条件属性による類別に基づく相対縮約の評価手法について

条件属性による類別に基づく相対縮約の評価手法について

その他（別言語等）のタイトル	On an Evaluation Method of Relative Reducts Based on Classification by Condition Attributes
著者	中浦　孝仁，工藤　康生，村井　哲也
雑誌名	ファジィシステムシンポジウム講演論文集
巻	25
発行年	2009-07
URL	http://hdl.handle.net/10258/2201

doi: info:doi/10.14864/fss.25.0.107.0

条件属性による類別に基づく相対縮約の評価手法について

On an Evaluation Method of Relative Reducts Based on Classification by Condition Attributes

中浦 孝仁　　工藤 康生　　村井 哲也
Takahito Nakaura　Yasuo Kudo　Tetsuya Murai
室蘭工業大学　　北海道大学
Muroran Institute of Technology　Hokkaido University

Abstract: In this paper, we propose a new evaluation method of relative reducts based on classification ability of condition attributes. Extraction of relative reducts and decision rules are important elements of the data analysis by rough set theory. As an evaluation method of relative reducts, the authors have proposed an evaluation criterion of relative reducts based on partitions by equivalence classes. However, this method needs to construct equivalence classes for evaluating relative reducts, and moreover, does not evaluate each condition attribute. The new method proposed in this paper evaluates relative reducts by combining evaluation results of classification ability of condition attributes.
Keywords： relative reduct, conditon attribute, classfication

1　はじめに

決定表に対する相対縮約および決定ルールの抽出は，ラフ集合 [4] によるデータ解析の重要な要素である．相対縮約の分析手法の一つとして，著者らは同値類による分割を用いた相対縮約の評価手法を提案した [2]．この手法では，相対縮約から得られる決定ルールの被覆度の平均値を用いて，相対縮約を評価する．しかし，評価の際に相対縮約から対象の同値類を作る必要があり，また，個々の条件属性の評価を行っていない．そこで，本研究では，個々の条件属性による類別能力の評価を用いて，相対縮約を評価する手法を提案する．

2　ラフ集合

文献 [3] に基づき，ラフ集合に関する基礎的な内容について簡略に説明する．

2.1　決定表と相対縮約

ラフ集合によるデータ解析では，複数の属性とそれらの値から与えられる決定表を対象とすることが多い．決定表は以下の 4 項組で定義される．

$$(U, C \cup D, V, \rho). \tag{1}$$

ここで，U は対象の有限集合，C は条件属性の集合，D は決定属性の集合，V は各属性 $a \in C \cup D$ が取り得る値の集合，$\rho : U \times C \cup D \to V$ は対象 $x \in U$ の属性 a における値 $\rho(x, a) \in V$ を表す関数である．任意の属性の部分集合 $B \subseteq C \cup D$ に対して，対象間の識別不能関係 R_B を次式で定義する．

$$xR_By \Leftrightarrow \forall a \in B, \quad \rho(x, a) = \rho(y, a) \tag{2}$$

表 1: 決定表の例

	c1	c2	c3	c4	c5	c6	d
x1	1	1	1	1	1	2	1
x2	2	2	1	1	1	2	1
x3	2	3	2	1	2	1	2
x4	2	2	2	2	2	1	2
x5	2	2	3	1	1	2	1
x6	1	2	1	1	2	2	3

識別不能関係は同値関係となる．xR_By は B に含まれる属性のみでは x と y を区別することができないことを表す．決定属性は全体集合 U の分割 $\{D_1, \ldots, D_m\}$ を与え，各 D_i は決定クラスと呼ばれる．

決定表の例を表 1 に示す．表 1 において，U={x1, . . . , x6}, C={c1,. . .,c6}, D={d}, V={1,2,3} であり，$\rho(x, a)$ は ρ(x1,c1)=1 のように，x を行，a を列に対応させ，求められる表中の値により定められる．また，決定クラスは D_1={x1,x2,x5},D_2={x3,x4},D_3={x6} の 3 個となる．

R_B を条件属性の部分集合 $B \subseteq C$ に基づく識別不能関係とする。各決定クラス D_i に対して，識別不能関係 R_B による下近似 $\underline{B}(D_i)$ および上近似 $\overline{B}(D_i)$ をそれぞれ以下の通り定義する．

$$\underline{B}(D_i) = \{x \in U | [x]_B \subseteq D_i\}, \tag{3}$$

$$\overline{B}(D_i) = \{x \in U | [x]_B \cap D_i \neq \phi\}. \tag{4}$$

ここで，$[x]_B$ は対象 x の識別不能関係 R_B による同値類である．下近似 $\underline{B}(D_i)$ は，属性集合 B に含まれる属性の値のみで確実に決定クラス D_i に分類できる対象の集合であり，上近似 $\overline{B}(D_i)$ は D_i に分類される可能性

