STRIM によるルール導出原理と適用範囲に関する考察*

加藤 裕一[†]・佐伯 徹郎[‡]・水野祥太郎[†]

Considerations on the Principle of Rule Induction by STRIM and its Capacity for Application*

Yuichi KATO[†], Tetsuro SAEKI[‡] and Shotaro MIZUNO[†]

STRIM (Statistical Test Rule Induction Method) has been proposed as a method to effectively induct if-then rules from the decision table, and its effectiveness have been confirmed by simulation experiments. However, the previous work has only pointed out the problems of the conventional methods and simply proposed the method to overcome them, and STRIM has yet to be applied to several problems and/or has contained some limitations in STRIM which must be solved before it can be applied to real-world datasets. This paper further examines the limitations of STRIM as presently introduced, and considers several conditions for its application and utilization. Specifically, these are to eliminate the limitation of the number of the decision attribute values, and to clarify the principle for STRIM to induct true rules, the size of the dataset needed, and the relationships and differences between STRIM and the conventional methods. Real-world datasets often contain missing values in the condition attributes, and contaminated values in the decision attribute, from various reasons. Based on the above considerations, this paper reports simulation experiments to examine the capacity of STRIM in such circumstances. The results show the method seems to be sufficiently robust for application to real-world datasets.

1. はじめに

1992 年に Pawlak[1] により提案されたラフ集合理論は、決定表とよばれるデータベースから評定構造や知識構造を、if-then ルール形式で表現し導出する手法として利用されている。このルール導出法は、識別、診断、決定問題などに利用される。このため決定表からのルール導出原理やその性質、効率的なアルゴリズムに関する多くの考察が報告されている [2–7]。しかしながら従来のルール導出法は、データはすでに与えられたものとして議論を出発させており、この枠組みで数学的な形式化と論理的な推論法の考察に重点を置いたものが多い。こ

のため実データを扱うために必要なデータ発生機構も含めた議論がなされていないように思える。そこでわれわれは、データ発生モデルとこれに基づく統計的ルール導出法、ならびにそのアルゴリズム STRIM を提案した [8,9]。文献 [8,9] では、あらかじめ設定したルールからシミュレーションデータを作成し、このデータを利用して従来法 [4,6,7] と STRIM とのルール導出性能の比較実験が示されている。その結果、従来法は 1) 数個の設定ルールを数 1000 倍の部分ルールで推定する傾向がある、2) 同一の設定ルールから発生させた異なる決定表に対して、異なるルールを推定する傾向があることを示している。すなわち従来法は標本に強く依存し、ほとんどルール推定手法の役割を果たしていないことを示している。一方 STRIM は従来法による問題点 1) 2) を画期的に改善して、あらかじめ設定した真のルールを推定できることを示している。

しかしながら文献 [8,9] での考察は、従来法の問題点の指摘と統計的手法としての STRIM による改善法の導入的考察に留まっており、実データに適用するうえで多くの課題が残されていた。具体的には、1) 適用できる決定属性値数に制約がない STRIM へ拡張すること（文

* 原稿受付 2013 年 6 月 21 日

[†] 島根大学 総合理工学部 Interdisciplinary Faculty of Science and Engineering, Shimane University; 1060 Nishikawatsu-cho, Matsue city, Shimane 690-8504, JAPAN[‡] 山口大学 工学部 Faculty of Engineering, Yamaguchi University; 2-16-1 Tokiwadai, Ube city, Yamaguchi 755-8611, JAPAN**Key Words:** rough sets, if-then rule, simulation experiment, missing and contaminated values, statistical test.

献 [8,9] では、決定属性値数は 2 で考察していた)、2) STRIM の統計的ルール導出原理と従来法との関連性を考察すること、3) STRIM を適用するための必要データサイズを考察することなどである。さらに実データは、条件属性値に欠損値を含む場合や決定属性値が何らかの理由で汚染されている場合がある。すなわち、4) 文献 [8,9] で提案したデータ発生モデルをより実システムに近づける改善などが残されていた。

本論文ではまず、観測システムと欠損値・汚染データ発生モデルを組み込んだより実システムに近いデータ発生モデルを提案する。このうえで具体的に決定属性値数が 6 のシミュレーションデータを発生し、このようなデータからルール導出可能であるように文献 [8,9] の STRIM を拡張する。またこの導出プロセスを通して、従来の STRIM を要約・整理する。具体的には、この手法は 2 段階からなっている。第 1 段階は、「真の if-then ルールは決定属性値の分布に偏りを生起させる」という原理に着目し、ルール候補を効率よく探索する段階である。第 2 段階はルール候補を整理して最終的なルールを決定する段階である。この整理の後、STRIM の第 1 段階のルール候補探索法と従来法で最もよく利用される VPRS (Variable Precision Rough Set) [5] によるルール導出法との関係を考察する。VPRS は許容矛盾率 ε を導入して、(accuracy, coverage) の指標の高いルールを選定する手法である。この考察で 1) accuracy は偏りの指標であり、VPRS は STRIM と同一のルール導出原理とみなせること、2) 両者の差異はこの偏りに検定保証があるか否か、すなわち、考察の背景にデータ発生モデルがあるか否かであること、3) 許容矛盾率 ε と STRIM で利用する統計的有意水準との対応関係を考察している。さらに coverage の指標は STRIM の検定モデルの適用条件に対応していることも考察している。ついで第 2 段階のルール候補の整理手法の合理性を、統計モデルに基づき考察して真のルールを推定する条件を考察している。

さらに上記第一段階で検定適用条件を満足するデータサイズ $N(w, RL)$ を、真のルールを推定する確率 w [%] とルール長 RL をパラメータとして導出している。この考察から、ほぼ $w = 100$ [%] で真のルールを導出する最小のデータサイズ $N_{lst} = N(w = 100[\%], RL)$ や、与えられたデータのサイズ N_{gvn} で真のルールを導出する確率 $w = N^{-1}(N_{gvn}, RL)$ を見積もることが可能となる。最後に、この考察に基づいて発生した N_{lst} サイズの決定表に欠損データや汚染データを発生させて、STRIM のルール導出能力を考察している。その結果、STRIM はそのようなノイズを含んだデータからのルール導出にかなりの頑健性をもつことを示している。

このような考察やシミュレーション実験結果は、従来法に比して STRIM のルール導出能力の優位性を示すとともに、実データに利用するうえでの条件や導出結果の妥当性を検討するうえで重要である。

