# New Rule Induction Method by Use of a Co-occurrence Set from the Decision Table

Yuichi Kato[1(✉)] and Tetsuro Saeki[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 as an if-then rule induction method from the decision table (DT) and has improved those methods by the conventional Rough Sets from a statistical view. The method recognizes condition attributes (CA) and the decision attribute (DA) in DT as random variables having the causality of an input-output relation, and uses the relation of transforming the inputs (outcomes of CA) into the outputs (those DA) through the rules for rule induction strategies. This paper reconsiders the conventional STRIM, proposes a new rule induction method and strategy named apriori-STRIM and confirms the validity and capacity by a simulation experiment. Specifically, the new method explores CA of causes after receiving outcomes of DA by use of co-occurrence sets of outcomes of CA. The co-occurrence set is a well-known concept in the association rule learning (ARL) field. This paper also clarifies the differences of rule induction methods and their capacities between apriori-STRIM and ARL by the same experiments.

**Keywords:** Rough Sets · If-then rule induction · apriori-STRIM · Simulation experiment

## 1 Introduction

The Rough Set (RS) theory was introduced by Pawlak [1] and used for inducing if-then rules from a dataset called the decision table (DT). To date, various methods and algorithms for inducing rules by the theory have been proposed [2–5] since the inducing rules are useful to simply and clearly express the structure of rating and/or knowledge hiding behind the table. The basic idea to induce rules is to approximate the concept in the DT by use of the lower and/or upper approximation sets which are respectively derived from the equivalence relations and their equivalence sets in the given DT. However, those methods and algorithms by RS paid little attention to the fact that the DT was just a sample set


V. Gutiérrez-Basulto et al. (Eds.): RuleML+RR 2020, LNCS 12173, pp. 54–69, 2020.
https://doi.org/10.1007/978-3-030-57977-7_4

gathered from the population of interest. If resampling the DT from the population or the DT by Bootstrap method for example, the new DT will change equivalence relations, their equivalence sets, and the lower and/or upper approximation sets, so the induced rules will change and fluctuate. Those methods and algorithms also had the problem that those induced rules were not arranged from the statistical views. Then, we proposed a rule induction method named STRIM (Statistical Test Rule Induction Method) taking the above mentioned problems into consideration [6–16]. Specifically, STRIM

(1) Proposed a data generation model for generating a DT. This model recognized the DT as an input-output system which transformed a tuple of the condition attribute's value occurred by chance (the input) into the decision attribute value (the output) through pre-specified if-then rules (generally unknown) under some hypotheses. That is, the input was recognized as an outcome of the random variables and the output was also the outcome of a random variable dependent on the input and the pre-specified rules. Accordingly, the pairs of input and output formed the DT containing rules.
(2) Assumed a trying proper condition part of if-then rules and judged whether it was a candidate of rules by statistically testing whether the condition caused bias in the distribution of the decision attribute's values.
(3) Arranged the candidates having inclusion relationships by representing them with one of the highest bias and finally induced if-then rules with a statistical significance level after systematically exploring the trying condition part of rules. The validity and capacity of STRIM have been confirmed by the simulation experiments that STRIM can induce pre-specified if-then rules from the DT proposed in (1). In this way, the conventional data generation model proposed in (1) also can be used for a verification system of a rule induction method (VSofRIM). The validity and capacity also secure a certain extent of the confidence of rules induced by STRIM from the DT of real-world datasets. The VSofRIM is also used for confirming the validity and capacity of other rule induction methods proposed previously [11,14].

The conventional STRIM systematically explores the domain of the condition attributes looking for rule candidates causing the bias and statistically judges their validity by use of the DT which is accumulated by rules intervening between the inputs of the condition attributes and the corresponding outputs of the decision attribute. This paper reconsiders the process after (2) from the view of Bayes's law which generally infers the causes from the results, and proposes a new rule induction method named apriori-STRIM. Specifically, the method explores a co-occurrence set of the condition attribute's value in the DT against a specific decision attribute's value. The concept of the co-occurrence set plays an important role in the association rule learning (ARL) field [17] and the set can be effectively found using the well-known Apriori algorithm [18]. That is, apriori-STRIM focuses on the property that the specific decision attribute's value will occur with the specific condition attribute values by the rules' intervention, although the conventional STRIM focuses on the bias. The rules' intervention also can be judged by a statistical test using the co-occurrence set in the DT.

The validity and capacity of apriori-STRIM is also confirmed by the same experiment as the conventional and the two-way confirmations by both STRIMs secure the validity and capacity for the rule induction method. This paper also shows interesting features of ARL by applying it to VSofRIM and clarifies the differences between apriori-STRIM and ARL.

## 2 Conventional Rough Sets and STRIM

The Rough Set theory is used for inducing if-then rules from a decision table $S$. $S$ is conventionally denoted by $S = (U, A = C \cup \{D\}, V, \rho)$. Here, $U = \{u(i)|i = 1, ..., |U| = N\}$ is a sample set, $A$ is an attribute set, $C = \{C(j)|j = 1, ..., |C|\}$ is a condition attribute set, $C(j)$ is a member of $C$ and a condition attribute, and $D$ is a decision attribute. Moreover, $V$ is a set of attribute values denoted by $V = \cup_{a \in A} V_a$ and is characterized by the information function $\rho$: $U \times A \rightarrow V$.

