body is at $LN = 19-33$) systematically forms a trying rule based on the dimension rule[ ] (condition part of a rule CP). At $LN = 25$, we examine the degree of the validity for the trying rule by the z-value, which is the degree of bias in the frequency distribution of D supposing the standard normal distribution and is used to select the rule as a candidate. The selected candidates are finally arranged into the induced rules at $LN = 12$.

Table 3 shows examples of the results of the arranged rules for $D = 1$ and the part of those for $D = 2$ in descending order of z-values for each D. For example, the first row $CP(1,1)$ of the table means the following: The condition part of the induced rule is $(C(2) = 1) \bigwedge (C(5) = 1)$. The frequency distribution of D $f = (n_1, ..., n_6)$ satisfying the condition is $(138, 3, 4, 4, 6, 6)$, which suggests the maximum frequency n_d of D is $n_{d=1} = 138$ and thus $D = 1$ is the decision part for the rule. The distribution of $z = \frac{n_d + 0.5 - np_d}{(np_d(1-p_d))^{0.5}}$ obeys the standard normal distribution under the null hypothesis H_0: CP is not a rule candidate (the alternative hypothesis H_1: CP is a rule candidate) and the testing condition [13]: $np_d \geq 5$ and $n(1-p_d) \geq 5$, where $n = \sum_{m=1}^{6} n_m$. The p-value corresponding to the z-value is the index of supporting H_0, and the accuracy and the coverage are also shown in the table.

Table 3 shows that the previous STRIM doesn't induce $R(1,1)$ of the pre-specified rules having $RL = 3$ in Table 2 but induces three rules $CP(1,1)$, $CP(1,3)$ and $CP(1,6)$ with $RL = 2$ including $R(1,1)$. Hereafter $R(1,1)$ is called a partial rule of them since it is a special case of them and conversely they are called a including rule of $R(1,1)$ respectively. The same results apply to $R(1,2)$ and applied to those for $D = 2, ..., 6$. Then, all the rule candidates for $D = 1$ were investigated as shown in Table 4 which shows $CndCP$ to distinguish the CP in Table 3. Table 4 shows the following:

(4-1) The rules $CndCP(1,1), ..., CndCP(1,6)$ with $RL = 2$ including $R(1,1)$ or $R(1,2)$ appear in descending order of z-values, which coincides with the CP in Table 3. They suggest us that a lot of inputs partially coinciding with the pre-specified rules by $Dgr = 2$ were transformed into the output by the use of their rules and PCH.

(4-2) The $CndCP(1,7), ..., CndCP(1,21)$ with $RL = 1$ including $R(1,1)$ or $R(1,2)$, or those straddling both rules with RL =2 appear in descending order of z-values. For example, the candidate $CndCP(1,10)$ with $RL = 2$ straddles both $CndCP(1,8)$ and $CndCP(1,7)$ of the rule including $R(1,1)$ and $R(1,2)$ respectively. They also suggest the same as that applied to (4-1) by $Dgr = 1$.

(4-3) All $CndCP(1,7), ..., CndCP(1,21)$ in Table 4 were arranged in Table 3, which was conducted at $LN = 12$ in Fig. 2. For example, $CndCP(1,10)$ is a partial rule of $CndCP(1,7)$ whereas the z-value of $CndCP(1,7)$ is larger than that of $CndCP(1,10)$. Accordingly, the previous STRIM made $CndCP(1,7)$ represent $CndCP(1,10)$ based on the index of z. In the same way, $CndCP(1,7)$ was represented by $CndCP(1,3)$. In this way, the previous STRIM arranged the rule candidates with inclusion relationships by their z-values.

The pre-specified rules $R(1,1)$ and $R(1,2)$ did not appear even as rule candidates respectively in Table 3 since each of them did not satisfy the testing condition at $LN = 24$. The following is a summary of the simulation studies using the previous STRIM for the PCH dataset:

(1) The previous STRIM can't induce the pre-specified rules with longer rule lengths since the datasets partially corresponding with those rules will cause increased growth, and overwhelmingly covers those completely corresponding with them which is the PCH effects. As the result, it induces a lot of rules including the pre-specified rules.