2. 決定表とデータ発生モデル

ラフ集合理論は、与えられた決定表 S に埋もれた知識を if-then ルール形式で導出する手法として利用される。ここで S はつぎのように形式表現される: $S = (U, A = C \cup \{D\}, V, \rho)$ 。ここで $U = \{u(i) | i = 1, \dots, N = |U|\}$ はサンプル集合、 A は属性集合、 $C = \{C(j) | j = 1, \dots, |C|\}$ は条件属性集合、 $C(j)$ は条件属性、 D は決定属性、 $V = \bigcup_{a \in A} V_a$ は属性値の集合、 $\rho: U \times A \rightarrow V$ は情報関数である。たとえば $a = C(j) \in A$ ($j = 1, \dots, |C|$) のとき $V_a = \{1, 2, \dots, M_{C(j)}\}$, $a = D$ とすると $V_a = \{1, 2, \dots, M_D\}$ となる。Table 1 は $|C| = 6$, $M_{C(j)} = 6$, $M_D = 6$, $\rho(u(1), C(1)) = 5$, $\rho(u(2), C(2)) = 5$, ... の一例である (発生法は後述する)。

STRIM [8,9] は、決定表のデータ $u(i)$ を Fig. 1 に示す Rule Box の入出力関係と Table 2 に示す決定属性値仮説から生成された標本データ集合と考えている。すなわち、決定表の標本 $u(i)$ はその条件属性値 $u^C(i)$ と決定属性値 $u^D(i)$ からなる。ここで $u^C(i)$ は Rule Box への入力であり、Rule Box 中の設定ルールと Table 2 の決定属性値仮説により $u^D(i)$ に変換される。さらに Fig. 1 には実システムに近づけるために、観測雑音 NoiseC と NoiseD を考慮して文献 [8,9] のモデルを拡張している。これらはそれぞれ条件属性値に欠損値を生起させ、決定属性値を汚染する。このようなデータを観測器 Observer は $u(i) = (u^C(i), u^D(i))$ として観測している。

たとえば

設定ルール :

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

$Rd = (C(1) = d) \wedge (C(2) = d) \vee (C(3) = d) \wedge (C(4) = d)$ を設定する。ここで、 $\wedge$ は連語、 $\vee$ は選語である。このうえで一様乱数を利用してサンプル $u(i)$ の条件部 $u^C(i) = (v_{C(1)}(i), v_{C(2)}(i), \dots, v_{C(|C|)}(i))$ を発生し、上記設定ルールならびに仮説を利用して、標本 $u(i)$ の決定属性値 $u^D(i)$ を決定する。Table 1 はこのように発生させた決定表の例である。ただし Table 1 は単純に考察するために観測雑音を混入させずに発生している。上記データ発生モデルは、実データの発生モデルとしてもそれほど奇異ではないと思われる。与えられた決定表から if-then ルールを導出する問題は、上記モデルを前提として観測入出力データ集合から Rule Box 内の設定ルールを同定する問題である。以降、上記設定ルール (真のルール) で発生したデータを利用して考察を進める。

3. STRIM によるルール推定手法の要約

STRIM [8,9] は Table 1 に示すような決定表の $u(i)$ ($i = 1, \dots, N = |U|$) を利用して、2 段階処理でルール推定する手法である。第 1 段階は無関心データ (Table 2 の仮説 2 により生成したデータ) と一意決定可能データおよび矛盾データ (Table 2 の仮説 1 および仮説 3 によ

Table 1 An example of a decision table

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

Fig. 1 Rough Sets system contaminated with noise: Input is a tuple of the condition attributes' value and its output is the decision attribute's value

Table 2 Hypothesis with regard to the decision attribute value

仮説 1	$u^C(i)$ は if-then ルールの条件部に当てはまり、結論部の決定属性値 $u^D(i)$ が一意に定まる (一意決定可能データ)
仮説 2	$u^C(i)$ は if-then ルールのどの条件部にも当てはまらず、 $u^D(i)$ をランダムに決めざるをえない (無関心データ)
仮説 3	$u^C(i)$ は複数の if-then ルールの条件部に当てはまり、一意に $u^D(i)$ を決定できず、複数の該当ルールの一つをランダムに選んで $u^D(i)$ を決定せざるをえない (矛盾データ)

り生成したデータ)とを統計的に識別する。具体的にはまずルール候補として、条件部 $CP(k) = \bigwedge_j (C(j_k) = v_j) (\in V_{C(j_k)})$ を仮定し、 $CP(k)$ を満足する $u(i)$ の集合 $U(CP(k)) = \{u(i) | u^C = CP(k)(i)\}$ 、 $D = m$ を満足する $u(i)$ の集合 $U(m) = \{u(i) | u^D = m(i)\} (m=1, \dots, M_D)$ を算定する。さらに $U(CP(k))$ に対する $u^D(i)$ の度数分布 $f: (n_1, n_2, \dots, n_{M_D})$ を算定する。ここで、 $n_m = |U(CP(k) \cap U(m))| (m=1, \dots, M_D)$ である。設定ルール Rd は一意決定可能データ (仮説 1) と矛盾データ (仮説 3) のみを生成するので、 $U(Rd)$ の f の分布には偏りが生じる。したがって $CP(k)$ が Rd の十分条件 ($U(Rd) \supseteq U(CP(k))$) である場合は、この $CP(k)$ の f の分布は、 $U(Rd)$ の部分集合としての偏りがある。 $CP(k)$ が設定ルール Rd の必要条件 ($U(CP(k)) \supseteq U(Rd)$) である場合は、一部は無関心データ (仮説 2) も含むが、 $U(Rd)$ の偏りの分布を含むので $U(CP(k))$ の f の分布にも偏りが生じる可能性が高い。一方 $CP(k)$ が設定ルール Rd の十分条件 ($U(Rd) \supseteq U(CP(k))$) あるいは必要条件 ($U(CP(k)) \supseteq U(Rd)$) でもない場合は、 $U(CP(k))$ は $U(Rd)$ の部分集合を含む場合もあるが、多くの無関心

データを中心に f が構成されて偏りが生じにくいと推察される。したがって $CP(k)$ をルール候補として採用するか否かの基準は、 $CP(k)$ による f の偏りを判定すればよい。 Rd との共通性 ($|U(CP(k) \cap U(Rd))|$) が大であれば、 f の分布の偏りが大きくなるはずである。このために帰無仮説 $H0: (n_1, n_2, \dots, n_{M_D})$ は偏りがない。対立仮説 $H1: (n_1, n_2, \dots, n_{M_D})$ は偏りがある、を設定して適当な統計量のもとで $H0$ を検定すればよい。