The conventional Rough Set theory first focuses on the following equivalence relation and the equivalence set of indiscernibility within the decision table $S$ of interest:

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

$I_B$ is an equivalence relation in $U$ and derives the quotient set $U/I_B = \{[u_i]_B | i = 1, 2, ..., |U| = N\}$. Here, $[u_i]_B = \{u(j) \in U | (u(j), u_i) \in I_B, u_i \in U\}$, $[u_i]_B$ is an equivalence set with the representative element $u_i$.

Let $X$ be an arbitrary subset of $U$ then $X$ can be approximated as $B_*(X) \subseteq X \subseteq B^*(X)$ through the use of the equivalence set. Here, $B_*(X) = \{u_i \in U | [u_i]_B \subseteq X\}$, and $B^*(X) = \{u_i \in U | [u_i]_B \cap X \neq \phi\}$, $B_*(X)$ and $B^*(X)$ are referred to as the lower and upper approximations of $X$ by $B$ respectively. The pair of $(B_*(X), B^*(X))$ is usually called a rough set of $X$ by $B$.

Specifically, let be $X = \{u(i) | \rho(u(i), D) = d\} = U(d) = \{u(i) | u^{D=d}(i)\}$ called the concept of $D = d$, and define a set of $u(i)$ as $U(CP) = \{u(i) | u^{C=CP}(i)$, meaning $CP$ satisfies $u^C(i)$, where $u^C(i)$ is the tuple of the condition attribute values of $u(i)\}$ and let it be equal to $B_*(X)$, then $CP$ can be used as the condition part of the if-then rule of $D = d$, with necessity. That is, the following expression of if-then rules with necessity is obtained: if $CP = \wedge_j (C(j_k) = v_{j_k})$ then $D = d$. In the same way, $B^*(X)$ derives the condition part $CP$ of the if-then rule of $D = d$ with possibility.

However, the approximation of $X = U(d)$ by the lower or upper approximation is respectively too strict or loose so that the rules induced by the approximations are often of no use. Then, Ziarko expanded the original RS by introducing an admissible error in two ways [4]: $\underline{B}_\epsilon(U(d)) = \{u(i) | accuracy \geq 1 - \varepsilon\}$, $\overline{B}_\varepsilon(U(d)) = \{u(i) | accuracy > \varepsilon\}$, where $\varepsilon \in [0, 0.5)$. The pair of $(\underline{B}_\epsilon(U(d)), \overline{B}_\varepsilon(U(d)))$ is called an $\varepsilon$-lower and $\varepsilon$-upper approximation which satisfies the following 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 $\varepsilon$-lower and/or $\varepsilon$-upper approximation induce if-then rules with admissible errors in the same way as the lower and/or upper approximation.

As mentioned above, the conventional RS theory basically focuses on the equivalence relation $I_B$ and its equivalence sets $U/I_B$ in $U$ given in advance and induces rules approximating the concept by use of the approximation sets derived from the $U/I_B$. However, $I_B$ is very dependent on the DT provided. Accordingly, every DT obtained from the same population is different from each other and, $I_B$, $U/I_B$ and the approximation sets are different from each other for each DT, which leads to inducing different rule sets. That is, the rule induction methods by the conventional RS theory lack statistical views.

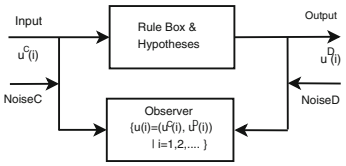


**Fig. 1.** A data generation model: Rule box contains if-then rules $R(d,k)$: if $sCP(d,k)$ then $D=d$ $(d=1,2,...,k=1,2,...)$.

**Table 1.** Hypotheses with regard to the decision attribute value.

| | |
|---|---|
| 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 $R(d,k)$ $(d=d1,d2,\ldots)$, 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) |

Then, STRIM [6,9,10,12,15] has proposed a data generation model (DGM) for the DT and a rule induction method based on the model. Specifically, STRIM considers the decision table to be a sample dataset obtained from an input-output system including 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 its decision attribute value $u^D(i)$. $u^C(i)$ is the input for the rule box, and is transformed into 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 input. They are uniquely determined, indifferent and conflicted cases (see Table 1). In contrast, $u(i)=(u^C(i),u^D(i))$ is measured by an observer, as shown in Fig. 1. The existence of NoiseC and NoiseD makes missing values in $u^C(i)$, and changes $u^D(i)$ to create another value for $u^D(i)$, respectively. Those noises bring the system

closer to a real-world system. The data generation model suggests that a pair of $(u^C(i), u^D(i))$, $(i = 1, \ldots, N)$, i.e. a decision table is an outcome of these random variables: $(C, D) = ((C(1), \ldots, C(|C|), D)$ observing the population.

Based on the data generation model, STRIM (1) extracted significant pairs of a condition attribute and its value like $C(j_k) = v_{j_k}$ for rules of $D = d$ by the local reduct [10,12,13], (2) constructed a tentatively trying condition part of the rules like $CP = \wedge_j (C(j_k) = v_{j_k})$ by use of the reduct results, and (3) investigated whether $U(CP)$ caused a bias at $n_d$ in the frequency distribution of the decision attribute values $f = (n_1, n_2, \ldots, n_{M_D})$ or not, where $n_m = |U(CP) \cap U(m)|$ $(m = 1, \ldots, |V_{a=D}| = M_D)$ and $U(m) = \{u(i)|u^{D=m}(i)\}$, since the $u^C(i)$ coinciding to $sCP(d, k)$ in the rule box is transformed into $u^D(i)$ based on Hypotheses 1 or 3. Accordingly, the $CP$ coinciding to one of rules in the rule box produces bias in $f$. Specifically, STRIM used a statistical test method for the investigation specifying a null hypothesis $H0$: $f$ does not have any bias, that is, $CP$ is not a rule and its alternative hypothesis $H1$: $f$ has a bias, that is, $CP$ is a rule, and a proper significance level, and tested $H0$ by use of the sample dataset, that is, the decision table and the proper test statistics, for example,