(2) In the case when N is not so large and the rule length of the pre-specified rules is long, the previous STRIM can't adopt them even as a rule candidate.

Table 2. An example of pre-specified rules in the rule box.

$R(d, k)$	$CP(d, k)$	$D = d$
$R(1, 1)$	110000	$D = 1$
$R(1, 2)$	001100	$D = 1$
$R(2, 1)$	220000	$D = 2$
$R(2, 2)$	002200	$D = 2$
...	...	...
$R(6, 1)$	660000	$D = 6$
$R(6, 2)$	006600	$D = 6$

```
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 testing
         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. An algorithm for STRIM including a reduct function.

4 Improved Algorithm Taking PCH into Account

The PCH effects derive a lot of including rules of the pre-specified rules as shown in (4-1) and (4-2) if the previous algorithm of STRIM is applied to the PCH dataset. In this section, we improve the algorithm based on the considerations

Table 3. Examples of finally induced rule using previous STRIM for the PCH dataset.

$CP(d,k)$	$C(1)C(2)$ ...$C(6)$	D	p-value(z)	Accuracy	Coverage	$f = (n_1, n_2, ..., n_6)$
$CP(1,1)$	010010	1	1.19E−123(23.62)	0.857	0.166	(138, 3, 4, 4, 6, 6)
$CP(1,2)$	000101	1	2.42E−120(22.30)	0.899	0.150	(125, 0, 3, 3, 3, 5)
$CP(1,3)$	100010	1	2.38E−97(20.90)	0.826	0.137	(114, 5, 5, 7, 1, 6)
$CP(1,4)$	001100	1	3.27E−90(20.11)	0.861	0.119	(99, 2, 2, 2, 3, 7)
$CP(1,5)$	001001	1	9.22E−84(19.36)	0.835	0.115	(96, 3, 7, 3, 3, 3)
$CP(1,6)$	110000	1	5.60E−78(18.66)	0.780	0.119	(99, 4, 4, 6, 7, 7)
$CP(2,1)$	002200	2	4.43E−130(24.24)	0.849	0.175	(8, 141, 1, 6, 6, 4)
$CP(2,2)$	000202	2	7.58E−111(22.33)	0.883	0.140	(2, 113, 5, 3, 4, 1)
...	...	...	...	...	...	...

obtained by the simulation experiment in Sect. 3. Figure 3 especially shows their relationships for the including rules of $D = 1$. For example, "110000(6)" denotes $CndCP(6)$ in Table 4. The solid line connects each other with one Hamming distance ($HD = 1$) which is considered to be the closest and solidest relationship since rule candidates derived from the pre-specified rules by the PCH effects as shown in (4-1) and (4-2). For example, one of the methods to estimate $R(d,1)$ or $R(d,2)$ is to make the groups of candidates connected to each other with $HD = 1$ in Table 4 and to make each group indicate the pre-specified rules for each $D = d$ as follows:

(Step1) Truncate Table 4 in descending order of z-value until the candidate with $RL = 1$ having the least z-value.
(Step2) Make the Hamming matrix (HM) having the (i,j) element of the HD between $CndCP(d,i)$ and $CndCP(d,j)$ by use of the truncated table. The HM is symmetric.
(Step3) Make the groups with $HD = 1$ by using the HM and a one-stroke sketch, and estimate the pre-specified rules.

In the case of $D = 1$, the last term of Table 4 to be truncated in (Step1) is $CndCP(1,14)$ and the HM obtained in (Step2) is Table 5 showing $HM(i,j)$ $(i,j = 1,...,14)$. For example, the $HM(1,2)$ $(= HM(2,1))$ is the HD between $CndCP(1,1) = 010010$ and $CndCP(1,2) = 000101$ and is found to be $HD = 4$. The following is the specific procedures of (Step3) by the use of Table 5:

(1) Find the i-th element in Table 4 corresponding with $CP(d = 1,k)$ in Table 3 and the least j-th with $HM(i,j) = 1$. Reserve the i for the starting point $i0$.
(2) Reset $HM(i,j) = 0$ and $HM(j,i) = 0$ to prevent a loop.
(3) Substitute i with j.