Table 3 は、Table 1 の決定表で $N = 10000$ としてデータを発生し、試みのルール条件部 $CP(k)$ とその $u^D(i)$ の分布 $f: (n_1, n_2, \dots, n_6)$ ならびに偏りの指標 z を具体的に記した一例である。同表の 1 行目: 100000 はルール条件部が $C(1) = 1$ の意であり、ルール長 $RL = 1$ のルールという。5 行目: 110000 は $C(1) = 1 \wedge C(2) = 1$ で $RL = 2$, 13 行目: 110010 は $C(1) = 1 \wedge C(2) = 1 \wedge C(5) = 1$ で $RL = 3$ のルール条件の意などである。これに対応して、 $n_m = |U(CP(k)) \cap U(m)| (m=1, \dots, M_D)$ を $(n_1, n_2, \dots, n_6)$ の列に記す。 z の列は n_d を半整数補正して、

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

を記している。ここで、 $n_d = \max(n_1, n_2, \dots, n_{M_D} = n_6)$, $(d \in \{1, 2, \dots, M_D = 6\})$, $p_d = P(D = d)$, $n = \sum_{m=1}^{M_D} n_m$ である。 $H0$ のもとで検定適用条件 [10]: $np_d n \geq 5$ かつ $n(1 - p_d) \geq 5$ を満足するとき、 z は近似的に標準正規分布に従う。いま、 $z \geq z_\alpha = 3.0$ と偏りの判定基準 (有意水準) を設定して $H0$ を棄却したとき、この候補は結論部 $D = d$ のルール候補となる。Table 3 の z 値とあらかじめ設定したルールから上記 $CP(k)$ による判別法の有効性をうかがうことができる。

第 2 段階は、 $H0$ を棄却したルール候補集合 $\{CP(k) | H0 \text{ を棄却した } CP(k) (k=k1, k2, k3, \dots)\}$ から、最終的に設定ルールを絞り込む段階である。 $CP(k)$ は有意水準以上の f の分布の偏りをもつ (第一段階で $H0$ を棄却して、 $H1$ を採択している) ルール候補であることから、これらの候補間に $CP(ki) \supseteq CP(kj) \supseteq CP(kl) \dots$ を満たす場合がある。たとえば、Table 3 でもわかるようにルールの条件部に包含関係: $100000 \supset 110000 \supset 110010$ が生じる場合がある。このような場合、STRIM は包含関係のルール条件部の中で偏りの最大のもの、すなわち最大の z のもので代表する。この例の場合は 110000 である。これは偶然であるがあらかじめ設定したルールと一致している。Fig. 2 に STRIM のアルゴリズムを示す (文献 [8,9] 参照)。

Table 4 に Table 1 の決定表 ($N = 10000$) に対する導出結果を示す。あらかじめ設定した 12 個のルールすべてと余分なルール一つを導出している。しかし両者の accuracy と coverage (定義は 4. を参照) には明らか

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

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

```

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

int rule_check(int tail,int rule[|C|]) { // Stage 1
for (ci=tail+1; cj<|C|; ci++) {
for (cj=1; cj<=|C[ci]|; cj++) {
rule[ci]=cj; // a trying rule sets for test
count frequency of the trying rule; // count n1, n2, ...
if (frequency>=NO) { //sufficient frequency ?
if (|z|>3.0) { //sufficient evidence ?
store necessary data such as rule, frequency of
n1 and n2, and z
} // end of if |z|
rule_check(ci,rule);
} // end of if frequency
} // end of for cj
rule[ci]=0; // trying rules reset
} // end of for ci
} // end of rule_check

```

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

に大きな差異が見て取れる。

4. STRIM の原理に関する考察

4.1 従来法と STRIM の関係

従来のラフ集合理論では文献[4]の LEM2 アルゴリズムや文献[6,7]の決定行列を用いたアルゴリズムによる下近似のルール導出法がよく利用されていた。しかしながら、これらにはつぎの問題点がある（文献[8,9]参照）。問題点 1: ルール長 RL が大の多くの部分ルールを推定する傾向がある（Table 4 から STRIM はこの問題点を画期的に解決していることがわかる）。問題点 2: 同一のルール集合から生成した異なる決定表に対して、異なるルールを推定する傾向になる。これは従来法が一標本の決定表に強く依存する問題点である（STRIM はこの問題点も解決していることを文献[8,9]で示している）。そこで Ziarko[5]は VPRS を提案している。VPRS は許容矛盾率 $\varepsilon \in [0, 0.5]$ を導入して、

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

あるいは、

$$\overline{C}_\varepsilon(U(d)) = \{u(i) | acc > \varepsilon\} \quad (3)$$

を満足する $CP(k)$ を求める。ここで $acc = |U(d) \cap U(CP(k))| / |U(CP(k))| = n_d / n$ はルールの正確度 accuracy で、 $\underline{C}_\varepsilon(U(d))$ は VPRS の ε -下近似、 $\overline{C}_\varepsilon(U(d))$ は ε -上近似とよばれる。 $\varepsilon = 0$ でそれぞれ通常の下近似、上近似に一致する。とくに ε -下近似は、被覆度 coverage: $cov = |U(d) \cap U(CP(k))| / |U(d)|$ を定義して、 cov のある程度大きな $CP(k)$ を選択すれば、問題点 1) を避けることができるとされてきた。従来の問題点 2) は Jaworski が考察している[11]。Jaworski は決定表を標

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

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

本空間から母集団に拡張して, acc と cov の信頼区間を推定することを提案して, 問題点 2) をある程度緩和した.

VPRS は 1993 年に Ziarko により提案されて以来, 通常のラフ集合の上記問題点を克服して ε -下近似は, 最もよく利用されるルール導出手法である. また考察対象や問題によってオリジナルの VPRS をさまざまに利用したものや修正した VPRS が提案されている [12–14]. しかしながら, どの程度に ε を設定すべきかの根拠ある指針が示されていない. また VPRS は $\varepsilon \in [0, 0.5]$ を導入したことにより, LEM2 のような系統的なルール導出アルゴリズムをわれわれの知る限りもっていない. そのため原則的には全数探索や発見的方法を利用している.

