知能と情報（日本知能情報ファジィ学会誌）Vol.32, No.4, pp.759－767（2020）

原著論文

ラフ集合における擬一般化動的縮約の抽出手法の改良†

工藤　康生*1・高橋　智*2・村井　哲也*3

本論文では，高橋らによる大規模な決定表からの擬一般化動的縮約の抽出手法の改良を試みる．高橋らの手法は主に，擬一般化動的縮約の候補の抽出フェーズおよび抽出された候補の確認フェーズの 2 段階で構成されているが，確認フェーズで用いるパラメータ ϵ の適切な事前設定が必要であること，および擬一般化動的縮約となる条件を満たす可能性が低い候補であっても，確認作業を早期に打ち切ることが困難であることなどの課題があった．本論文ではこれらの課題を解決することを目的に，パラメータ ϵ の動的更新および二項検定を用いた確認作業の早期打ち切りを導入することで，擬一般化動的縮約の候補の確認作業を改良することを試みる．更に，抽出された擬一般化動的縮約について，表内に存在しないデータに対する頑健性を検証する．

キーワード：ラフ集合，縮約計算，擬一般化動的縮約

1. はじめに

Pawlak が提唱したラフ集合理論 [1,2] は，考慮する対象全体の集合の部分集合として表される概念に対する集合論的な近似の理論であると共に，表形式データに対する分析手法の数学的基礎も与えている．データ分析手法としてのラフ集合では，分析対象データを表す決定表から，データに内在する規則性を決定ルールと呼ばれる if-then 形式の表現として抽出することが最重要課題の 1 つであり，ルールとして内容が正しく，決定ルールの前提（if-then 形式の if の部分）が短い決定ルールの抽出に関連して，相対縮約の抽出に関する研究が多数行われている（例えば [3–8]）．Bazan et al. は表内に存在しないデータに対してより頑健な分類能力を持つ決定ルールを得ることを目的に，動的縮約（dynamic reduct）およびその一般化である一般化動的縮約（generalized dynamic reduct，以下 GDR と略記）を提案した [9–11]．また，Kudo and Murai は条件属性の個数が多い決定表からできるだけ多数の相対縮約を抽出するアルゴリズムを提案し [12]，このアルゴリズムと Bazan の GDR 抽出手法を組み合わせた手法の基礎部分を提案した [13, 14]．更に高橋らは，Kudo and Murai の手法 [13, 14] を改良し，対象の個数および属性の個数が共に多い大規模な決定表から，GDR と同程度の分類能力を有する属性集合（擬一般化動的縮約，詳細は後述）を抽出する手法を提案した [15, 16].

高橋らの手法は主に，擬一般化動的縮約の候補の抽出フェーズおよび抽出された候補の確認フェーズの 2 段階で構成されている．高橋らの当初の手法 [15] では，確認フェーズにおける計算時間が全体の計算時間のボトルネックとなっていたため，改良手法 [16] では識別行列 [3] を用いた確認手法を用いることで，大幅な計算時間の短縮を実現した．しかし，この改良手法にはなお，確認フェーズで用いるパラメータ ϵ の適切な事前設定が必要であること，および擬一般化動的縮約となる条件を満たす可能性が低い候補であっても，確認作業を早期に打ち切ることが困難であることなどの課題が残されていた．

本論文では，高橋らによる擬一般化動的縮約抽出の改良手法における，上述の 2 つの課題を解決することを目的に，パラメータ ϵ の動的更新および二項検定を用いた確認作業の早期打ち切りを導入することで，擬一般化動的縮約の候補の確認作業を改良することを試みる．更に，抽出された擬一般化動的縮約について，表内に存在しないデータに対する分類能力を検証する．

2. ラフ集合

本節では文献 [17] に基づき，ラフ集合について概略を述べる．

2.1 決定表と相対縮約

ラフ集合によるデータ分析では，分析対象は決定表で表される．決定表は 4 項組 $(U, C \cup D, V, \rho)$ で定義される．ここで，U は有限集合でありその要素を対象と呼ぶ．C は条件属性の有限集合，D は決定属性の有限集合で $C \cap D = \emptyset$ である．V は属性の値の集合であり，各属性 $a \in C \cup D$ の値の集合 V_a を用いて $V = \bigcup_{a \in C \cup D} V_a$ と定義される．$\rho : U \times (C \cup D) \to V$ は対象 $x \in U$ の属性 $a \in C \cup D$ における値 $\rho(x, a) \in V_a$ を表す関数である．

属性の任意の部分集合 $E \subseteq C \cup D$ に対して，以下の 2 項関係

$$IND(E) \stackrel{\text{def}}{=} \{(x, y) \in U \times U \mid \rho(x, a) = \rho(y, a), \forall a \in E\} \quad (1)$$

は対象の集合 U 上の同値関係となり，識別不能関係と呼ばれる．識別不能関係 $IND(E)$ による，すべての対象 $x \in U$ の同値類 $[x]_E$ の集合は，U の分割 $U/IND(E)$ となる．特に，決定属性集合 D による識別不能関係 $IND(D)$ から得られた同値類は

† Improvement of an Extraction Method of Pseudo-Generalized Dynamic Reducts in Rough Sets
Yasuo KUDO, Satoshi TAKAHASHI, and Tetsuya MURAI

*1 室蘭工業大学大学院工学研究科
Graduate School of Engineering, Muroran Institute of Technology
*2 クオリサイトテクノロジーズ株式会社
Qualysite Technologies Inc.
*3 公立千歳科学技術大学理工学部
Faculty of Science and Technology, Chitose Institute of Science and Technology

決定クラスと呼ばれる．決定表 $A = (U, C \cup D, V, \rho)$ の決定クラスの集合を $\mathcal{D}_A \stackrel{\text{def}}{=} U/IND(D)$ と表す．

決定表 A における対象の任意の部分集合 $X \subseteq U$ に対して，属性の任意の部分集合 $E \subseteq C \cup D$ による，X の下近似 $\underline{E}(X)$ と上近似 $\overline{E}(X)$ はそれぞれ以下のように定義される：

$$\underline{E}(X) \stackrel{\text{def}}{=} \{x \in U \mid [x]_E \subseteq X\}, \tag{2}$$

$$\overline{E}(X) \stackrel{\text{def}}{=} \{x \in U \mid [x]_E \cap X \neq \emptyset\}. \tag{3}$$

特に，決定表 A の各決定クラス $D_i \in \mathcal{D}_A$ $(i = 1, \ldots, p)$ に対して，条件属性の部分集合 $R \subseteq C$ による下近似 $\underline{R}(D_i)$ の和集合

$$Pos_R(\mathcal{D}_A) \stackrel{\text{def}}{=} \bigcup_{D_i \in \mathcal{D}_A} \underline{R}(D_i) \tag{4}$$