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

where $n_d = \max_m f = (n_1, \ldots, n_m, \ldots, n_{M_D})$, $p_d = P(D = d)$, $n = \sum_{j=1}^{M_D} n_j$. $z$ obeys the standard normal distribution under test conditions: $np_d \geq 5$ and $n(1 - p_d) \geq 5$ [19] and is considered to be an index of the bias of $f$. (4) If $H0$ is rejected then the assumed $CP$ becomes a candidate for the rules in the rule box. (5) After repeating the processes from (1) to (4) and obtaining the set of rule candidates, STRIM arranged their rule candidates and induced the final results (see literatures [12,13] for details).

To summarize, STRIM directly induces rules with statistical significance level assuming the condition part of rules: $CP = \wedge_j (C(j_k) = v_{j_k})$ and statistically testing it by use of $U$. STRIM does not require the basic concept of the approximation which is the point for the rule induction by RS theory. Conversely, RS theory has nothing directly to do with statistical significance.

## 3 Studies on STRIM by Simulation Experiment

We implemented the data generation process and verified the capacity of inducing the rules by the conventional STRIM as follows: (1) Specified the rules in Table 2 (the number of rules $(N_{rule}) = 10$) in the rule box in Fig. 1, where $|C| = 6$, $V_a = \{1, 2, \ldots, 5\}$ $(a = C(j)$, $(j = 1, \ldots, |C|)$, $a = D)$, and $sCP(1, 1) = 110000$ denoted $sCP(1, 1) = (C(1) = 1) \wedge (C(2) = 1)$ and was called a rule of the rule length 2 $(RL = 2)$, having two conditions. (2) Generated $v_{C(j)}(i)$ $(j = 1, \ldots, |C| = 6)$ with a uniform distribution and formed $u^C(i) = (v_{C(1)}(i), \ldots, v_{C(6)}(i))$ $(i = 1, \ldots, N = 10{,}000)$. (3) Transformed $u^C(i)$ into $u^D(i)$ using the pre-specified rules in Table 2 and hypotheses in Table 1,

**Table 2.** An example of pre-specified rules $R(d,k)$ in the rule box: if $sCP(d,k)$ then $D=d$ $(d=1,\ldots,5,\ k=1,2)$.

| $R(d,k)$ | $sCP(d,k)$ | $D=d$ |
|---|---|---|
| $R(1,1)$ | 110000 | $D=1$ |
| $R(1,2)$ | 001100 | $D=1$ |
| $R(2,1)$ | 220000 | $D=2$ |
| $R(2,2)$ | 002200 | $D=2$ |
| $R(3,1)$ | 330000 | $D=3$ |
| $R(3,2)$ | 003300 | $D=3$ |
| $R(4,1)$ | 440000 | $D=4$ |
| $R(4,2)$ | 004400 | $D=4$ |
| $R(5,1)$ | 440000 | $D=5$ |
| $R(5,2)$ | 004400 | $D=5$ |

**Table 3.** An example of estimated rules by the conventional STRIM for the DT with $N_B$ = 5,000 generated by the data generation model in Fig. 1 and the pre-specified rules in Table 2.

| Rule no. | Estimated rules $(C(1)\ldots C(6)D)$ | $f=(n_1,n_2,n_3,n_4,n_5)$ | $p$-value | Accuracy | Coverage |
|---|---|---|---|---|---|
| 1 | (0022002) | $(4,216,8,5,4)$ | 5.87E−173 | 0.911 | 0.223 |
| 2 | (0011001) | $(207,3,4,2,3)$ | 8.26E−162 | 0.945 | 0.200 |
| 3 | (0055005) | $(8,4,7,5,211)$ | 1.87E−159 | 0.898 | 0.212 |
| 4 | (4400004) | $(5,6,5,187,4)$ | 5.73E−150 | 0.903 | 0.195 |
| 5 | (1100001) | $(190,1,6,3,4)$ | 1.86E−145 | 0.931 | 0.184 |
| 6 | (5500005) | $(5,8,6,5,191)$ | 2.73E−142 | 0.888 | 0.192 |
| 7 | (0044004) | $(4,3,3,167,3)$ | 8.99E−140 | 0.928 | 0.174 |
| 8 | (3300003) | $(5,6,193,7,3)$ | 1.98E−139 | 0.902 | 0.186 |
| 9 | (2200002) | $(3,167,6,1,5)$ | 7.37E−136 | 0.918 | 0.172 |
| 10 | (0033003) | $(3,4,185,10,2)$ | 6.03E−135 | 0.907 | 0.178 |

without generating NoiseC and NoiseD for a plain experiment and then generated the decision table. After randomly selecting samples by $N_B$ = 5,000 samples, newly forming the DT and applying STRIM to the DT, Table 3 was obtained. The table shows us the following: For example, the estimated Rule No. 1 $(RN=1)$ "0022002" denotes if $(C(3)=2)\wedge(C(4)=2)$ then $D=2$, has $f=(n_1,n_2,\ldots,n_5)=(4,216,8,5,4)$ and the bias at $D=2$. The outcome probability to cause such a bias is around 5.87E-173 under $H0$, which leads to rejecting $H0$ and adopting $H1$. As the result, "0022002" was adopted as a rule. It should be noted that the reason it was adopted as the rule was not the high $accuracy = 216/237 = 0.911$. STRIM just induced all the pre-specified rules in

Table 2. This experiment suggests that conventional STRIM works well and the DGM in Fig. 1 can be useful as a verification system of a rule induction method.