(2),(3) 式から ε -下近似・上近似の差異は, acc の範囲にあるが, 共通に $acc \geq (>) acc0$ (ε -下近似の場合: $acc0 = 1 - \varepsilon$, ε -上近似の場合: $acc0 = \varepsilon$) を満足する $U(CP(k))$ すなわち $CP(k)$ を求めることは同一である (以下 $\geq$ で代表して議論する). この共通性は $z \geq z_\alpha$ を満足する $CP(k)$ を求める STRIM と符合とする. そこでいま無作為に $CP(k)$ を作成すれば, この $U(CP(k))$ が $acc \geq acc0$ の要件を満足する場合もある. このときは多くの $U(CP(k))$ が $U(d)$ の要素となっている場合である. すなわち, $U(CP(k))$ の決定属性が $D=d$ に偏っている場合である. 逆に満足しないときは, $U(CP(k))$ の決定属性値が $D=d$ に偏らず $n_m = |U(CP(k)) \cap U(m)|$ ($m=1, \dots, M_d$), すなわち $(n_1, n_2, \dots, n_{M_d})$ の分布がばらついている場合である. このようなルール導出法の考え方は STRIM とまったく同一である. 両者の差異は偏りの判定に acc を用いるか統計モデルを導入して $(n_1, n_2, \dots, n_{M_d})$ の偏りを判定するかである. 後者は統計的根拠を与える手法である.

いま, $acc \geq acc0$ を満足する n_d は $acc0$ を利用すれば, $n_d \geq n \cdot acc0$ を満足する. この n_d が出現する確率はつぎのように評価される.

$$P(n_d \geq n \cdot acc0) = P\left(\frac{n_d + 0.5 - np_d}{(np_d(1-p_d))^{0.5}} \geq \frac{n \cdot acc0 + 0.5 - np_d}{(np_d(1-p_d))^{0.5}}\right)$$

したがって, (1) 式と比較すれば, $z_\alpha = \frac{n \cdot acc0 + 0.5 - np_d}{(np_d(1-p_d))^{0.5}}$ の関係が成立している. したがって

$$acc0 = p_d - 0.5/n + z_\alpha \left(\frac{p_d}{n}(1-p_d)\right)^{0.5} \quad (4)$$

の関係があることがわかる. STRIM は VPRS の許容矛盾率 ε の設定の根拠を与えることができる. これは, STRIM が統計モデルに根ざす大きな利点の一つである.

また VPRS の cov の指標は, ルールとして承認されるためのデータ数として, 分子の $n_d = |U(d) \cap U(CP(k))|$ が分母の $|U(d)|$ に比して, ある程度は必要であるという意味に解釈できる. これは, STRIM の検定統計量の適用条件: $p_d n \geq 5$ かつ $n(1-p_d) \geq 5$ に対応することもある. STRIM には $CP(k)$ が検定条件を満たすという明確な統計的意義があるが, VPRS の cov には accuracy の場合と同様にそのような基準や意味はない.

Table 4 の $R(i=13)$ の例でみると, $n = 254$, $p_d = 1/6$, $z_\alpha = 3.0$ を (4) 式に代入して, $acc0 = 0.235 = \varepsilon$ となる. これは $acc = 0.240 \geq \varepsilon$ を満足している. この場合は ε -上近似のルールを推定しているが, $acc = 0.240 \not\geq 1 - \varepsilon = 0.765$ となり ε -下近似のルールにはならない. 同様に $R(i=12)$ は, $acc0 = 0.234 = \varepsilon$ で $acc = 0.866 > \varepsilon$ となり ε -上近似のルールであると同時に $acc = 0.866 > 1 - \varepsilon = 0.766$ となり ε -下近似のルールでもある. さらに, $R(i=12)$ の観測値 z は 30.47 なので仮に $z_\alpha = 30.0$ と設定すると $acc0 = 0.855$ となり, この場合は $\varepsilon = 1 - acc0 = 0.145$ と設定して, $acc = 0.866 \geq acc0 = 0.855 = 1 - \varepsilon$ となり, ε -下近似のルール推定を目的としたことになる. 3. では $z_\alpha = 3.0$ と設定してルール導出例を示したが, どの程度に z_α を, すなわちどの程度に $acc0$ を設定してルールを推定するかは, 考

察対象に見識をもつ分析者に委ねられる問題である。

4.2 STRIM の第 2 段階に関する考察

ここでは、第 2 段階でのルール候補選択に関する問題を考察する。すなわち、 H_0 のもとで仮説検定を実行して $CP(ki) \supset CP(kj) \supset CP(kl) \dots$ を満たすルール候補をいくつか検出した場合、STRIM にはつぎの問題が生じる。

問題： STRIM で利用する偏りの指標 z は、ルール候補が真の設定ルールに一致するとき、最大となることを保証するか。あるいは、どのような条件下で保証できるか。

この節では一般性を失うことなく単純化して上記問題を考察するために、サンプルは独立で同一分布 (*i.i.d.*) に従うことを仮定して、つぎのように記号を統一する。

Rule Box 内の設定ルール (真のルール) の条件部の一つを $CP(F)$ として、これによって生成される U の集合を $U(F)$ とする。 $U(F)$ の真部分集合 $U(G)$ は、 $CP(G)$ から生成される。また $U(F)$ を真部分集合として含む $U(E)$ は、 $CP(E)$ から生成されるとする。さらに、 $U(EF) = U(E) - U(F)$ は多くの矛盾・無関心データを含むと仮定する。上記例で示せば、それぞれ $CP(F) = "110000"$, $CP(G) = "110010"$, $CP(E) = "100000"$ の場合である。またこれらは $D = 1$ で最大度数 ($n_{\max} = n_1$) となるとする。集合 $U(F)$ での D の分布 f を $(n_1^F, n_2^F, \dots, n_{M_D}^F)$, z 値を $z^F = \frac{n_1^F + 0.5 - n^F p_1}{(n^F p_1 (1 - p_1))^{0.5}} \approx \frac{n^F (a^F - p_1)}{\sigma^F}$ とする。ここで、 $0.5 \ll n^F p_1$ として 0.5 を無視し、 $n_1^F = a^F n^F$ ($0 < a^F \leq 1$), $n^F = \sum_{j=1}^{M_D} n_j^F$, $\sigma^F = (n^F p_1 (1 - p_1))^{0.5}$ とおいている。