は，R による決定表 A の正領域と呼ばれる．正領域 $Pos_R(\mathcal{D}_A)$ は，属性集合 R に含まれる条件属性の値を調べることで，決定クラスに正しく分類できる対象の集合である．分割 $\mathcal{D}_A$ に対する，属性集合 $R \subseteq C$ による近似の質 $\gamma_R(\mathcal{D}_A)$ は次式で定義され，属性集合 R による情報により正しく分類できる対象の割合を表す：

$$\gamma_R(\mathcal{D}_A) \stackrel{\text{def}}{=} \frac{\left|\displaystyle\sum_{D_i \in \mathcal{D}_A} \underline{R}(D_i)\right|}{|U|} = \frac{|Pos_R(\mathcal{D}_A)|}{|U|}. \tag{5}$$

ここで，$|X|$ は集合 X の要素数を表す．

分割 $\mathcal{D}_A$ の正領域を用いて，決定表 $A = (U, C \cup D, V, \rho)$ における分割 $\mathcal{D}_A$ に関する相対縮約 $R \subseteq C$ （誤解の恐れがない場合は，決定表 A の相対縮約と呼ぶ）は以下の条件を満たす条件属性の部分集合として定義される：

$$\begin{aligned} &Pos_R(\mathcal{D}_A) = Pos_C(\mathcal{D}_A) \text{ かつ} \\ &\nexists a \in R;\ Pos_{R-\{a\}}(\mathcal{D}_A) = Pos_R(\mathcal{D}_A). \end{aligned} \tag{6}$$

すなわち，決定表 A の相対縮約 $R \subseteq C$ は，A の条件属性をすべて用いた際の対象の分類能力と同等の分類能力を持つ，必要最小限の条件属性のみからなる集合である．一般的に，決定表には複数個の相対縮約が存在しうる．決定表 A のすべての相対縮約の集合を $RED(\mathcal{D}_A)$ と表す．

2.2 識別行列

決定表 A のすべての相対縮約を求める方法として，識別行列 [3] を用いる方法が知られている．決定表 $A = (U, C \cup D, V, \rho)$ の識別行列は，i 行 j 列目の成分 δ_{ij} が次式で定義される $|U| \times |U|$ 行列である：

$$\delta_{ij} = \begin{cases} \{a \in C \mid \rho(x_i, a) \neq \rho(x_j, a)\}, \\ \quad \exists d \in D, \rho(x_i, d) \neq \rho(x_j, d) \text{ かつ} \\ \quad \{x_i, x_j\} \cap Pos_C(\mathcal{D}_A) \neq \emptyset \text{ のとき}, \\ *, \text{ その他}. \end{cases} \tag{7}$$

この定義より，任意の i, j $(1 \leq i, j \leq |U|)$ について，明らかに $\delta_{ii} = *$ かつ $\delta_{ij} = \delta_{ji}$ であるため，識別行列は上三角部分または下三角部分のみ考慮すれば十分である．本論文では下三角部分のみ考慮する．

条件属性の集合 $\delta_{ij}(\neq *)$ は，対象 x_i と x_j の決定クラスが互いに異なり，かつ少なくとも一方は対応する決定クラスに正しく分類可能である状況で，互いに値が異なる条件属性の集合である．よって，少なくとも 1 個の条件属性 $a \in \delta_{ij}$ の値を調べることで，対象 x_i と x_j を区別することが可能であるため，識別行列のすべての要素 $\delta(\neq *)$ と空でない共通部分を持ち，集合の包含関係で極小となる属性集合は，すべて決定表 A の相対縮約となる．

2.3 縮約計算の計算量

識別行列を用いることで，理論的には与えられた決定表のすべての相対縮約を求めることができる．しかし，決定表のすべての相対縮約を求める問題は NP 困難であることが証明されている [3]．よって，ある程度以上の対象件数および属性件数を持つ決定表について，すべての相対縮約を求めることは現実的ではないため，何らかのヒューリスティクスに基づいて相対縮約の候補となる条件属性集合を，少数個または 1 個のみ抽出する手法が多数提案されている（例えば [4–8]）．

3. 一般化動的縮約

Bazan et al. [9] は，ラフ集合分析により得られた決定ルールを表内に存在しないデータに対して適用する際に，そのようなデータに対してより頑健な分類能力を持つ決定ルールを得ることを目的とし，動的縮約（dynamic reduct）およびその一般化である一般化動的縮約（GDR）を提案した．本節では文献 [9–11] に基づき，一般化動的縮約について概略を述べる．

決定表 $A = (U, C \cup D, V, \rho)$ に対して，$U' \subseteq U$ である任意の決定表 $B = (U', C \cup D, V, \rho)$ を決定表 A の部分表（subtable）と呼ぶ．決定表 A のすべての部分表の集合を $P(A)$ とし，その任意の空でない部分集合 $\mathcal{S} \subseteq P(A)$ に対して，以下の条件を満たす条件属性集合 $G \subseteq C$ を $(\mathcal{S}, \epsilon)$-一般化動的縮約（$(\mathcal{S}, \epsilon)$-GDR）と呼ぶ：

$$\frac{|\{B \in \mathcal{S} \mid G \in RED(\mathcal{D}_B)\}|}{|\mathcal{S}|} \geq 1 - \epsilon. \tag{8}$$

ここで，値 ϵ $(0 \leq \epsilon \leq 1)$ は，属性集合 G が相対縮約とならない部分表 $B \in \mathcal{S}$ の割合に関する閾値となるパラメータである．集合 $RED(\mathcal{D}_B)$ は部分表 $B = (U', C \cup D, V, \rho)$ $(U' \subseteq U)$ における，決定クラスの集合 $\mathcal{D}_B \stackrel{\text{def}}{=} U'/IND(D)$ による分割に対するすべての相対縮約の集合である．以上より，$(\mathcal{S}, \epsilon)$-GDR である条件属性集合 $G \subseteq C$ は，与えられた部分表の集合 $\mathcal{S}$ に含まれる部分表の $100 \times (1 - \epsilon)\%$ 以上で相対縮約となる条件属性集合である．$(\mathcal{S}, \epsilon)$-動的縮約は $\epsilon = 0$ の場合に相当し，$\mathcal{S}$ に含まれるすべての部分表で相対縮約となる条件属性集合である．

一般化動的縮約 G に対して，以下の値

$$\frac{|\{B \in \mathcal{S} \mid G \in RED(\mathcal{D}_B)\}|}{|\mathcal{S}|} \tag{9}$$

は G の（$\mathcal{S}$ に対する）安定係数（stability coefficient）と呼ばれる．GDR の安定係数は以下の性質を満たす [10]．この性質は，GDR の表内に存在しないデータに対する頑健性を裏付けている．

定理 1 (最尤推定量としての安定係数) 以下を仮定する：

- $\mathbf{W} = (W, C \cup D, V, \rho)$：すべての可能な対象からなる決定表（*universal decision table*）．ここで，集合 W は対象領域

（*universe of discourse*）である．

- $A = (U, C \cup D, V, \rho)$： 与えられた決定表 ($U \subseteq W$).
- $P(\mathbf{W})$： 母集団 （$\mathbf{W}$ から得られるすべての可能な部分表）.
- $\mathcal{S} \subseteq P(A)$： 与えられた決定表 A から得られた部分表の集合.
- G： ある値 ϵ ($0 \leq \epsilon \leq 1$) の下での $(\mathcal{S}, \epsilon)$-GDR.