(4) If $i = i0$ then go to (6) else go to (5).
(5) Find the least j-th element with $HM(i,j) = 1$ if there are and go to (2), else go to (6).

(6) If $i = i0$ is satisfied then construct the pre-specified rule by use of the above sequence candidates else discard the sequence.

For example, execute procedure (1) by $k = 1$ in Table 3 then $i = 1$ is found in Table 4, $(i, j) = (1, 7)$ is obtained and $i0 = 1$ is the starting point in Table 5 since $(1, 7)$ is the least j satisfying $HM(1, j) = 1$. Execute the procedures (2)–(5) and then the sequence of $H(i, j)$ is $(i, j) = (1, 7) \rightarrow (7, 3) \rightarrow (3, 11) \rightarrow (11, 6) \rightarrow (6, 13) \rightarrow (13, 1)$ and $i = i0$ is satisfied. The sequence is proved to be the one-stroke sketch of the rule candidates with $HD = 1$ of $R(1, 1)$ (trace the sequence in Fig. 3) and then $R(1, 1)$ is reconstructed. In the same way, for $k = 2$, the sequence satisfying $i = i0$: $(i, j) = (2, 8) \rightarrow (8, 5) \rightarrow (5, 14) \rightarrow (14, 4) \rightarrow (4, 9) \rightarrow (9, 2)$ is obtained and is proved to be that of $R(1, 2)$ (see Fig. 3). The $k = 3$ in Table 3 derives $R(1, 1)$. All of the k in Table 3 derives $R(1, 1)$ and $R(1, 2)$ by three respectively. The same applied to $D = 2, ..., 6$.

Table 4. Rule candidates for $D = 1$ induced by the previous STRIM for the PCH dataset.

$CndCP(d, k)$	$C(1)C(2)...C(6)$	D	p-value(z)
$CndCP(1, 1)$	010010	1	1.19E−123(23.62)
$CndCP(1, 2)$	000101	1	2.42E−120(23.30)
$CndCP(1, 3)$	100010	1	2.38E−97(20.91)
$CndCP(1, 4)$	001100	1	3.27E−90(20.10)
$CndCP(1, 5)$	001001	1	9.22E−84(19.36)
$CndCP(1, 6)$	110000	1	5.60E−78(18.66)
$CndCP(1, 7)$	000010	1	5.21E−70(17.65)
$CndCP(1, 8)$	000001	1	4.19E−68(17.40)
$CndCP(1, 9)$	000100	1	3.52E−55(15.60)
$CndCP(1, 10)$	000011	1	1.83E−54(15.50)
$CndCP(1, 11)$	100000	1	5.83E−54(15.42)
$CndCP(1, 12)$	001010	1	1.52E−51(15.06)
$CndCP(1, 13)$	010000	1	2.57E−50(14.87)
$CndCP(1, 14)$	001000	1	5.11E−47(14.35)
$CndCP(1, 15)$	100001	1	2.15E−44(13.93)
$CndCP(1, 16)$	000110	1	1.33E−42(13.63)
$CndCP(1, 17)$	010100	1	2.39E−41(13.41)
$CndCP(1, 18)$	100100	1	2.31E−37(12.72)
$CndCP(1, 19)$	011000	1	1.82E−36(12.56)
$CndCP(1, 20)$	101000	1	2.08E−34(12.18)
$CndCP(1, 21)$	010001	1	3.84E−34(12.13)

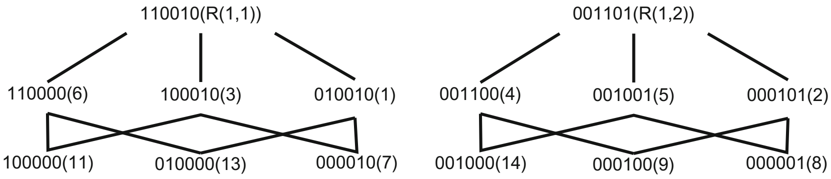

Fig. 3. Derived rules from the pre-specified rules for $D = 1$ with one Hamming distance.

Adding to an algorithm implementing the above procedure under $LN = 12$ in Fig. 2, STRIM can adapt the PCH dataset and results in a new algorithm we call STRIM2.