同様に G についても $z^G = \frac{n_1^G + 0.5 - n^G p_1}{(n^G p_1 (1 - p_1))^{0.5}} \approx \frac{n^G (a^G - p_1)}{\sigma^G}$ である。ここで z^F と z^G の関係を考察すれば、 $U(F) \supset U(G)$ で $a^G \approx a^F$ であり、 $n^F > n^G$ であることから、 $n^G = r n^F$, $0 < r < 1$ となる r が存在する。これらを考慮すれば、 z^G の近似値 z_{app}^G は $z_{app}^G = \frac{n^G (a^G - p_1)}{\sigma^G} = \frac{r^{0.5} n^F (a^F - p_1)}{\sigma^F} = r^{0.5} z^F$ となる。これは、 $z^G < z^F$ を意味しており、STRIM の第 2 段階処理では $CP(G)$ ではなく $CP(F)$ が選択されることを意味する。

同様に E についても、 $z^E = \frac{n_1^E + 0.5 - n^E p_1}{(n^E p_1 (1 - p_1))^{0.5}}$ である。ここで z^F と z^E の関係を考察するために、 $n_1^E = n_1^F + n_1^{EF}$, $n^E = a^F n^F + a^{EF} n^{EF}$, $n^E = n^F + n^{EF}$ とおく。ここで、 $n^{EF} = |U(EF)|$ とおいている。この差集合 $U(EF)$ は多くの無関心データ (仮説 2 に従うデータ) を含む。 z^G と同様に 0.5 を無視して、 E と F の関係を考慮すると z^E の近似式 z_{app}^E は次式となる。 $z_{app}^E =$

Fig. 3 Relationship between z^F and z^E

$$\frac{n^F (a^F - p_1)}{\sigma^F} + \frac{n^{EF} (a^{EF} - p_1)}{\sigma^{EF}} = \frac{\sigma^F}{\sigma^E} z^F + \frac{\sigma^{EF}}{\sigma^E} z^{EF}$$

$$= s_E^F z^F + s_E^{EF} z^{EF}$$
ここで、 $n^F < n^E$, $n^{EF} < n^E$ より、 $\frac{\sigma^F}{\sigma^E} = s_E^F < 1$, $\frac{\sigma^{EF}}{\sigma^E} = s_E^{EF} < 1$ となっている。 z^E と z^F の近似的関係を Fig. 3 に示す。同図には、

I : $z^E = s_E^F z^F + s_E^{EF} z^{EF}$

II : $z^E = z^F$

を示している。I と II の交点は $\left(z_0^F = \frac{s_E^{EF}}{1 - s_E^F} z^{EF}, z_0^E \right)$ である。したがって、STRIM の第 1 段階で $z^F \geq z_0^F$ となれば、必ず $z^E < z^F$ となり、これは STRIM の第 2 段階処理で $CP(E)$ ではなく $CP(F)$ が選択されることを意味する。問題に対する一つの結論として、 $z^F \geq z_0^F$ であれば、STRIM は Rule Box 内の設定ルール (真のルール) を導出する。以下具体的に Table 1 の決定表で z_0^F の算定事例を示す。Table 1 では、設定ルールと条件属性値数 $M_{C(j)} = M_C \geq 2$ と設定し、各条件属性値は均等に生起することを仮定している。このとき、 $s_E^F = \left(\frac{1}{M_C} \right)^{0.5}$, $s_E^{EF} = \left(1 - \frac{1}{M_C} \right)^{0.5}$ となり、さらに $x = \frac{1}{M_C}$ とおけば、 $\frac{s_E^{EF}}{1 - s_E^F} = \frac{\sqrt{1-x}}{1-\sqrt{x}} = f(x)$ とおき、 $0 < x \leq 0.5$ の範囲 ($M_C \geq 2$) で $f(x)$ は単調増加して $1 < f(x) < 2.5$ である。 $M_C = 6$ のとき、 $f(0.1667) = 1.542$ 程度である。また z^{EF} は無関心データとして標準正規分布に従うとして、1 [%] 未満の過誤を許して $|z^{EF}| < z_\alpha = 3.0$ が成立する。したがって、 $z^F \geq z_0^F = \frac{s_E^{EF}}{1 - s_E^F} z^{EF} \approx 5$ であれば十分に STRIM の第 2 段階処理で $CP(E)$ ではなく $CP(F)$ が選択される。

この考察を Table 3 と Table 4 でとくに (100000) に関して確認する。Table 3 で (100000) $\supset$ (110000) $\supset$ (110010), (110020), (110050), (110003) は $E - F - G$ の関係になっていることがわかる。したがって、 F の (110000) が最大の z 値として選定され、Table 4 に $R(i=11)$ として現れる。他のルールについても同様である。また、Table 4 で $R(i=13)$ を除いて、 $z \geq 5.0$ を満足していることもわかる。すなわち、 F が選定される条件を満たしていることがわかる。

5. データサイズ N の考察

本章では、STRIM が利用する検定法の適用条件: $p_{dn} \geq 5$, $n(1 - p_d) \geq 5$ を満足する決定表のデータサイズ N を考察する。この考察は決定表 $S = (U, A = C \cup \{D\}, V, \rho)$

のうち、未知データが $|U| = N$ (したがって ρ も未知) の場合、真のルールを推定するためにはどの程度のサンプル数を必要とするかの指針となる。また逆に S が与えられている場合 (したがって $|U| = N$, ρ も既知), この決定表からどの程度の確率で、どの程度の範囲で真のルールを推定できるかの指針を得ることができれば、この考察は有意義である。

いま、検定適用条件を満足する最小のデータ数を N_0 として、つぎの確率事象を考察する。

$$P(n \geq N_0) = P(z \geq z_0) = w \quad (5)$$

ここで、

$$z = \frac{n + 0.5 - Np_C}{\sqrt{Np_C(1-p_C)}}, \quad z_0 = \frac{N_0 + 0.5 - Np_C}{\sqrt{Np_C(1-p_C)}} \quad (6)$$