このとき，式 (9) で求められる G の安定係数は，すべての可能な部分表の中で G が相対縮約となる部分表の割合

$$\frac{|\{B \in P(\mathbf{W}) \mid G \in RED(\mathcal{D}_B)\}|}{|P(\mathbf{W})|} \tag{10}$$

の最尤推定量である．ここで，$RED(\mathcal{D}_B)$ は部分表 $B = (W', C \cup D, V, \rho)$ ($W' \subseteq W$) における，決定クラスの集合 $\mathcal{D}_B \stackrel{def}{=} W'/IND(D)$ による分割に対するすべての相対縮約の集合である．

定理 1 より，安定係数の値が大きい $(\mathcal{S}, \epsilon)$-GDR ほど，未知の部分表に対して相対縮約となる確率が高い，すなわち表内に存在しないデータに対する頑健性が高いことが期待できるため，安定係数は GDR の有用な評価指標である（定理 1 の証明は Bazan [10] を参照のこと）．

安定係数を用い，信頼係数 $1 - \alpha$ で母比率の区間推定を考える．推定の許容誤差を ΔMLE とすると，部分表の集合 $\mathcal{S}$ が含むべき部分表の個数を次式で見積もることができる [10]：

$$|\mathcal{S}| \geq \frac{t_\alpha^2}{4 \cdot (\Delta MLE)^2}. \tag{11}$$

ここで，値 t_α は標準正規分布表から読み取る値であり，次式を満たす：

$$1 - \alpha = \frac{1}{\sqrt{2\pi}} \int_{-t_\alpha}^{t_\alpha} \exp\left(-\frac{t^2}{2}\right) dt. \tag{12}$$

例えば，信頼係数を $1 - \alpha = 0.9$，推定の許容誤差を $\Delta MLE = 0.05$ とすると，標準正規分布表より $t_\alpha = 1.64$ となり，必要となる部分表の個数として，$|\mathcal{S}| \geq 1.64^2/(4 \cdot 0.05^2) \approx 268.96$ が得られる [10].

4. 大規模データからの擬一般化動的縮約の抽出手法

本節では，擬一般化動的縮約（pseudo-generalized dynamic reduct，以下 pGDR と略記）を導入し，文献 [13–16] に基づき大規模データからの pGDR の抽出手法について概略を述べる．なお，文献 [13–16] では擬一般化動的縮約の略称として GDR を用いている．

4.1 擬一般化動的縮約

3 節で述べたように，与えられた決定表 $A = (U, C \cup D, V, \rho)$ の部分表の集合 $\mathcal{S}$ とパラメータ ϵ の下での $(\mathcal{S}, \epsilon)$-GDR は，集合 $\mathcal{S}$ 内の $100 \times (1 - \epsilon)$% 以上の部分表で相対縮約となる条件属性集合 $G \subseteq C$ である．G が部分表 $B \in \mathcal{S}$ の相対縮約となるためには，相対縮約の定義より，(i) B 内で正しく分類できる対象は G が持つ属性ですべて分類可能であり，(ii) G は冗長な属性を含まないこと，が必要となる．

パラメータ ϵ の値を 0 に近い小さい値に設定すると，$(\mathcal{S}, \epsilon)$-GDR の条件を満たすために $\mathcal{S}$ 内の多数の部分表で相対縮約となることが求められるが，一般的に，多数の部分表で対象を正しく分類するためには多数の属性が必要となりうる一方で，属性の冗長性を排除するためにはできるだけ少数の属性が望ましいため，上述の条件 (i) と条件 (ii) を両立させることは困難と考えられる．これに対し，パラメータ ϵ が 1 に近い値であると，3 節で述べた通り，安定係数が小さい属性集合 G でも $(\mathcal{S}, \epsilon)$-GDR となりえるため，表内に存在しないデータに対する頑健性は疑わしい．

そのため，本論文では GDR の条件を緩和し，集合 $\mathcal{S}$ 内の $100 \times (1 - \epsilon)$% 以上の部分表で上述の条件 (i) を満たす条件属性集合を $(\mathcal{S}, \epsilon)$-擬一般化動的縮約（$(\mathcal{S}, \epsilon)$-pGDR）と呼び，以下のように定義する．

定義 1 決定表 $A = (U, C \cup D, V, \rho)$ の部分表の集合 $\mathcal{S}$ と，パラメータ ϵ ($0 \leq \epsilon \leq 1$) に対して，$(\mathcal{S}, \epsilon)$-擬一般化動的縮約（$(\mathcal{S}, \epsilon)$-pGDR）は，以下の条件を満たす条件属性の部分集合 $G \subseteq C$ である：

$$\frac{|\{B \in \mathcal{S} \mid Pos_G(\mathcal{D}_B) = Pos_C(\mathcal{D}_B)\}|}{|\mathcal{S}|} \geq 1 - \epsilon. \tag{13}$$

ここで，$Pos_X(\mathcal{D}_B)$ は部分表 $B = (U', C \cup D, V, \rho)$ ($U' \subseteq U$) における，決定クラスの集合 $\mathcal{D}_B \stackrel{def}{=} U'/IND(D)$ による分割の，属性集合 $X \subseteq C$ による正領域である．

また，$(\mathcal{S}, \epsilon)$-pGDR である属性集合 $G \subseteq C$ に対して，次式で定義される値を G の（$\mathcal{S}$ に対する）安定係数と呼ぶ：

$$\frac{|\{B \in \mathcal{S} \mid Pos_G(\mathcal{D}_B) = Pos_C(\mathcal{D}_B)\}|}{|\mathcal{S}|}. \tag{14}$$

4.2 大規模データからの pGDR の抽出方法

本節では，大規模データからの pGDR 抽出方法について概略を述べる．

大規模データからの pGDR 抽出手法の全体の流れを Algorithm 1 に示す [14]．この抽出手法ではまず，ステップ 4~17 で，与えられた決定表から作成した部分表から，その相対縮約を pGDR の候補として多数抽出する．次に，ステップ 19 で，抽出した pGDR 候補が式 (13) で定義される pGDR の条件を満たすかを確認することにより，与えられた決定表の pGDR の集合を抽出する．この手法では，部分表の集合 $\mathcal{S}$ とパラメータ ϵ の下で，すべての $(\mathcal{S}, \epsilon)$-pGDR を求めるのではなく，多数生成した pGDR の候補から，pGDR を確実に抽出することを意図している．

ステップ 2 における部分表の作成では，Algorithm 2 により，層化抽出法を用いて与えられた決定表における各決定クラスの要素の割合を反映した部分表を，与えられた決定表から所定の個数作成している．Algorithm 2 による部分表の作成方法により，与えられた決定表 A におけるすべての対象 $x \in U$ に対してある部分表 $A_i \in \mathcal{S}$ が存在し，その対象の集合 U_i について $x \in U_i$ となる．また，作成された各部分表は，決定表 A の各決定クラスから少なくとも 1 個以上の対象が選択されている．

部分表からできるだけ多数の相対縮約を抽出するために，部分表から更にその小規模決定表（reduced decision table）[12] を多数作成し，小規模決定表からその相対縮約を抽出する．このように抽出した相対縮約が，元の部分表の相対縮約となるこ