Table 5. Examples of Hamming distance against rule candidates for $D = 1$.

$HM(i,j)$	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]
[1]	0	4	2	4	4	2	1	3	3	2	3	2	1	3
[2]	4	0	4	2	2	4	3	1	1	2	3	4	3	3
[3]	2	4	0	4	4	2	1	3	3	2	1	2	3	3
[4]	4	2	4	0	2	4	3	3	1	4	3	2	3	1
[5]	4	2	4	2	0	4	3	1	3	2	3	2	3	1
[6]	2	4	2	4	4	0	3	3	3	4	1	4	1	3
[7]	1	3	1	3	3	3	0	2	2	1	2	1	2	2
[8]	3	1	3	3	1	3	2	0	2	1	2	3	2	2
[9]	3	1	3	1	3	3	2	2	0	3	2	3	2	2
[10]	2	2	2	4	2	4	1	1	3	0	3	2	3	3
[11]	3	3	1	3	3	1	2	2	2	3	0	3	2	2
[12]	2	4	2	2	2	4	1	3	3	2	3	0	3	1
[13]	1	3	3	3	3	1	2	2	2	3	2	3	0	2
[14]	3	3	3	1	1	3	2	2	2	3	2	1	2	0

5 Another Type of Pre-specified Rule

In order to confirm the availability of the algorithm studied in Sect. 4, let us study it by modifying the rules in Table 2 like $R(d,2) = 00dd0d \rightarrow R(d,2) = 0d0d0d$ ($d = 1, ..., 6$). Having the same condition attribute value like $C(2) = d$ in $R(d,1)$ and $R(d,2)$ is the feature of the modified rules. Generating the PCH dataset based on the modified rules in Fig. 1, and applying STRIM2 to the dataset, Table 6 for $D = 1$ was obtained by arranging the interim results. Table 6 contains the set of $CndCP(1,k)$ which is ordered in descending order of the z-value and truncated at the least z-value of the candidate with $RL = 1$ corresponding to the front side of Table 5, and the HM which corresponds to Table 5 and was constructed by the set of $CndCP(1,k)$. Here, three $CndCP(1,k)$ $(k = 1, 5, 7)$ with an "$*$" are the candidates corresponding to $CP(1,k)$ in Table 3.

In the same way as Table 5, STRIM2 induced the rules from Table 6 as follows: By use of $CndCP(1,1) = 010000(*1)$, the sequence: $(1,2) \rightarrow (2,9) \rightarrow (9,5) \rightarrow (5,8) \rightarrow (8,3) \rightarrow (3,1)$ induced $010101 = R(1,2)$ although $CndCP(1,1)$ is also the including rule of $R(1,1)$. In the same way, $CndCP(1,5) = 000101(*2)$ derived the sequence: $(5,8) \rightarrow (8,3) \rightarrow (3,1) \rightarrow (1,2) \rightarrow (2,9) \rightarrow$(9,5) and induced $010101 = R(1,2)$. However, $CndCP(1,7) = 100010(*3)$ derived the sequence: $(7,10) \rightarrow (10,4) \rightarrow (4,1) \rightarrow (1,2) \rightarrow (2,9) \rightarrow (9,5) \rightarrow (5,8) \rightarrow (8,3) \rightarrow (3,1) \rightarrow (1,6) \rightarrow (6,12) \rightarrow (12,7)$ and induced "110111," which was the compound of $R(1,1)$ and $R(1,2)$. Inspecting the sequence in detail, it started from $CndCP(1,7)$ of the including rule of $R(1,1)$, and on the way changed that of $R(1,1)$ into that of $R(1,2)$ like $(1,2) \rightarrow (2,9)$ and again changed into that of $R(1,1)$. That is why STRIM2 induced the compound rule. It should be noted that the case when STRIM2 cannot induce the pre-specified rules but the compound rules may happen in the case when they have more than two $CP(d,k)$ $(k = 1, 2, ...)$ and the same condition attribute value like $C(2) = d$ for the same decision attribute value, and/or their including rules are not separated from each other (see Fig. 3).

Table 6. Rule candidates and Hamming distance induced by STRIM2 for the dataset generated by the rules modifying Table 2.