## 4 New Rule Induction Method by Co-occurrence Set

As mentioned in Sect. 2, conventional STRIM regards the condition attributes $C$ and the decision attribute $D$ as random variables, and $D$ of the output depends on $C$ of the input, rules and hypotheses and the rule induction method focuses on $P(D = d|CP)$ of $P(CP, D = d) = P(CP)P(D = d|CP)$. That is, STRIM regards $P(D = d|CP)$ as $P($ if $CP$ then $D = d)$ and explores $CP = \wedge_j(C(j_k) = v_{j_k})$ which causes bias at $n_d$ in $f = (n_1, n_2, \ldots, n_{M_D})$. The bias can be detected by use of the DT and the statistical test.

```
Line No. Algorithm to induce if-then rules by STRIM with apriori function
   1     install.packages("arules") # import package arules
   2     library(arules) # load package arules
   3     input data # input Decision Table
   4       for (iD in 1: M_D) {# proceed co-occurrence item set of iD
   5       dataiD<-data[data[, (length(C) +1)]==iD,] # extract dataset of Decision
         attribute value of iD
   6       dataCiD<-dataiD[1: length(C) ,] # extract its Condition attributes value part
   7       CiD.tra<-transform dataCiD # CiD of transaction form
   8       CiD.ap<-apriori(CiD.tra, parameter=list(support=supp0,
         target='frequent itemset')) # explore freqent item set more than supp0
   9       SFIS<-inspect(CiD.ap) # output the set of frequent item set
  10       for (iCo in 1: nrow(SFIS) ) {# proceed each frequent item set
  11         calculate p-value of SFIS(iCo)
  12         if p-value < p-value0, save the SFIS(iCo) as a rule candidate with necessary
         index
  13       }# end of for of iCo
  14       arrange the rule candidates of iD
  15     }# end of for of iD
```

**Fig. 2.** An algorithm for apriori-STRIM written in R language style.

From the view of Bayes's law, however, another strategy of focusing on $P(CP|D = d)$ of $P(CP, D = d) = P(D = d)P(CP|D = d)$ can be considered for the rule induction. That is, after receiving the outputs of $D = d$, the strategy exploring and estimating $CP = \wedge_j(C(j_k) = v_{j_k})$ of the corresponding inputs can be also valid. Specifically, when receiving the outputs, the corresponding inputs are classified into two cases: One is the uniquely determined and/or conflicted cases and the other is the indifferent case (see Table 1). Both cases can be easily distinguished from each other by use of a statistical test specifying the null hypothesis $H0$: the event $D = d$ has occurred by chance (the indifferent case) and the alternative hypothesis $H1$: the event $D = d$ hasn't occurred by chance (the uniquely and/or conflicted case). Under $H0$, $P(CP|D = d) = P(CP)$ and the intervention of rules transforming the inputs into the output is denied. If $H0$ is denied, $H1$ is adopted as a rule candidate, which means $P(CP|D = d) \neq P(CP)$. Such hypothesis testing can be easily executed by finding the co-occurrence set with the event $D = d$ since the concept of the co-occurrence set is well-known in

**Table 4.** An example of $FIS$ extracted from $U(D = 1)$ of the DT corresponding to Table 3.

| No. of FIS | Items | Support | Count |
|---|---|---|---|
| [1] | $\{15\}$ | 0.138 | 143 |
| [2] | $\{22\}$ | 0.143 | 148 |
| [27] | $\{11\}$ | 0.337 | 348 |
| [31] | $\{15, 22\}$ | 0.032 | 33 |
| [32] | $\{15, 24\}$ | 0.028 | 29 |
| [350] | $\{11, 21\}$ | 0.184 | 190 |
| [351] | $\{31, 41\}$ | 0.200 | 207 |
| [353] | $\{21, 31\}$ | 0.101 | 104 |
| [354] | $\{15, 31, 41\}$ | 0.029 | 30 |
| [387] | $\{31, 52, 63\}$ | 0.024 | 25 |
| [403] | $\{21, 31, 41\}$ | 0.044 | 45 |

the field of association rule learning (ARL) [17] and the finding is effectively executed by the Apriori algorithm [18]. Then, this paper names this rule induction method apriori-STRIM. The ARL and the Apriori algorithm are summarized in Appendix A and the differences of the idea for the rule induction method between ARL and apriori-STRIM are shown through the same experiment in Sect. 3, that is, by VSofRIM. See Appendix A to easily understand the following.