Algorithm 1 大規模データからの pGDR 抽出アルゴリズム

Require: 決定表 $A = (U, C \cup D, V, \rho)$，パラメータ ϵ，部分表の個数 ST_num，小規模決定表の個数 RDT_num，閾値 th
Ensure: pGDR の集合 $\mathcal{G}$
1: $C \leftarrow \emptyset$, $\mathcal{G} \leftarrow \emptyset$, $\mathcal{S} \leftarrow \emptyset$, $\mathcal{M} \leftarrow \emptyset$
2: $\mathcal{S} \leftarrow$ 決定表 A の部分表 $A_i = (U_i, C \cup D, V, \rho_i)$ $(U_i \subseteq U,\ 1 \leq i \leq ST\_num)$ を ST_num 個作成
3: /* pGDR 候補抽出フェーズ */
4: **for** $i \leftarrow 1$ **to** ST_num **do**
5: $M_i \leftarrow$ 部分表 A_i の識別行列を作成し，冪等律，吸収律を用いて冗長な要素を削除
6: $\mathcal{M} \leftarrow \mathcal{M} \cup \{M_i\}$
7: **for** $j \leftarrow 1$ **to** RDT_num **do**
8: 部分表 A_i の小規模決定表 $B_j = (U_i, C_j \cup D, V, \rho_i^j)$ $(C_j \subseteq C, \rho_i^j : U_i \times C_j \rightarrow V)$ を作成
9: /* C_j の要素数に応じて相対縮約の抽出法を切り替える */
10: **if** $|C_j| \leq th$ **then**
11: $Red \leftarrow$ 小規模決定表 B_j のすべての相対縮約の集合
12: **else**
13: $Red \leftarrow$ 縮約抽出のヒューリスティックアルゴリズムを用いて B_j から抽出した相対縮約の候補の集合
14: **end if**
15: $C \leftarrow C \cup Red$
16: **end for**
17: **end for**
18: /* pGDR 候補確認フェーズ */
19: $\mathcal{G} \leftarrow C$ 内の各候補に対して，許容誤差のパラメータ ϵ の下で，識別行列を用いて確認を行う
20: **return** $\mathcal{G}$

Algorithm 2 層化抽出法を用いた部分表の作成アルゴリズム

Require: 決定表 A = $(U, C \cup D, V, \rho)$，部分表の個数 ST_num
Ensure: 部分表の集合 $\mathcal{S}$
1: $\mathcal{S} \leftarrow \emptyset$
2: /* 各部分表の対象数の目安 OBJ_num を設定する */
3: $OBJ\_num = \left\lceil |U| \times \frac{1}{ST\_num} \right\rceil$
4: i 番目 $(1 \leq i \leq ST\_num)$ の部分表 A_i の対象の集合 U_i を $U_i \leftarrow \emptyset$ で初期化
5: **for** $j \leftarrow 1$ to $|\mathcal{D}_A|$ **do**
6: /* 部分表における各決定クラス D_j の要素数 n_j を定める */
7: $n_j \leftarrow \left\lceil OBJ\_num \times \frac{|D_j|}{|U|} \right\rceil$
8: **end for**
9: **for** $j \leftarrow 1$ **to** $|\mathcal{D}_A|$ **do**
10: $Tmp \leftarrow D_j$
11: **for** $i \leftarrow 1$ **to** ST_num **do**
12: /* A_i の j 番目の決定クラスの要素を U_i に追加する */
13: $U_i^j \leftarrow Tmp$ から n_j 個の対象を非復元抽出で抽出
14: $U_i \leftarrow U_i \cup U_i^j$
15: **if** $Tmp \neq \emptyset$ **and** $|Tmp| < n_j$ **then**
16: $Tmp' \leftarrow D_j \setminus Tmp$ から $n_j - |Tmp|$ 個の対象を非復元抽出で抽出
17: $Tmp \leftarrow Tmp \cup Tmp'$
18: **end if**
19: **if** $Tmp = \emptyset$ **then**
20: $Tmp \leftarrow D_j$
21: **end if**
22: **end for**
23: **end for**
24: /* i 番目の部分表 A_i を作成し S に追加
25: **for** $i \leftarrow 1$ **to** ST_num **do**
26: $A_i \leftarrow (U_i, C \cup D, V, \rho|_{U_i})$ /* $\rho|_{U_i} : U_i \times C \rightarrow V$ */
27: $S \leftarrow S \cup \{A_i\}$
28: **end for**
29: **return** $\mathcal{S}$

とは理論的に保障されている [12]．これにより，部分表における条件属性の個数が多くても，多数の相対縮約を pGDR の候補として抽出することができる．

ステップ 19 における pGDR 候補の確認フェーズでは，Algorithm 3 により，ステップ 5 で作成した各部分表の識別行列を用いて各 pGDR 候補の部分表に対する分類能力を確認し，$(\mathcal{S}, \epsilon)$-pGDR の条件を満たさない候補は Algorithm 3 のステップ 12 で pGDR の候補の集合 C から削除される．

以上より，与えられたパラメータ ϵ に基づき，決定表 A の $(\mathcal{S}, \epsilon)$-pGDR の集合 $\mathcal{G}$ が得られる．

なお，pGDR 候補抽出フェーズおよび pGDR 候補確認フェーズのどちらでも，各部分表 A_i に対する処理は他の部分表 A_j $(j \neq i)$ への処理と独立に実行可能であるため，並列化による計算の高速化を行っている [15]．並列化には OpenMP [18] を使用している．

5. pGDR 候補の確認手法の改良

本節では，Algorithm 3 で示した従来の pGDR 候補確認手法の問題点を 2 点指摘し，これらを解決する新たな pGDR 候補確認手法を提案する．

5.1 従来の確認手法の問題点

3 節で述べた大規模データからの pGDR 抽出手法の問題点として，pGDR の候補に対する確認作業における，以下の 2 点が挙げられる：

1. パラメータ ϵ の適切な設定が困難．
2. 確認作業における計算の冗長性．

従来の確認手法では，$(\mathcal{S}, \epsilon)$-pGDR を抽出する際に，パラメータ ϵ の値を予め設定する必要があり，pGDR の抽出・確認作業の途中で ϵ の値を変更することは不可能であった．そのため，与えられた決定表に対して適切でないパラメータ ϵ を設定すると，$(\mathcal{S}, \epsilon)$-pGDR をまったく抽出できない場合が存在した．

更に，従来の確認手法では，$(\mathcal{S}, \epsilon)$-pGDR の条件を満たさない候補を確実に排除するため，すべての pGDR の候補 $G \in C$ に対して，G ですべての分類可能な対象を正しく分類することができなかった部分表が $\epsilon \cdot |\mathcal{S}|$ 個以上になるか，または $\mathcal{S}$ のすべての部分表を確認し終えるまで，候補 G に対する確認作業を続ける．よって，各候補について少なくとも $\epsilon \cdot |\mathcal{S}|$ 個の部分表に対する確認作業を行うため，パラメータ ϵ $(0 \leq \epsilon < 1)$ の値を 1 に近い値に設定すると，候補ごとに多数の部分表に対する確認作業が必要となる．しかし，実際には pGDR の候補の大半は確認作業の早い段階で，ごく少数個の部分表でしか対象を正しく分類できず，$(\mathcal{S}, \epsilon)$-pGDR となる条件を明らかに