$CndCP(1,k)$:	HM											
$(1,1)$: 010000(*1)	0	1	1	1	3	1	3	2	2	2	2	2
$(1,2)$: 010001	1	0	2	2	2	2	4	3	1	3	1	3
$(1,3)$: 010100	1	2	0	2	2	2	4	1	3	3	1	3
$(1,4)$: 110000	1	2	2	0	4	2	2	3	3	1	3	3
$(1,5)$: 000101(*2)	3	2	2	4	0	4	4	1	1	3	1	3
$(1,6)$: 010010	1	2	2	2	4	0	2	3	3	3	3	1
$(1,7)$: 100010(*3)	3	4	4	2	4	2	0	3	3	1	5	1
$(1,8)$: 000100	2	3	1	3	1	3	3	0	2	2	2	2
$(1,9)$: 000001	2	1	3	3	1	3	3	2	0	2	2	2
$(1,10)$: 100000	2	3	3	1	3	3	1	2	2	0	4	2
$(1,11)$: 010101	2	1	1	3	1	3	5	2	2	4	0	4
$(1,12)$: 000010	2	3	3	3	3	1	1	2	2	2	4	0

6 Application of STRIM2 to a Real-World Dataset

The Rakuten Institute of Technology provides an open dataset of Rakuten Travel [14]. This dataset contains about $6,200,000$ questionnaire survey ratings $A = \{C(1) = \text{Location}, C(2) = \text{Room}, C(3) = \text{Meal}, C(4) = \text{Bath (Hot Spring)}, C(5) = \text{Service}, C(6){=}\text{Amenity}, D = \text{Overall}\}$ for about $130,000$ travel facilities using a set of categorical values $V_a = \{$Dissatisfied (1), Somewhat dissatisfied (2), Neither satisfied nor dissatisfied (3), Satisfied (4), Very Satisfied (5)$\}$, $\forall a \in A$, that is, $|V_{a=D}| = |V_{a=C(j)}| = 5$. We constructed a decision table of $N = 10,000$ surveys by randomly selecting $2,000$ samples, each with $D = m$ $(m = 1, ..., 5)$, from about $400,000$ surveys of the 2013–2014 dataset because there were heavy biases with respect to the frequency of $D = m$. Finally, we randomly sampled $N_B = 5,000$ from the $10,000$ surveys and re-constructed the decision table.

We applied STRIM2 to the decision table and Table 7 shows the interim results corresponding to Table 3. The HM corresponding to Table 5 or Table 6 is omitted since its size is so large, for example, 62×62 for $D = 1$. Table 8 shows the final results by STRIM2 obtained in the same procedures as the simulation experiments in Sects. 4 and 5. Here, the results are shown as $CP2(d, k)$ to distinguish the final from the interim. Although the Rakuten Travel dataset is not clear whether it obeys PCH or not, and no one knows the original rules since it is not a simulation experiment, Table 8 suggests the following based on the results obtained from the simulation experiments:

(1) For $D = 1$, both of $CP(1, 1)$ and $CP(1, 2)$ with $RL = 1$ induced the same rule $CP2(1, 1)$ with $RL = 3$ respectively. That is, STRIM2 induced the partial rule $CP2(1, 1)$ of $CP(1, 1)$ and $CP(1, 2)$ which represented $CP2(1, 1)$ by use of the previous STRIM and moreover found another factor $C(6) = 1$ affecting $D = 1$. The result seems not to be so strange.

(2) For $D = 2$, STRIM2 induced the same rule as $CP(2, 1)$ of which accuracy is not so high to compare with the other rules. The frequency distribution of $CP(2, 1)$ spreads widely from $D = 1$ to $D = 3$, which seemed to be caused by the hard decision of "Somewhat dissatisfied." Accordingly, it is supposed that the original rule of $D = 2$ could not make the one-strike sketch by the including rules with $RL = 1$

(3) STRIM2 induced $CP2(3, 1)$ with $RL = 4$ from $CP(3, 1)$ with $RL = 2$ and $CP2(4, 1)$ with $RL = 3$ from $CP(4, 1)$ with $RL = 2$, which seems not to be so strange taking the simulation studies into account.