The specific procedure for apriori-STRIM is shown in Fig. 2 where the procedure is shown in R language style since the Apriori algorithm is already implemented by the language as an apriori() function which has a good reputation. The outline is the following: Line No. 1 ($LN = 1$) installs the package of ARL as "arules" [20] via the internet and $LN = 2$ loads it as the library "arules". $LN = 3$ inputs the DT as "data". From $LN = 4$ to $NL = 15$, every $iD$ of the decision attribute ($= 1 \sim M_D$) is proceeded. $LN = 5$ substitutes $U(iD) = \{u(i)|u^{D=iD}(i)\}$ with data$iD$, and its condition part data$CiD$ is extracted ($LN = 6$), and transformed into the transaction form $CiD.tra$ at $LN = 7$. $LN = 8$ extracts co-occurrence sets of $CiD.tra$, that is, co-occurrence sets of the condition attributes' values of $U(iD)$ as frequent item sets according to parameters which specify them as greater than or equal to $supp0$ and substitute them for $CiD.ap$. $LN = 9$ extracts the set of frequent item set ($SFIS$). From $LN = 10$ to $LN = 13$, every $p$-value of $SFIS(iCo)$ ($iCo = 1, \ldots, |SFIS|$) is calculated and tested for whether its $p$-value is less than a pre-specified $p$-value0 and $SFIS(iCo)$ is saved as a candidate if it satisfies the condition. $LN = 14$ arranges the candidates having inclusion relationship by representing the candidate with the least $p$-value and finally decides rules for $iD$.

## 5 Studies on Apriori-STRIM by Simulation Experiment

An experiment result for the rule induction by the conventional STRIM was shown in Table 3 by applying it to the DT containing the pre-specified rules. We also applied apriori-STRIM for the same DT and show the results with the process in Fig. 2.

**Table 5.** An example of rule candidates extracted from Table 4.

| Rule no. | Estimated rules $(C(1)\ldots C(6)D)$ | Count | $p$-value | Accuracy | Coverage |
|---|---|---|---|---|---|
| 1 | 0011001 | 207 | 1.93E−81 | 0.945 | 0.200 |
| 2 | 1100001 | 190 | 1.48E−68 | 0.931 | 0.184 |
| 3 | 0100001 | 372 | 5.78E−33 | 0.360 | 0.360 |
| 4 | 0001001 | 366 | 7.14E−31 | 0.366 | 0.354 |
| 5 | 0010001 | 366 | 7.14E−31 | 0.366 | 0.354 |
| 6 | 0011031 | 57 | 4.44E−29 | 0.966 | 0.055 |
| 7 | 1000001 | 348 | 5.42E−25 | 0.360 | 0.337 |
| 8 | 0511001 | 52 | 7.38E−25 | 0.945 | 0.050 |
| 9 | 0011201 | 52 | 7.38E−25 | 0.963 | 0.050 |

Table 4 shows the part of $SFIS$ obtained at $LN = 8$ and 9 for $iD = 1$, that is, $D = 1$. Here, $supp0 = 5 \cdot |V_a|/|U(iD)|$, $\forall a \in C$ was used for exploring $FIS$. This specification secure $freq(FIS) \geq 5 \cdot |V_a| = count0$ for the hypothesis testing at $LN = 11$ and 12 and induced $|SFIS| = 403$. The table shows: No. of $FIS = 1 - 30$ ($NFIS = 1 - 30$) is $FIS(|items| = 1)$, $NFIS = 31 - 353$ is $FIS(|items| = 2)$ and $NFIS = 354 - 403$ is $FIS(|items| = 3)$, and for example, $NFIS = 387$ indicates that the co-occurrence set of $items = \{C(3) = 1, C(5) = 2, C(6) = 3\}$ that is, the pattern $CP = (C(3) = 1) \wedge (C(5) = 2) \wedge (C(6) = 3)$ occurred $count = 25$ times in $|U(D = 1) = \{u(i)|u^{D=1}(i)\}| = 1,033$. The pre-specified rules for $D = 1$ in Table 2 are $R(1, 1)$ and $R(1, 2)$, which appear in $NFIS = 350$ and 351 respectively.

$LN = 12$ induced significant $FIS$ patterns in Table 4 by the hypothesis testing under $H0$. The frequency $X$ of the co-occurrence pattern $CP$ obeys Binominal distribution $Bn(n, p)$ having the expectation $np$ where $n = |U(D = 1)|$ and $p = \prod_{a \in CP} \frac{1}{|V_a|}$. For example, $p = (\frac{1}{5})^3$ at $NFIS = 387$ due to $RL = 3$. One specification for $supp0$ was to satisfy the requirement $np = Xp \geq 5$ for $RL = 1$ as well as the conventional STRIM (see the test conditions [19] in Sect. 2). That is, $\min X = count0 = \frac{5}{p} = 5 \cdot |V_a|$. $count \geq count0 = 25$ is satisfied in Table 4. As shown in Appendix A, the small $count0$ tends to generate a large number of meaningless $FIS$, and conversely, the large increases the possibility to miss the valid $FIS$s.

$LN = 12$ induced the number of 46 rule candidates from 403 in Table 4 using $p$-value $0 = 1.0E - 10$ this time and saved them with $p$-value, accuracy, coverage

**Table 6.** Finally estimated rules by apriori-STRIM for the DT corresponding to those of Table 3.

| Rule no. | Estimated rules $(C(1)\ldots C(6)D)$ | Count | $p$-value | Accuracy | Coverage |
|---|---|---|---|---|---|
| 1 | 0011001 | 207 | 1.93E−81 | 0.945 | 0.200 |
| 2 | 1100001 | 190 | 1.48E−68 | 0.931 | 0.184 |
| 3 | 0022002 | 216 | 4.19E−94 | 0.911 | 0.223 |
| 4 | 2200002 | 167 | 2.65E−56 | 0.918 | 0.172 |
| 5 | 3300003 | 193 | 2.42E−70 | 0.902 | 0.186 |
| 6 | 0033003 | 185 | 1.64E−64 | 0.907 | 0.178 |
| 7 | 4400004 | 187 | 1.36E−71 | 0.903 | 0.195 |
| 8 | 0044004 | 167 | 6.91E−57 | 0.928 | 0.174 |
| 9 | 0055005 | 211 | 1.37E−87 | 0.898 | 0.212 |
| 10 | 5500005 | 191 | 6.45E−72 | 0.888 | 0.192 |

and so on. Table 5 shows the first nine candidates after sorting them in ascending order of $p$-value. $RN = 1$ and 2 coincide with $R(1, 2)$ and $R(1, 1)$ respectively. $RN = 3$ can be represented by $RN = 2$ with the smaller $p$-value and in the same manner all the following candidates were arranged and represented by $RN = 1$ or 2 at $LN = 14$. Table 6 shows the final rule induction results including those of $D = 2, \ldots, 5$ by apriori-STRIM.