Algorithm 3 識別行列を用いた pGDR 候補の確認アルゴリズム

Require: pGDR の候補の集合 C，部分表の識別行列の集合 $\mathcal{M}$，パラメータ ϵ
Ensure: pGDR の集合 $\mathcal{G}$

```
G ← C
for i ← 1 to |C| do
    e ← 0    /* i 番目の pGDR 候補の誤分類数を初期化 */
    for j ← 1 to |M| do
        /* 部分表 A_j の識別行列 M_j を用いて，候補 G_i が A_j の
           要素を正しく分類できるか確認 */
        for k ← 1 to |M_j| do
            /* 識別行列の要素 δ_k ∈ M_j と候補 G_i に共通部分がな
               ければ，G_i は A_j の要素を正しく分類できない */
            if δ_k ≠ * and δ_k ∩ G_i = ∅ then
                e ← e + 1
                if e ≥ ε · |S| then
                    /* G_i は pGDR の条件を満たせないので除去 */
                    G ← G \ {G_i}
                    break
                end if
            end if
        end for
    end for
end for
return G
```

満たさないと予測できることが多い．そのため，このような候補に対して確認作業を打ち切らず最後まで行うことは，計算リソースの観点から無駄が多いとみなすことができる．

以上を踏まえ，本節で提案する新たな pGDR 候補確認手法では，パラメータ ϵ の値を動的に更新しつつ，$(\mathcal{S},\epsilon)$-pGDR となる条件を満たす可能性が低い pGDR の候補に対する確認作業を早期に打ち切ることを行う．

5.2 パラメータ ϵ の動的更新

提案手法では，パラメータ ϵ の初期値を 1 に近い値とし，微小な非負の値 $\Delta\epsilon$ を ϵ の更新幅として設定する．ある pGDR の候補 $G \in C$ が現時点でのパラメータ ϵ の下で $(\mathcal{S},\epsilon)$-pGDR となる条件を満たしたとき，G が対象を正しく分類できた部分表の個数 n が以下の条件

$$\frac{n}{|\mathcal{S}|} \geq 1-(\epsilon-\Delta\epsilon) \tag{15}$$

を満たすならば，$\epsilon-\Delta\epsilon<\epsilon$ であることから，G は対象を正しく分類できた部分表の個数がより多い $(\mathcal{S},\epsilon-\Delta\epsilon)$-pGDR となる条件を満たしているとみなせる．よって，これまでに抽出された G 以外の $(\mathcal{S},\epsilon)$-pGDR をすべて破棄し，次の候補 $G' \in C$ からは $(\mathcal{S},\epsilon-\Delta\epsilon)$-pGDR となる条件を満たすか確認するために，パラメータ ϵ を $\epsilon-\Delta\epsilon$ に更新する．

5.3 二項検定を用いた確認作業の早期打ち切り

片側二項検定による母比率の検定を用いて，$(\mathcal{S},\epsilon)$-pGDR となる条件を満たす可能性が低い候補について確認作業を早期に打ち切ることにより，確認作業に要する計算リソースの無駄を減らすことを試みる．

ある $(\mathcal{S},\epsilon)$-pGDR の候補 $G \in C$ について，$\mathcal{S}$ に含まれる部分表全体に対する，G がすべての対象を正しく分類できる部分表の割合を母比率 P_G とし，帰無仮説 $H_0: P_G = 1-\epsilon$ および対立仮説 $H_1: P_G < 1-\epsilon$，有意水準 α を設定する．帰無仮説 H_0 は pGDR の候補 G が $(\mathcal{S},\epsilon)$-pGDR の条件を満たすことを，対立仮説 H_1 は条件を満たさないことをそれぞれ表す．

pGDR の候補 G と $\mathcal{S}$ に含まれる部分表について，G で対象を正しく分類できるか確認を終えた部分表の個数を m，その中で G が実際に対象を正しく分類できた部分表の個数を n とする．帰無仮説 H_0 の下で，m 個の部分表を確認した時点で G が n 個の部分表においてすべての対象を正しく分類できる確率 $\widehat{P}_G^m$ は，二項分布 $B(m,1-\epsilon)$ を用いて次式で表すことができる：

$$\widehat{P}_G^m = {}_mC_n(1-\epsilon)^n\epsilon^{m-n}. \tag{16}$$

確率 $\widehat{P}_G^m$ は組み合わせの数 ${}_mC_n$ を含むため，一般的には m や n の値が大きい場合は計算が困難である．しかし，提案手法では $\mathcal{S}$ に含まれる部分表に対して 1 個ずつ確認作業を行うため，確率 $\widehat{P}_G^m$ の値は以下の漸化式で逐次的に更新することができる：

$m+1$ 個目の部分表で G が対象を正しく分類できる場合

$$\begin{aligned}\widehat{P}_G^{m+1} &= {}_{m+1}C_{n+1}(1-\epsilon)^{n+1}\epsilon^{(m+1)-(n+1)}\\ &= \frac{(m+1)!}{(m-n)!(n+1)!}(1-\epsilon)^{n+1}\epsilon^{(m+1)-(n+1)}\\ &= \frac{m+1}{n+1}(1-\epsilon)\cdot\frac{m!}{(m-n)!n!}(1-\epsilon)^n\epsilon^{(m-n)}\\ &= \frac{m+1}{n+1}(1-\epsilon)\cdot\widehat{P}_G^m.\end{aligned} \tag{17}$$

$m+1$ 個目の部分表で G が対象を正しく分類できない場合

$$\begin{aligned}\widehat{P}_G^{m+1} &= {}_{m+1}C_n(1-\epsilon)^n\epsilon^{(m+1)-n}\\ &= \frac{(m+1)!}{((m+1)-n)!n!}(1-\epsilon)^n\epsilon^{(m+1)-n}\\ &= \frac{m+1}{m-n+1}\epsilon\cdot\frac{m!}{(m-n)!n!}(1-\epsilon)^n\epsilon^{(m-n)}\\ &= \frac{m+1}{m-n+1}\epsilon\cdot\widehat{P}_G^m.\end{aligned} \tag{18}$$

$\mathcal{S}$ に含まれる部分表に対して確認作業を行うごとに，式 (17) または式 (18) を用いて確率 $\widehat{P}_G$ を更新する．G による実際の結果について $n/m<1-\epsilon$ であり，かつ確率 $\widehat{P}_G$ が有意水準 α より小さくなった時点で，片側二項検定の結果として帰無仮説 H_0 は棄却され，対立仮説 H_1 が採択されるため，pGDR の候補 G は $(\mathcal{S},\epsilon)$-pGDR の条件を満たさないと判断することができる．これにより，実際には $(\mathcal{S},\epsilon)$-pGDR の条件を満たさない可能性が高い候補について，確認作業を行った部分表の個数 m が $\epsilon\cdot|\mathcal{S}|$ より少なくても，確認作業を打ち切ることが可能となる．一方，確率 α で，実際には $(\mathcal{S},\epsilon)$-pGDR である候補 G について，確認作業を打ち切ってしまい抽出できなくなる可能性がある．