(4) STRIM2 induced $CP2(5, 1)$ with $RL = 3$ from $CP(5, 1)$ with $RL = 1$ and $CP2(5, 2)$ with $RL = 4$ from $CP(5, 2)$ with $RL = 2$, and the former rule includes the latter, which remind us of the studies in Sect. 5. However, STRIM2 suggested that the factors: $C(2) = 5$, $C(3) = 5$, $C(5) = 5$, $C(6) = 5$ have an important effect on $D = 5$ while the previous STRIM indicates only the partial effect.

Table 7. Induced interim rules from Rakuten Travel dataset by STRIM2.

$CP(d,k)$ by STRIM	$C(1)C(2)$ ...C(6)	D	p-value(z)	Accuracy	Coverage	$f=(n_1,n_2,...,n_6)$
$CP(1,1)$	000010	1	0.00(40.50)	0.761	0.639	(654,187,16,1,1)
$CP(1,2)$	010000	1	4.01E−236(32.79)	0.683	0.509	(521,200,39,3,0)
$CP(2,1)$	020000	2	4.44E−79(18.79)	0.488	0.335	(160,339,169,29,4)
$CP(3,1)$	030030	3	2.47E−165(27.38)	0.634	0.390	(31,97,373,82,5)
$CP(4,1)$	040040	4	1.50E−184(28.95)	0.725	0.351	(7,16,47,350,63)
$CP(5,1)$	000050	5	0.00(44.94)	0.758	0.790	(17,21,31,186,800)
$CP(5,2)$	055000	5	0.00(43.36)	0.874	0.580	(11,12,5,57,588)

Table 8. Induced final rules from Rakuten Travel dataset by STRIM2.

$CP2(d,k)$ by STRIM2	$C(1)C(2)$ ...$C(6)$	D	p-value(z)	Accuracy	Coverage	$f=(n_1,n_2,...,n_6)$
$CP2(1,1)$	010011	1	8.14E−185(28.97)	0.940	0.231	(236,15,0,0,0)
$CP2(2,1)$	020000	2	4.44E−79(18.79)	0.488	0.335	(160,339,163,29,4)
$CP2(3,1)$	033033	3	3.26E−135(24.72)	0.811	0.207	(8,15,198,23,0)
$CP2(4,1)$	040044	4	4.97E−162(27.10)	0.796	0.262	(4,8,18,261,37)
$CP2(5,1)$	055050	5	0.00(43.24)	0.939	0.515	(3,4,0,27,522)
$CP2(5,2)$	055055	5	0.00(40.20)	0.977	0.419	(2,2,0,6,424)

7 Conclusion

This paper experimentally studied an algorithm to adapt PCH datasets and improved the previous STRIM. Specifically, this paper focused on rule candidates derived by the STRIM, proposed a method to group them by the solid relationship of a one-stroke sketch having one Hamming distance ($HD = 1$) and made the groups estimate the pre-specified rules. STRIM incorporating this function was named STRIM2 which clarified its performance and cautions for use by applying it in two typical simulation experiments. STRIM2 was applied to a real-world dataset, that is, Rakuten Travel dataset and the induced rules were considered from the view of those studied by the simulation so that the results were roughly shown to be valid and were full of interesting suggestions although no one knew the pre-specified rules and the domain-knowledge was needed for the review.

References

1. 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
2. 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
3. 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
4. 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)
5. 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
6. 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
7. 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
8. 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, part I. LNCS (LNAI), vol. 10313, pp. 418–431. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-60837-2_35
9. Pawlak, Z.: Rough sets. Int. J. Inform. Comput. Sci. **11**(5), 341–356 (1982)
10. 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, vol. 11, pp. 3–18. Springer, Dordrecht (1992). https://doi.org/10.1007/978-94-015-7975-9_1
11. Ziarko, W.: Variable precision rough set model. J. Comput. Syst. Sci. **46**, 39–59 (1993)
12. Laboratory of Intelligent Decision Support System (IDSS). http://idss.cs.put.poznan.pl/site/139.html
13. 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)
14. http://rit.rakuten.co.jp/opendataj.html