To compare Table 6 by apriori-STRIM with Table 3 by the conventional STRIM, the following is seen:

(1) Both methods statistically induce the pre-specified rules in Table 2 in proper quantities and justly coincide with each other in corresponding figures.
(2) The differences between two tables are their surface caput of $f = (n_1, n_2, \ldots, n_5)$ and *count*. The former focuses on $P(D = d|CP)$ and adopts the strongest bias of the distribution of $D$ by $CP$. The latter focuses on $P(CP|D = d)$ and adopts the strongest intervention by rules, which appears in the $p$-value of the co-occurrence set (pattern) in Table 5.

In the same way, to compare Table 6 and/or Table 3 with Table 9 and/or Table 10 by the associate rule learning (ARL), the following is seen:

(3) ARL first focuses on the co-occurrence set of $(CP, D)$ and its count directly connects to $P(CP, D)$ and induces rules by use of parameters of support, confidence, and so on. However, ARL has no way of distinguishing the co-occurrence sets by rules from those by chance since ARL doesn't have any models for the distinction.
(4) Connecting to (3), ARL also has no way of arranging a large number of rule candidates as shown in Appendix A although it has useful indexes of support, confidence, lift, and so on. That is, ARL based on $P(CP, D)$ seems not to closely focus on inducing if-then rules although it can induce the co-occurrence set.

## 6 Conclusion

This paper summarized the rule induction methods by the conventional RS and their statistically improved method STRIM showing the validity and capacity of STRIM in VSofRIM. The conventional STRIM focused on $P(D = d|CP)$ which can be recognized as a probabilistic structure transforming the input $CP$ into the corresponding output $D$, and used the structure and the DT for inducing the if-then rules that causes the bias in the distribution of $D$. From this view, another new rule induction method focusing on $P(CP|D = d)$ was proposed. Specifically, the method estimated the inputs after receiving the outputs by exploring the co-occurrence set of $U(d) = \{u(i)|u^{D=d}(i)\}$ and executing statistical testing with regard to the explored set under $H0$: $P(CP|D = d) = P(CP)$. The exploration was executed by Apriori algorithm developed in the field of ARL. The new method was named apriori-STRIM. The validity and capacity for apriori-STRIM was confirmed by applying it to the same DT as the conventional STRIM. The similarities and differences between the conventional, apriori-STRIM and ARL were clarified through the same simulation dataset, that is, VSofRIM.

**Table 7.** An example of transaction dataset.

| Transaction | Record |
|---|---|
| $tr(1)$ | $1, 2, 5, 6, 7, 9$ |
| $tr(2)$ | $2, 3, 4, 5$ |
| $tr(3)$ | $1, 2, 7, 8, 9$ |
| $tr(4)$ | $1, 7, 9$ |
| $tr(5)$ | $2, 3, 7, 9$ |

**Table 8.** An example of the set of FIS (SFIS) for Table 7 ($\theta_0 = 3$).

| |
|---|
| SFIS(1) = $\{\{1\}, \{2\}, \{7\}, \{9\}\}$ |
| SFIS(2) = $\{\{1, 2\}, \{1, 7\}, \{1, 9\}, \{2, 7\}, \{2, 9\}, \{7, 9\}\}$ |
| $\Rightarrow\{\{1, 7\}, \{1, 9\}, \{2, 7\}, \{2, 9\}, \{7, 9\}\}$ |
| SFIS (3) = $\{\{1, 2, 7\}, \{1, 2, 9\}, \{1, 7, 9\}, \{2, 7, 9\}\}$ |
| $\Rightarrow\{\{1, 7, 9\}, \{2, 7, 9\}\}$ |
| SFIS(4) = $\{1, 2, 7, 9\}$ |
| $\Rightarrow\{\phi\}$ |

Focus for future studies:

(1) The differences of performance evaluation between the above three methods were considered by the plain data generation model. To examine them in a much closer model to the real-world.
(2) To apply three methods to the real-world dataset after finishing the studies (1).
(3) To expand the DT to the transaction database and study if both STRIMs can be applied to such a database and work effectively.