Algorithm 4 改良した pGDR 候補の確認アルゴリズム

Require: pGDR の候補の集合 C，部分表の識別行列の集合 $\mathcal{M}$，パラメータの初期値 ϵ，ϵ の更新幅 $\Delta\epsilon$，有意水準 α
Ensure: pGDR の集合 $\mathcal{G}$

1: $\mathcal{G} \leftarrow \emptyset$
2: **for** $i \leftarrow 1$ **to** $|C|$ **do**
3: $m \leftarrow 0, n \leftarrow 0, prec \leftarrow 1.0$
4: **for** $j \leftarrow 1$ **to** $|\mathcal{M}|$ **do**
5: /* 部分表 A_j の識別行列 M_j を用いて，候補 G_i が A_j の要素を正しく分類できるか確認 */
6: flagDisernibility $\leftarrow$ *true*
7: $m \leftarrow m + 1$
8: **for** $k \leftarrow 1$ **to** $|M_j|$ **do**
9: /* 識別行列の要素 $\delta_k \in M_j$ と候補 G_i に共通部分がなければ，G_i は A_j の要素を正しく分類できない */
10: **if** $\delta_k \neq *$ **and** $\delta_k \cap G_i = \emptyset$ **then**
11: /* 式 (18) で確率 *prec* を更新 */
12: $prec \leftarrow \dfrac{m+1}{m-n+1}\epsilon \cdot prec$
13: flagDisernibility $\leftarrow$ *false*
14: break /* A_j の確認を打ち切る */
15: **end if**
16: **end for**
17: **if** flagDisernibility = *true* **then**
18: /* A_j の要素を正しく分類できているので，式 (17) で確率 *prec* を更新 */
19: $n \leftarrow n + 1$
20: $prec \leftarrow \dfrac{m+1}{n+1}(1-\epsilon) \cdot prec$
21: **end if**
22: /* 帰無仮説 H_0 を棄却してよいか判断 */
23: **if** $\dfrac{n}{m} < 1 - \epsilon$ **and** $prec < \alpha$ **then**
24: /* G_i は pGDR ではないと判断し，G_i の確認作業を打ち切る */
25: break
26: **end if**
27: **end for**
28: /* ϵ を更新できるか判断 */
29: **if** $\dfrac{n}{m} \geq 1 - (\epsilon - \Delta\epsilon)$ **then**
30: $\epsilon \leftarrow \epsilon - \Delta\epsilon$
31: $\mathcal{G} \leftarrow \emptyset$
32: **end if**
33: /* G_i を $(\mathcal{S}, \epsilon)$-pGDR として $\mathcal{G}$ に追加 */
34: $\mathcal{G} \leftarrow \mathcal{G} \cup \{G_i\}$
35: **end for**
36: **return** $\mathcal{G}$

5.4 提案手法のアルゴリズム

提案手法のアルゴリズムを Algorithm 4 に示す．このアルゴリズムではステップ 8 からステップ 21 で，5.3 節で述べた手法を用いて，式 (17) または式 (18) による確率 $\widehat{P}_G$ の更新を行い，ステップ 23 で pGDR の候補に対する確認作業を打ち切るかを判定している．また，ステップ 29 からステップ 32 で，5.2 節で述べた手法を用いて，パラメータ ϵ の更新条件を満たした場合に ϵ を $\epsilon - \Delta\epsilon$ に更新している．その結果，Algorithm 4 はアルゴリズム中で最後に設定されたパラメータ ϵ による，$(\mathcal{S}, \epsilon)$-pGDR の集合 $\mathcal{G}$ を出力する．

表 1 使用したデータセット

データセット名	対象数	属性数	決定クラス数
covtype	581012	55	7
diabetes	101766	49	4
GISETTE	6000	5001	2
Internet-Advertise（IA）	3279	1558	2
Nomao	34465	119	2

表 2 パラメータ

部分表の個数 ST_num	300
小規模決定表の個数 RDT_num	100
縮約抽出手法の切り替えの閾値 th	20
ϵ の初期値（Algorithm 4）	0.9
ϵ の更新幅 $\Delta\epsilon$（Algorithm 4）	0.1
有意水準 α（Algorithm 4）	0.1

6. 評価実験

6.1 評価実験の概要

識別行列を用いた pGDR 候補の確認手法（Algorithm 3）に替えて提案手法（Algorithm 4）を用いることによる，pGDR 候補確認フェーズでの計算コスト改善の効果を検証するため，評価実験を行った．また，抽出された pGDR の表内に存在しないデータに対する分類能力を検証した．

従来手法による pGDR 抽出法と提案手法による抽出法をそれぞれ実装し実験を行った．実験環境は CPU: Inter(R) Xeon(R)CPU E5-2650 v2@2.60GHx(16cores)，Memory: 132GB，OS: CentOS release 6.5 である．また，使用言語は C++(gcc 4.4.7) である．なお，この実装では，Algorithm 1 のステップ 13 で用いる縮約抽出のヒューリスティックアルゴリズムとして，識別行列を用いて相対縮約の候補を 1 個抽出する Zhang et al. のアルゴリズム [8] を使用している．

6.2 評価実験 1

従来手法による pGDR 抽出法と提案手法による抽出法について，各データセットに対して，確認部分および全体の計算に要する時間を比較した．使用したデータセットを表 1 に示す．データセットは UCI Machine Learning Repository [19] から取得した．また，実験で用いたパラメータを表 2 に示す．従来手法ではデータセット毎にパラメータ ϵ の値を予め適切に設定する必要があるため，データセット毎に予備実験を行い，各データセットのパラメータ ϵ の値を求めた．

結果を表 3 に示す．なお，小数点以下は切り捨てた．また，1 秒に満たない部分は $*$ とした．

6.3 評価実験 2

提案手法を用いて抽出された pGDR の，表内に存在しないデータに対する分類能力を検証する実験を行った．実験は 10 分割交差検定を用い，以下の手順で行った：

1. 決定表 $A = (U, C \cup D, V, \rho)$ の対象の集合を 10 分割し，集合 $U_1, \ldots, U_{10}$ を用いて，集合 U_i を対象の集合とする評価用の決定表 A_i および集合 $U \setminus U_i$ を対象の集合とする

表 3 計算時間（秒）

	従来手法		提案手法	
	確認部分	全体	確認部分	全体
covtype	4	2874	*	2037
diabetes	457	1091	36	688
IA	19	56	2	39
GISETTE	61	162	6	107
Nomao	32	84	6	74