である。また、 $p_C = P(C = CP(k)) = \prod_j P(C(j_k) = v_k)$ はルール条件部 $CP(k)$ の生起確率である。たとえば、 $CP(K) = (C(1) = 1) \wedge (C(2) = 1)$ ($RL = 2$) であれば、 $p_C = P(C(1) = 1) \cdot P(C(2) = 1)$ である。 z が標準正規分布で十分近似できるという仮定のもとでは、(5) 式は次式となる。

$$P(z \geq z_0) = \int_{z_0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = w \quad (7)$$

w が与えられると、(7) 式を満足する z_0 がわかる。したがって、(6) 式より N に関する 2 次方程式を解いて、(5) 式を満足する最小のデータサイズ N_{lst} は次式で与えられる。

$$N_{lst} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad (8)$$

ここで符号は、+: $z_0 \leq 0$, -: $z_0 > 0$ である ((7) 式で $w \geq 0.5$ で $z_0 \leq 0$ となり、より大きなデータ数 N を必要とするので)。また $a = p_C^2$, $b = -(2p_C(N_0 + 0.5) + z_0^2 p_C(1 - p_C))$, $c = (N_0 + 0.5)^2$ である。以上の考察から (8) 式を満足する N_{lst} は主として RL と w に依存していることがわかる。そこで (8) 式の N_{lst} を $N_{lst}(w, RL)$ と記す。Fig. 4 は 2. の設定: $P(C(j) = v_k) = 1/6$ ($j = 1, \dots, |C| = 6$), $p_d = P(D = d) = 1/6$, したがって $N_0 = 30$ で、 $w = 0.1$ [%] ($z_0 = 3.0$), $= 2.3$ [%] ($z_0 = 2.0$), $= 15.9$ [%] ($z_0 = 1.0$), $= 50.0$ [%] ($z_0 = 0.0$), $= 84.1$ [%] ($z_0 = -1.0$), $= 97.7$ [%] ($z_0 = -2.0$), $= 99.9$ [%] ($z_0 = -3.0$) の 7 点で、 $RL = 1, 2, 3$ の場合のそれぞれについて (8) 式を評価したものである。同図からつぎの有用な情報が読み取れる。

- 1) 真のルールは $RL = 2$ と仮定して、 $w = 99.9$ [%] (ほぼ 100 [%]) で真のルールを導出するためには、

Fig. 4 Theoretical $N(w, RL)$ evaluated by (4) at $w = 0.1$ [%] ($z_0 = 3.0$), $= 2.3$ [%] ($z_0 = 2.0$), $= 15.9$ [%] ($z_0 = 1.0$), $= 50.0$ [%] ($z_0 = 0.0$), $= 84.1$ [%] ($z_0 = -1.0$), $= 97.7$ [%] ($z_0 = -2.0$), $= 99.9$ [%] ($z_0 = -3.0$) (◆: $RL = 1$, ■: $RL = 2$, ▲: $RL = 3$)

Fig. 5 Comparison of $N(w, RL = 2)$ between theoretical and experimental values at $w = 0.1$ [%] ($z_0 = 3.0$), $= 2.3$ [%] ($z_0 = 2.0$), ..., $= 99.9$ [%] ($z_0 = -3.0$) (◆: experimental value, ●: theoretical value)

少なくとも $N = 1865$ は必要である。これを $1865 = N_{lst}(w = 100$ [%], $RL = 2)$ と記す。

- 2) $N = 1000$ の決定表が与えられた場合、 $RL = 2$ と仮定すると真のルールを導出する確率は、30 [%] 程度である。これを 30 [%] $= w = N_{lst}^{-1}(N_{gvn} = 1000, RL = 2)$ と記す。

この考察を確認するために、Table 1 で発生した $N = 10000$ 個のデータを利用して、以下に記す手順に従いシミュレーション実験を実施した。

Step1 : Table 1 から $u(i)$ を $N_{lst}(w, RL = 2)$ 個無作為に抽出して、新たな決定表を作成する。

Step2 : Step1 で作成した決定表から STRIM を利用してルール導出する。導出ルールの中の真のルール数をカウントする。

Step3 : Step1 と Step2 を N_r 回繰り返す。

Step4 : N_r 回の実験での真のルール導出率を算定する。

Fig. 5 は $N_r = 100$ と設定し、 $N_{lst}(w, RL = 2)$ サイズの決定表から導出した真のルールの導出率とその理論値 ((8) 式) との比較を $w = 0.1$ [%] ($z_0 = 3.0$), $= 2.3$ [%] ($z_0 = 2.0$), ..., $= 99.9$ [%] ($z_0 = -3.0$) で示したものである。

実験の導出率は、7 点ではほぼ理論値どおりの値になっていることがわかり、 $N_{lst}(w, RL)$ の考察の正当性の一端をうかがうことができる。

われわれは従来のラフ集合論で、このような必要データ数に関する議論のできるルール推定手法をほとんど知らない。決定表を母集団からの標本データとして、このような定量的な議論ができるのも STRIM の利点である。

6. 欠損・汚染データに対するルール導出性能の実験的検討

6.1 条件属性値に欠損値を含む場合

実データは条件属性値に欠損値を含むことがある。文献 [15] は従来の欠損値の扱いに関する研究を 50 近くの文献とともにつぎのように整理している。

- 1) 従来の欠損値の扱いに関する研究は二つのタイプ: “sequential method” と “parallel method” に分類される。
- 2) “sequential method” は欠損値を何らかの方法であらかじめ補う, あるいは推定する前処理を伴う手法である。前処理を行って欠損値を埋めた後に, 通常のルール導出手法を利用する。文献 [15] は 10 を超える前処理手法を紹介している。
- 3) “parallel method” は 2) のような前処理は行わず, 欠損値をそのまま受け入れてルール導出する手法である。その一方で欠損値をつぎの二つのタイプ: “lost” と “do not care” に分類して, その扱い方を区別している。“lost” はその条件属性の値域から外れた値, “do not care” は値域の任意の値として扱う。

本節では, Fig. 1 の NoiseC が条件属性値に欠損値を作り, この欠損値を含む決定表から STRIM がどの程度の成績でルール導出するかを実験的に検討する。ここでは STRIM のルール導出能力を検討する目的で, “parallel method” で検討する。これは実験成績が前処理に依存せず, STRIM 本来のルール導出能力を考察するためである。以下簡易に表記するために, “lost” は “?” “do not care” は “*” と記す。まず STRIM が欠損値を扱えるように, 文献 [15] にならい $CP(k) = \bigwedge_j (C(j_k) = v_j) (\in V_{C(j_k)})$ に対する $U(CP(k))$ をつぎのように修正した。

$$\forall x \in U, \forall a \in C(j_k) \Rightarrow$$

$$\text{if } \rho(x, a) = ? \text{ then } x \notin U(CP(k)) \text{ and}$$

$$\text{if } \rho(x, a) = * \text{ then } x \in U(CP(k))$$

この扱いで $CP(k)$ に対する $f: (n_1, n_2, \dots, n_{M_D})$ を作成して, STRIM は容易に “?” あるいは “*” を含む決定表からルール導出が可能となる。