## A Transaction Database and Association Rule Learning [21]

Transaction database ($TrD$) is defined as a set of records called transaction ($tr$): $TrD = \{tr(i)|i = 1, \ldots, m\}$. Here, each $tr(i)$ is a subset of an item set defined with $Itm = \{itm(j)|j = 1, \ldots, n\}$. One of the examples is shown in Table 7 where $m = 5$ and $Itm = \{itm(j) = j|j = 1, \ldots, n = 9\}$. Now let be $\forall X \subseteq Itm$ then $Occ(X) = \{tr(i)|X \subseteq tr(i)\}$ is called the occurrence set of $X$ and its frequency is denoted $freq(X) = |Occ(X)|$. For example, let be $X = \{1\}$ in Table 7 then $Occ(X) = \{tr(1), tr(3), tr(4)\}$ and $freq(X) = 3$. $\exists X \subseteq Itm$ whose occurrence set is often found in $TrD$ is called a frequent item set ($FIS$). Table 8 arranges $FIS$ of $X$ with $freq(X) \geq \theta_0 = 3$ in Table 7 and shows the set of $FIS$ ($SFIS(|X|)$) every $|X|$. For example, $SFIS(|X| = 1)$ in Table 8 can be easily obtained by tallying the frequency of $TrD$ with $X$. $SFIS(|X| = 2)$ should be constructed by every combination of the element of $SFIS(|X| = 1)$ and confirm them by use of $TrD$ then $freq(\{1, 2\}) = 2 \not\geq 3$ and $\{1, 2\}$ is deleted. The result is shown after the symbol "⇒". In the same way, $SFIS(|X| = 3)$ should be constructed by every combination of items in $SFIS(|X| = 2)$. However, $\{1, 2, 7\}$ or $\{1, 2, 9\}$ should be deleted since $SFIS(|X| = 2)$ doesn't include $\{1, 2\}$, which is called downward closure property of frequency ($DCPF$). The rest of $FIS(|X| = 3)$ are confirmed by use of $TrD$. As well as the following.

The algorithm that effectively generates $SFIS(|X| = l+1)$ from $SFIS(|X| = l)$ by use of $DCPF$ is called Apriori algorithm [18] and implemented by R language as apriori() function in the library "arules" and is often used for association rule learning [17] problems.

Now let $\forall X, \forall Y \subseteq Itm$ and $X \cup Y \in FIS \subseteq SFIS$ then $X$ and $Y$ often simultaneously occur, which is called a co-occurrence set with $freq(X \cup Y)$ and induce rules called the association rule (AR): if $X$ then $Y$, or: if $Y$ then $X$. The following three indexes: support, confidence and lift are often referred as the quality of AR:

$$supp(X) = \frac{freq(X)}{|TrD|} = P(X),$$

$$conf(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)} = P(Y|X),$$

$$lift(X \to Y) = \frac{supp(X \cup Y)}{supp(X) \cdot supp(Y)} = \frac{conf(X \to Y)}{supp(Y)} = \frac{P(Y|X)}{P(Y)},$$

where "→" denotes implication. Support is an index of how frequency the item set appears in the dataset and confidence how often the rule has been found to be true. Lift implies the degree to which $X$ and $Y$ are dependent on one another. If $lift(X \to Y) = 1$ then $X$ and $Y$ are independent of each other and AR has no sense.

For example, $\{1, 7, 9\}$ in Table 8 is a co-occurrence set with $freq(\{1, 7, 9\}) = 3$, which induces AR: if $\{1, 7\}$ has occurred then $\{9\}$ will occur with $supp(\{1, 7\}) = 3/5$. $conf(\{1, 7\} \to \{9\}) = 3/3 = 1$ and $lift(\{1, 7\} \to \{9\}) = 5/4$. This AR is valid to some extent since the $lift > 1$.

There are various kinds of $TrD = \{tr(i)|i = 1, \ldots, m\}$ and DT: $S = (U, A = C \cup \{D\}, V, \rho)$ can be regarded as one of $TrD$ with corresponding relationships: $N \to m$, $u(i) = (\rho(u(i), C(1)) \ldots \rho(u(i), C(|C|)) \ \rho(u(i), D)) \to tr(i)$ and $V = \cup_{a \in A} V_a \to Itm$. For example, if $u(1) = (1234512)$ in the specification of Sect. 3 then $\{11, 22, 33, 44, 55, 61, 72\}$ corresponds to $tr(1)$. In this way, the $U$ with $N =$ 5,000 corresponding to Table 3 can be transformed into the $TrD$ form: d.tran and the if-then rules behind the $U$ can be induced by ARL of the following statement:

apriori(d.tran, parameter = list(support = 0.003, confidence = 0.80, maxlen = 5)).

This example induces rules satisfying $conf(X \to Y) \geq 0.80$ after finding the co-occurrence set satisfying the condition $supp(X \cup Y) \geq 0.003$ and $|X \cup Y| \leq 5$. The part of the number of 240 ARs induced is shown in Table 9 after sorting them in descending order of $lift$. In the surface caput of Table 9, Rule No. shows the descending order, $lhs$ and $rhs$ are abbreviations of left hand side and right hand side of an if-then rule respectively, and count is $freq(lhs \cup rhs)$.

**Table 9.** An example of estimated rules by apriori function for the dataset corresponding to that in Table 3.

| Rule no. | lhs | | rhs | Support | Confidence | Lift | Count |
|---|---|---|---|---|---|---|---|
| 1 | {24, 34, 44} | => | {74} | 0.008 | 1.000 | 5.20 | 40 |
| 5 | {14, 24, 42, 65} | => | {74} | 0.003 | 1.000 | 5.20 | 15 |
| 6 | {12, 22, 32} | => | {72} | 0.007 | 1.000 | 5.15 | 36 |
| 14 | {15, 21, 32, 42} | => | {72} | 0.003 | 1.000 | 5.15 | 15 |
| 22 | {15, 25, 55, 63} | => | {75} | 0.003 | 1.000 | 5.02 | 17 |
| 23 | {11, 23, 35, 45} | => | {75} | 0.003 | 1.000 | 5.02 | 15 |
| 41 | {24, 42, 63, 72} | => | {32} | 0.003 | 0.938 | 4.85 | 15 |
| 42 | {11, 31, 41} | => | {71} | 0.010 | 1.000 | 4.84 | 49 |
| 55 | {23, 33, 43} | => | {73} | 0.006 | 1.000 | 4.81 | 32 |
| 56 | {13, 23, 43} | => | {73} | 0.006 | 1.000 | 4.81 | 29 |
| 75 | {12, 22} | => | {72} | 0.033 | 0.918 | 4.73 | 167 |
| 76 | {11, 21, 32} | => | {71} | 0.008 | 0.976 | 4.72 | 40 |