表 4 抽出された pGDR の結果

	ϵ	pGDR 候補の個数	pGDR の個数	評価値（訓練）	評価値（テスト）
covtype	0.138	10155.9	7.6	1.00	0.995
diabetes	0.1	86375.2	9.4	1.00	0.985
GISETTE	0.372	10127.2	12.1	0.92	0.751
IA	0.1	21689.2	2051.8	0.988	0.753
Nomao	0.1	89952.0	30.8	0.999	0.97

pGDR 抽出用の決定表 A_{-i} を作成する．

2. 4 節で述べた手法および 5 節で提案した pGDR の確認手法を用いて，決定表 A_{-i} から作成した部分表の集合 $\mathcal{S}_{-i}$ における $(\mathcal{S}_{-i}, \epsilon)$-pGDR を抽出する．
3. 2. で抽出した $(\mathcal{S}_{-i}, \epsilon)$-pGDR から，正しく分類できた部分表の個数が多い上位 10 件の $(\mathcal{S}_{-i}, \epsilon)$-pGDR を選択する．件数が 10 件に満たない場合はすべて選択する．
4. 選択された $(\mathcal{S}_{-i}, \epsilon)$-pGDR G について，各部分表での分割による近似の質（式 (5)）の平均値 $m\gamma_G(\mathcal{S}_{-i})$ を以下の式 (19) で求める：

$$m\gamma_G(\mathcal{S}_{-i}) = \frac{\sum_{B \in \mathcal{S}_{-i}} \gamma_G(\mathcal{D}_B)}{|\mathcal{S}_{-i}|}. \tag{19}$$

5. 4. で求めた値を訓練データに対する各 $(\mathcal{S}_{-i}, \epsilon)$-pGDR の評価値とし，上位 3 件の $(\mathcal{S}_{-i}, \epsilon)$-pGDR を選択する．件数が 3 件に満たない場合はすべて選択する．
6. 5. で選択した各 $(\mathcal{S}_{-i}, \epsilon)$-pGDR G について，評価用の決定表 A_i での，分割による近似の質 $\gamma_G(\mathcal{D}_{A_i})$ を式 (5) で求め，その平均値をテストデータに対する pGDR 適用結果の評価値とする．
7. 上記の 1. ~ 6. を $1 \leq i \leq 10$ で実行し，10 回試行での各評価値の平均値を求める．

上述の手順 2 で得られた pGDR の個数の平均値および pGDR 候補の個数の平均値，pGDR のパラメータ ϵ の平均値，手順 5 で最終的に選択された最大 3 件の pGDR による訓練データでの式 (19) による評価値の平均値，手順 6 におけるテストデータに対する式 (5) による評価値の平均値をそれぞれ求めた．結果を表 4 に示す．実験で用いたパラメータは表 2 と同様である．なお，精度に関しては小数点第 4 位以降を切り捨てた．

6.4 考察

まず，評価実験 1 の結果について考察を行う．表 3 から，提案手法を用いることで確認部分を 5 倍から 10 倍程度に高速化できたことが確認できる．その理由として，確認の途中打ち切りの効果が考えられる．一般的に，pGDR の候補は大量に存在するものの，実際に pGDR の条件を満たす候補はごく少数である（表 4 の項目「pGDR 候補の個数」と「pGDR の個数」からもこの傾向が読み取れる）．従来手法では，大量に存在する pGDR 候補のそれぞれに対して，少なくとも（部分表の個数 $\times \epsilon$）個の部分表について必ず確認の計算を行う必要があった．提案手法では，これらの候補の大半について，確認作業を早期で打ち切ることにより，確認フェーズを高速化できていると考えられる．

次に，評価実験 2 について考察を行う．表 4 から，covtype データセットおよび diabetes データセット，Nomao データセットでは，抽出された pGDR は表内に存在しないデータに対しても高い精度を保ち，頑健な分類能力を有することが確認できる．実験では部分表を 300 個生成しているため，これら 3 種類のデータセットから作成された個々の部分表は含む対象の個数が十分多く（3 種類のデータセットでは最も対象の個数が少ない Nomao データセットでも，Algorithm 2 で用いる対象の個数の目安 OBJ_num は $\lceil(34465 - 3447)/300\rceil = 104$ である），データセット全体の特徴がある程度反映されていたため，抽出された pGDR もデータセット全体の分類に有用な属性を含んでいた結果，表内に存在しないデータに対しても高い分類能力を有することができたと考えられる．

一方，GISETTE データセットおよび IA データセットでの，表内に存在しないデータに関する評価値は，他のデータセットに比べて低い．表 1 より，これらのデータセットは他の 3 種類のデータセットと比べて，対象の個数が少ないデータセットである．各部分表の対象の個数の目安 OBJ_num は，GISETTE データセットでは $\lceil(6000 - 600)/300\rceil = 18$，IA データセットでは $\lceil(3279 - 328)/300\rceil = 10$ となり，作成される各部分表における対象の個数が非常に少ないことがわかる．そのため，個々の部分表は，データセット全体の特徴を反映しているとは見なし難く，これらのデータセットから抽出された pGDR は，データセット全体について分類能力が高い属性を必ずしも含んでいない可能性がある．よって，そのような pGDR は，大半の部分表に対しては対象を正しく分類できていても，データセット全体の特徴を捉え切れていないため，表内に存在しないデータに対しては分類能力が高くない結果となったと考えられる．GISETTE データセットや IA データセットのように，データの個数が多くないデータセットからの部分表の作成方法は，今後の課題とする．

更に，covtype データセットおよび GISETTE データセットでは，パラメータ ϵ の平均値がそれぞれ 0.138 および 0.372 であった．これは，ϵ の値を予め固定する必要がある従来手法で，例えば $\epsilon = 0.1$ と設定して実行した場合，pGDR の抽出に失敗するケースが存在したことを表す．このことから，5.2 節で提案したパラメータ ϵ の動的更新手法の有効性が示唆される．

7. まとめ

本論文では，高橋ら [16] による擬一般化動的縮約抽出の改良手法に対して，確認フェーズで用いるパラメータ ϵ の動的更

新と，二項検定を用いた確認作業の早期打ち切りを導入することで，高橋らの手法に内在する 2 つの課題の解決を試みた．また，抽出された擬一般化動的縮約について，表内に存在しないデータに対する頑健性を検証した．計算機実験の結果から，本論文で導入した手法を用いることで，パラメータ ϵ のデータセットに応じた適切な事前設定が不要となり，また確認作業の早期打ち切りによる計算時間の更なる高速化が確認された．更に，対象の個数が多いデータセットから抽出された擬一般化動的縮約は，表内に存在しないデータに対しても頑健であり，高い分類能力を持つことが確認された．

今後の課題として，対象の個数が比較的少ないデータセットに対する部分表の作成方法の改良，従来手法と提案手法での擬一般化動的縮約の抽出性能の比較，および実データを含めた多様なデータセットを用いた，提案手法の幅広い検証などが挙げられる．