欠損値のあるデータの実験は, 5. でのシミュレーション実験手順のうち, Step1 のみをつぎのように修正して同様に実施した。

Step1 : Table 1 ($N = 10000$) から $u(i)$ を $1865 = N_{lst}(w=100, RL=2)$ 個無作為に抽出する。この $u(i)$ ($i=1, \dots, 1865$) の条件属性値が確率 q [%] (混入率) で N_{miss} 個の欠損値を含んでいるか否かを決定する。欠損値がある場合は, 含む条件属性を無作為に N_{miss} 個選んで “?” あるいは “*” で置き換えて新たな決定表を作成する。

Fig. 6 Experimental results with respect to “?”: ■, “*”: ◆

Table 5 は, 導出したルール条件部が “110000” である場合での $q = 0, 50, 100$ [%] での f, p -value (z 値), accuracy, coverage の一例を示している。この表から $N_{lst}(w=100, RL=2) = 1865$ の最小個数での実験であるため, 多少のバラツキは見られるものの, つぎの傾向が読み取れる。

- 1) “?”について, 混入率 q [%] が大きくなれば, $x \notin U(CP(k))$ の影響が現れて f の分布がいわば痩せてくる。このため z 値や accuracy はあまり変動しないが, 検定適用条件を満足できなくなる傾向となる。また coverage も低下傾向になる。
- 2) “*”について, 混入率 q [%] が大きくなれば, $x \in U(CP(k))$ の影響が現れて f の分布がいわば膨張して, 特定の決定属性値での偏りが低下して, すなわち z 値が低下してルールとして検出できなくなる傾向になる。また accuracy も低下する。一方 coverage は増加傾向になる。

1), 2) の傾向は, 文献 [15] の “?” や “*” の扱い方から当然予測された結果である。さらに 5. と同様に $N_r = 100$ と設定し, q [%] に対する真のルールの導出率 w [%] を Fig. 6 に $N_{miss} = 1, 2, 3$ ごとに示す。Fig. 6 から $N_{lst}(w=100, RL=2) = 1865$ となる最小のデータサイズにもかかわらず, STRIM は真のルールを $w = 95$ [%] 以上で, $N_{miss} = 1$ で $q = 60$ [%], $N_{miss} = 2$ で $q = 30$ [%], $N_{miss} = 3$ で $q = 20$ [%] あたりまで導出していることがわかる。NoiseC による欠損値に対してかなり頑健なルール導出法であることがわかる。

Table 5 Example of influences by “?” or “*” for the rule induction by STRIM in $R(i) = 110000$, $N_{miss} = 1$, $q = 0, 50, 100$

q [%]	missing value	$f : (n1, n2, \dots, n6)$	$p\text{-value}(z)$	accuracy	coverage
0	-	(49, 0, 0, 1, 2, 1)	0(14.99)	0.925	0.151
50	“?”	(32, 0, 0, 1, 3, 0)	0(11.85)	0.888	0.111
	“*”	(57, 18, 20, 20, 19, 18)	0(7.00)	0.375	0.198
100	“?”	(29, 2, 0, 1, 1, 0)	0(29.25)	0.935	0.112
	“*”	(71, 12, 16, 17, 11, 13)	0(10.92)	0.507	0.209

Fig. 7 Experimental results with respect to noise contaminated in decision attribute

6.2 決定属性値が汚染されている場合

Fig. 1 の NoiseD により、決定属性値が汚染されている場合を考察する。このような決定表からラフ集合手法に基づいてルール導出実験した例は、著者らの知る限り、ほとんど報告されていない。具体的には、6.1 と同様に 5. の Step1 をつぎのように修正して、汚染データ実験を実施した。

Step1 : Table 1 ($N = 10000$) から $u(i)$ を $1865 = N_{lst}(w = 100, RL = 2)$ 個無作為に抽出する。この $u(i)$ ($i = 1, \dots, 1865$) の決定属性値が混入率 q [%] で汚染されているか否かを決定する。汚染されている場合は、決定属性値 $\{1, 2, \dots, M_D\}$ から無作為に一つ選び $u^D(i)$ の値を書き換えて新しい決定表を作成する。

この汚染された決定表に STRIM を適用して、 $N_r = 100$ でルール導出性能を検討した結果を Fig. 7 に示す。Fig. 7 から、 $N_{lst}(w = 100, RL = 2)$ となる最小のデータサイズにもかかわらず、STRIM は $w = 95$ [%] 以上の確率で混入率 $q = 60$ [%] あたりまで真のルールを導出していることがわかる。決定属性値が汚染されている場合についても、STRIM はかなり頑健なルール導出法であることがわかる。

7. おわりに

本論文では、文献 [8, 9] での STRIM の今後の課題として残されていた問題：

- 1) 決定属性値が 3 以上で適用できるモデルへ拡張すること、
- 2) ルール導出過程の原理的背景を明瞭化すること、
- 3) データサイズに関する適用条件を明瞭化すること、
- 4) 実データへの適用を検討するうえで、条件属性値に欠損値を含む、あるいは決定属性値が汚染されている場合でのルール導出能力を検討すること、を考察した。

具体的にはまずデータ発生モデルを、観測機構を付加したモデルへ拡張するとともに、具体的に $|V_D| = 6$ の場合でシミュレーションデータを発生した。一方で従来の STRIM のルール導出原理を整理・要約して問題 1) を解決するとともに、発生データを利用してルール導出原理の正当性を確認した。問題 2) ではルール導出手法を整理する過程で、VPRS による accuracy の指標に基づくルール導出法は、STRIM の考え方と同一で、「決定属性値の偏り」の指標に基づく手法とみなせる。しかしこの指標は標本データごとに変動する指標であり、またルール選定基準が明確でないことも指摘した。そこで STRIM の有意水準 z_α に基づいて選定基準を与えることができることを示した。また VPRS に代表される従来法にはデータ発生モデルがないことが、このような問題の背景であることも指摘した。さらに、STRIM のルール候補選択基準の妥当性をデータ発生モデルに基づいて考察した。この考察はモデル選択結果の妥当性の検討にも利用できることを示した。