Table 9 shows the following: For example, $RN = 1$ indicates if $C(2) = 4 \wedge C(3) = 4 \wedge C(4) = 4$ then $D = 4$ and $count = freq(lhs \cup rhs) = support \cdot |TrD| = 0.008 \cdot 5000 = 40$. $RN = 41$ is an interesting case that $lhs$ includes the decision attribute value 72 ($D = 2$) and $rhs$ is the condition attribute 32 ($C(3) = 2$). Such rules should be deleted when inducing rules from DT by ARL since $TrD$ has neither the explanatory nor response variables. When inducing rules from an information table which doesn't have such a distinction, ARL can be used. $RN = 75$ coincides with $R(2, 1)$ in Table 2. However, most of ARs from the DT is such rules adding a pair of the condition attribute and its value to the pre-specified rules in Table 2, that is, the part of pre-specified rules or those having no sense against them.

**Table 10.** An example of estimated rules of $D = 1$ by apriori function for the dataset corresponding to that in Table 3.

| Rule No. | lhs | | rhs | Support | Confidence | Lift | Count |
|---|---|---|---|---|---|---|---|
| 42(1) | {11, 31, 41} | => | {71} | 0.010 | 1.000 | 4.84 | 49 |
| 47(6) | {11, 21, 31, 43} | => | {71} | 0.003 | 1.000 | 4.84 | 17 |
| 48(7) | {11, 21, 43, 52} | => | {71} | 0.003 | 1.000 | 4.84 | 15 |
| 49(8) | {14, 25, 31, 41} | => | {71} | 0.004 | 1.000 | 4.84 | 18 |
| 124(24) | {25, 31, 41} | => | {71} | 0.010 | 0.945 | 4.58 | 52 |
| 125(25) | {31, 41} | => | {71} | 0.041 | 0.945 | 4.58 | 207 |
| 143(30) | {11, 21} | => | {71} | 0.038 | 0.931 | 4.51 | 190 |
| 208(46) | {11, 21, 63} | => | {71} | 0.008 | 0.891 | 4.31 | 41 |
| 211(47) | {11, 21, 43} | => | {71} | 0.010 | 0.889 | 4.30 | 48 |
| 231(48) | {11, 21, 53} | => | {71} | 0.006 | 0.857 | 4.15 | 30 |
| 237(49) | {11, 21, 33} | => | {71} | 0.006 | 0.829 | 4.01 | 29 |

Table 10 shows the part of 49 ARs of $D = 1$ extracted from the above 240 ARs where the number within parentheses in Rule No. shows the $lift$ order in the 49 ARs of $D = 1$. This table shows us the following:

(1) The pre-specified rules in Table 2 appears in $RN = 125(25)$ and $143(30)$ although there is no objective criterion or standard to adopt them by use of support, confidence or $lift$ and so on since ARL has no way of arranging a lot of ARs based on an objective principle.
(2) Accordingly, an analysist can't help but subjectively adopt several ARs by his own domain knowledge referring to indexes like $lift$ and so on. Incidentally, the above apriori() function also had difficulties specifying its parameters. For example, when specified $support = 0.001$ or $0.008$ fixing the other parameters, the function induced ARs of the number of $1,893$ or $92$. This example suggests that the specification for its parameters including their combinations will puzzle the analysist and he can't help but subjectively

specify them based on his domain knowledge after many trials when analyzing the real-world $TrD$.

## References


1. Pawlak, Z.: Rough sets. Int. J. 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.) Intelligent Decision Support, Handbook of Application and Advances of Rough Set Theory, pp. 331–362. Kluwer Academic Publishers, Dordrecht (1992)
3. 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, Dordrecht (1992)
4. Ziarko, W.: Variable precision rough set model. J. Comput. Syst. Sci. **46**, 39–59 (1993)
5. Shan, N., Ziarko, W.: Data-based acquisition and incremental modification of classification rules. Comput. Intell. **11**(2), 357–370 (1995)
6. 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
7. 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
8. 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
9. 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)
10. 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
11. 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
12. 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
13. 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

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. 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. J. **73**, 933–942 (2018)
16. Kato, Y., Saeki, T.: Studies on reducing the necessary data size for rule induction from the decision table by STRIM. In: Mihálydeák, T., et al. (eds.) IJCRS 2019. LNCS (LNAI), vol. 11499, pp. 130–143. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22815-6_11
17. Piatetsky-Shapiro, G.: Discovery, analysis, and presentation of strong rules. In: Piatetsky-Shapiro, G., Frawley, W.J. (eds.) Knowledge Discovery in Databases. AAAI/MT Press, Cambridge (1991)
18. Agrawal, R., Srikant, R.: Fast algorithm for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases, VLDB 1994, pp. 487–499 (1994)
19. 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)
20. https://www.rdocumentation.org/packages/arules/versions/1.5-5
21. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data Mining Practical Machine Learning Tools and Techniques, 4th edn., pp. 120–127. Morgan Kaufmann Publishers, Burlington (2017)