謝辞

本研究は JSPS 科研費 JP25330315 の助成を受けたものです．

参 考 文 献

[1] Z. Pawlak: “Rough Sets,” *Int. J. of Computer and Information Sciences (IJCIS)*, Vol.11, pp. 341-356, 1982.
[2] Z. Pawlak: *Rough Sets: Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers, 1991.
[3] A. Skowron and C. M. Rauszer: “The discernibility matrices and functions in information systems,” in *Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory*, R. Słowiński ed., Kluwer Academic Publishers, pp. 331-362, 1992.
[4] A. Chouchoulas and Q. Shen: “Rough Set-Aided Keyword Reduction for Text Categorization,” *Applied Artificial Intelligence*, Vol.15, No.9, pp. 843-873, 2001.
[5] K. Hu, L. Diao, Y. Lu, and C. Shi: “A Heuristic Optiaml Reduct Algorithm,” *Intelligent Data Engineering and Automated Learning – IDEAL 2000: 2nd Int. Conf. on Intelligent Data Engineering and Automated Learning, Data Mining, Financial Engineering, and Intelligent Agents IDEAL 2000 Proc.*, Dec. 13-15, Hong Kong, pp. 139-144, 2000.
[6] F. Hu, G. Wang, and L. Feng: “Fast Knowledge Reduction Algorithms Based on Quick Sort,” *Rough Sets and Knowledge Technology: 3rd Int. Conf. RSKT 2008 Proc.*, May 17-19, Chengdu, pp. 72-79, 2008.
[7] Y. Kudo and T. Murai: “Heuristic Algorithm for Attribute Reduction Based on Classification Ability by Condition Attributes,” *J. Adv. Comput. Intell. Intell. Inform.*, Vol.15, No.1, pp. 102-109, 2011.
[8] J. Zhang, J. Wang, D. Li, H. He, and J. Sun: “A New Heuristic Reduct Algorithm Based on Rough Sets Theory,” *Advances in Web-Age Information Management: 4th Int. Conf. on Web-Age Information Management WAIM 2003 Proc.*, Aug. 17-19, Chengdu, pp. 247-253, 2003
[9] J. G. Bazan, A. Skowron, and P. Synak: “Dynamic Reducts as a Tool for Extracting Laws from Decisions Tables,” *Methodologies for Intelligent Systems: 8th Int. Symp. on Methodologies for Intelligent Systems ISMIS’94 Proc.*, Oct. 16-19, North Carolina, pp. 346-355, 1994.
[10] J. G. Bazan: “Dynamic Reducts and Statistical Inference,” *Proc. of the 6th Int. Conf., Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU’96)*, Jul. 1-5, Granada, pp. 1-5, 1996.
[11] J. G. Bazan: “A Comparison of Dynamic and Non-Dynamic Rough Set Methods for Extracting Laws from Decision Tables,” in *Rough Sets in Knowledge Discovery*, L. Polkowski ed., Physica-Verlag, pp. 321-365, 1998.
[12] Y. Kudo and T. Murai: “An Attribute Reduction Algorithm by Switching Exhaustive and Heuristic Computation of Relative Reducts,” *Proc. of the 2010 IEEE Int. Conf. on Granular Computing (GrC 2010)*, Aug. 14-16, San Jose, pp. 265-270, 2010.
[13] 工藤康生, 村井哲也: “ラフ集合および統計的手法に基づく大規模データからの縮約抽出に関する一考察,” 日本知能情報ファジィ学会第 28 回ファジィシステムシンポジウム講演論文集, 9 月 12-14 日, 名古屋, pp. 759-760, 2012.
[14] Y. Kudo and T. Murai: “An Attempt of Hybridization of Generalized Dynamic Reducts and A Heuristic Attribute Reduction Using Reduced Decision Tables,” *Proc. of IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE 2013)*, Jul. 7-10, Hyderabad, 2013.
[15] 高橋智, 工藤康生, 村井哲也: “ラフ集合における Generalized Dynamic Reduct の並列抽出,” 第 25 回ソフトサイエンスワークショップ講演論文集, 3 月 10-11 日, 下関, 2015.
[16] 高橋智, 工藤康生, 村井哲也: ラフ集合における Generalized Dynamic Reduct の抽出手法の改良, “日本知能情報ファジィ学会第 31 回ファジィシステムシンポジウム講演論文集,” 9 月 2-4 日, 調布, pp. 773-776, 2015.
[17] 森典彦, 田中英夫, 井上勝雄編: ラフ集合と感性―データからの知識獲得と推論, 海文堂出版, 2004.
[18] OpenMP: http://openmp.org/ [accessed Sep. 15, 2014]
[19] UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/ [accessed Oct. 1, 2012]

（2019 年 8 月 27 日　受付）
（2019 年 10 月 2 日　採録）

［問い合わせ先］
〒050-8585　北海道室蘭市水元町 27-1
室蘭工業大学大学院工学研究科しくみ解明系領域
工藤　康生
E-mail: kudo@csse.muroran-it.ac.jp

著 者 紹 介

工藤（くどう）　康生（やすお）［正会員］

2000 年北海道大学大学院工学研究科博士後期課程修了．同年室蘭工業大学サテライト・ベンチャー・ビジネス・ラボラトリー博士研究員．2003 年室蘭工業大学工学部助手．2009 年同大学大学院工学研究科助教．2010 年同准教授．2016 年同教授，現在に至る．博士（工学）．ラフ集合理論，非古典論理，データマイニング，情報推薦システムなどの研究に従事．日本知能情報ファジィ学会，人工知能学会，日本感性工学会，IEEE などの会員．

高橋（たかはし）　智（さとし）［非会員］

2015 年室蘭工業大学工学部情報電子工学系学科卒業．2017 年室蘭工業大学大学院工学研究科情報電子工学系専攻修士課程修了．同年，クオリサイトテクノロジーズ株式会社入社，現在に至る．在学中はラフ集合の縮約計算に関する研究に従事．

村井 哲也［正会員］

1987 年北海道大学大学院情報工学専攻博士後期課程中途退学後，札幌医科大学衛生短期大学部，北海道教育大学函館校，北海道大学大学院工学研究科・情報科学研究科を経て，2016 年より公立千歳科学技術大学理工学部情報システム工学科教授．1995 年 博士（工学）（北海道大学）．日本知能情報ファジィ学会，人工知能学会，日本感性工学会各会員．

Improvement of an Extraction Method of Pseudo-Generalized Dynamic Reducts in Rough Sets

by

Yasuo KUDO, Satoshi TAKAHASHI, and Tetsuya MURAI

Abstract:

In this paper, we improve Takahashi et al.'s method for extracting pseudo-generalized dynamic reducts (pGDRs) from a decision table with numerous objects and attributes. Takahashi et al.'s method consists of pGDR candidates extraction phase and pGDR confirmation phase using training datasets. However, a parameter ϵ used in the confirmation phase is required to set appropriately before starting the confirmation phase. Moreover, it is difficult to interrupt the confirmation processes for a pGDR candidate G even though it is expected that G does not satisfy the condition of pGDR. To solve these two issues, a dynamic update method of the parameter ϵ and an interruption method of the confirmation processes based on binomial test are introduced to the confirmation phase. Moreover, robustness of the extracted pGDRs to test datasets is examined.

Keywords: rough set, attribute reduction, pseudo-generalized dynamic reduct

Contact Address: **Yasuo KUDO**
Muroran Institute of Technology
27-1 Mizumoto, Muroran, Hokkaido 050-8585, Japan
E-mail: kudo@csse.muroran-it.ac.jp