従来の研究では、真のルールを導出するための必要データサイズの考察がなされていない。本論文では、STRIM が確率 w [%] で真のルールを導出する最小のデータサイズ $N_{lst}(w, RL)$ や、与えられたデータサイズ N_{gvn} に対して真のルールを導出する確率 $w = N^{-1}(N_{gvn}, RL)$ をルール長 RL との関係で考察して、シミュレーション実験でその妥当性を検証した (問題 3)。さらに $N = N_{lst}(w = 100, RL = 2)$ の条件下で、条件属性に欠損値がある場合を欠損数 N_{miss} と混入率 q [%] をパラメータとして、真のルール導出確率 w [%] をシミュレーション実験で検討した。また同様に決定属性値が汚染されている場合も検討した。この結果、STRIM はこのようないわばノイズを含むデータに対して、かなり頑健な手法であるとの知見が得られた (問題 4)。

以上の結果は、あらかじめ設定したルールがすべて $RL = 2$ で、各条件属性の属性値も一様分布する比較的平易な設定での検討であった。今後の課題として、1) ルール長の異なるルールが混在する場合、2) この特別な場合として、完全な分類ルールである場合 (矛盾や無関心データが発生しない場合)、3) 矛盾と無関心データが区分できないような場合などで、各条件属性値の出現率も一様性を仮定しない実問題を想定した設定で、STRIM のルール導出性能を検討することがある。本論文の STRIM はルール候補選択基準として、偏りの基

準 z 値を利用している. 一方統計分野では, モデル選択基準として AIC(Akaike Information Criterion)[16] や MDL(Minimum Description Length)[17] などが知られている. このような基準と STRIM の偏り基準との関係を上記 1), 2), 3) の課題でさらに検討するとともに, STRIM を実測データに適用し, ルール導出結果を吟味して, STRIM の有効性を確認することなどが今後の課題である.

参考文献

- [1] Z. Pawlak: Rough sets; *Internat. J. Inform. Comput. Sci.*, Vol. 11, No. 5, pp. 341–356 (1982)
- [2] A. Skowron and C. M. Rauser: The discernibility matrix and functions in information systems; *Intelligent Decision Support, Handbook of Application and Advances of Rough Set Theory* (R. Słowiński (ed)), Kluwer Academic Publishers, pp. 331–362 (1992)
- [3] Y. G. Bao, X. Y. Du, M. G. Deng and N. Ishii: An efficient method for computing all reducts; *Transaction of the Japanese Society for Artificial Intelligence*, Vol. 19, No. 3, pp. 166–173 (2004)
- [4] J. W. Grzymala-Busse: LERS—A system for learning from examples based on rough sets; *Intelligent Decision Support, Handbook of Applications and Advances of the Rough Sets Theory* (R. Słowiński (ed.)), Kluwer Academic Publishers, pp. 3–18 (1992)
- [5] W. Ziarko: Variable precision rough set model; *Journal of Computer and System Science*, Vol. 46, pp. 39–59 (1993)
- [6] N. Shan and W. Ziarko: Data-based acquisition and incremental modification of classification rules; *Computational Intelligence*, Vol. 11, No. 2, pp. 357–370 (1995)
- [7] T. Nishimura, Y. Kato and T. Saeki: Studies on an effective algorithm to reduce the decision matrix; *RSFDGrC 2011, LNAI*, Vol. 8743, pp. 240–243 (2011)
- [8] T. Matsubayashi, Y. Kato and T. Saeki: A new rule induction method from a decision table using a statistical test; T. Li et al. (Eds.): *RSKT 2012, LNAI*, Vol. 7414, pp. 81–90, Springer (2012)
- [9] 水野, 加藤, 佐伯: 統計的検定法を用いた決定表からのルール導出法; システム制御情報学会論文誌, Vol. 26, No. 8, pp. 297–305 (2013)
- [10] R. E. Walpole, R. H. Myers, S. L. Myers and K. Ye: *Probability and Statistics for Engineers and Scientists*, Eighth edition, Pearson Prentice Hall, pp. 187–191 (2007)
- [11] W. Jaworski: Rule induction: Combining rough sets and statistical approaches; C. -C, Chan et al. (eds.): *RSCTC 2008, LNAI*, Vol. 5306, pp. 170–180, Springer (2008)
- [12] G. Xiw, J. Zhang, K. K. Lai and L. Yu: Variable precision rough set group decision-making: An application; *International Journal of Approximate Reasoning*, Vol. 49, pp. 331–343 (2008)
- [13] M. Inuiguchi, Y. Yoshioka and Y. Kusunoki: Variable-precision dominance-based rough set approach and attribute reduction; *International Journal of Approximate Reasoning*, Vol. 50, pp. 1199–1214 (2009)
- [14] K. Y. Huang, Ting-Hua Chang and Ting-Cheng Chang: Determination of the threshold β of variable precision rough set by fuzzy algorithms; *International Journal of Approximate Reasoning*, Vol. 52, pp. 1056–1072 (2011)
- [15] J. W. Grzymala-Busse and W. J. Grzymala-Busse: Handling missing attribute values; O. Maimon and L. Rokach (eds.), *Data Mining and Knowledge Discovery Handbook*, 2nd ed., Springer, pp. 33–49 (2010)
- [16] H. Akaike: A new look at the statistical model identification; *IEEE Transactions Automatic Control*, Vol. 19, pp. 716–723 (1974)
- [17] J. Rissanen: Modeling by shortest data description; *Automatica*, Vol. 14, pp. 465–471 (1978)

著者略歴

か とう ゆう いち (正会員)
加 藤 裕 一

1979 年 広島大学大学院工学研究科博士課程前期修了. 同年ダイセル化学工業(株)入社. 1983 年 呉工業高等専門学校電気工学科助手. 同講師・助教授を経て, 1996 年より 島根大学総合理工学部教授. 主として, 確率システム・環境システムなどの研究に従事. 京都大学博士(工学). 日本知能情報ファジィ学会, 応用統計学会などの会員.

さ えき てつ ろう (正会員)
佐 伯 徹 郎

1992 年 豊橋技術科学大学大学院工学研究科修士課程修了. 同年 宇部短期大学助手, 1994 年 山口大学工学部助手を経て, 2007 年より 山口大学大学院理工学研究科准教授. 主として, 音環境の計測, 予測, 評価および制御に関する研究に従事. 博士(工学). 1996 年 システム制御情報学会学会賞奨励賞・砂原賞受賞. 日本音響学会, 電子情報通信学会などの会員.

みず の しょう た ろう
水 野 祥 太 郎

2005 年 山口大学大学院理工学研究科知能情報システム工学専攻博士前期課程修了. 同年 株式会社クリック入社, 2008 年退社. 同年より島根大学総合情報処理センター技術職員. 2012 年島根大学総合理工学研究科電子機能システム工学専攻博士後期課程入学. 電子情報通信学会学生会員.