表 2: 表 1 の識別行列

	x1	x2	x3	x4	x5	x6
x1	$*$					
x2	$*$	$*$				
x3	{c1,c2,c3,c5,c6}	{c2,c3,c5,c6}	$*$			
x4	{c1,c2,c3,c4,c5,c6}	{c3,c4,c5,c6}	$*$	$*$		
x5	$*$	$*$	{c2,c3,c5,c6}	{c3,c4,c5,c6}	$*$	
x6	{c2,c5}	{c1,c5}	{c1,c2,c3,c6}	{c1,c3,c4,c6}	{c1,c3,c5}	$*$

がある対象の集合である．

決定属性の下近似と同じ分割を与える最小限の属性の集合を相対縮約と呼ぶ．相対縮約は各決定クラス $D_1, \ldots, D_m$ に対して，以下の 2 条件を満たす条件属性集合 $A \subseteq C$ である．

1. $\underline{A}(D_i) = \underline{C}(D_i)$.
2. 任意の $B \subset A$ に対して $\underline{B}(D_i) \neq \underline{C}(D_i)$.

例として，表 1 の相対縮約は {c1, c5}, {c3, c5}, {c5, c6}, {c1, c2, c3}, {c1, c2, c4},{c1, c2, c6},{c2, c4, c5} の 7 個である．

2.2 識別行列

すべての相対縮約を計算する方法として，識別行列を使った計算法 [5] が知られている．

決定表 $(U, C \cup D, V, \rho)$ が与えられたとき，決定属性集合 $B \subseteq D$ に関する識別行列は，(i, j) 成分 δ_{ij} が次の集合あるいは $*$ で定義される $|U| \times |U|$ 行列である．

$$\delta_{ij} = \begin{cases} \{a \in C | \rho(x_i, a) \neq \rho(x_j, a)\}; \\ \quad \text{決定属性の値が異なる場合}, \\ *; \quad \text{その他}. \end{cases} \tag{5}$$

ここで，$|U|$ は集合 U の要素数であり，$*$ は don't care を意味する．δ_{ij} は，決定属性の値が等しくない場合に属性値が異なる条件属性の集合を表している．i 番目の対象と j 番目の対象の決定属性の値が等しくない理由が δ_{ij} 内のいずれかの属性に対する値の相違にあることを示し，δ_{ij} 内のいずれかの属性を用いれば，i 番目の対象と j 番目の対象を区別できることを示している．$\delta_{ij} = \delta_{ji}, \delta_{ii} = *$ であることに注意すると，上三角あるいは下三角の部分が与えられれば十分である．

したがって，条件属性の集合 $Q \subseteq C$ が，任意の $(i, j)(1 \leq i, j \leq n)$ に対して，条件

$$\delta_{ij} \neq * \quad \text{ならば} \quad \delta_{ij} \cap Q \neq \phi \tag{6}$$

を満足するとき，Q 内のすべての条件属性を用いることにより，条件属性集合 C による近似の質を低下させることなく，決定属性集合 B の属性値を判定することができる．ゆえに，条件 (6) を満たす極小な属性集合 Q が相対縮約となる．

表 2 は，表 1 の識別行列である．この識別行列より，以下の論理式を構成し，主加法形に変形すると，次のようになる．

$$\begin{gathered}(\mathrm{c1}\vee\mathrm{c2}\vee\mathrm{c3}\vee\mathrm{c5}\vee\mathrm{c6})\wedge\cdots\wedge(\mathrm{c1}\vee\mathrm{c3}\vee\mathrm{c5}) \\ \equiv(\mathrm{c1}\wedge\mathrm{c5})\vee(\mathrm{c3}\wedge\mathrm{c5})\vee(\mathrm{c5}\wedge\mathrm{c6})\vee(\mathrm{c1}\wedge\mathrm{c2}\wedge\mathrm{c3})\vee \\ (\mathrm{c1}\wedge\mathrm{c2}\wedge\mathrm{c4})\vee(\mathrm{c1}\wedge\mathrm{c2}\wedge\mathrm{c6})\vee(\mathrm{c2}\wedge\mathrm{c4}\wedge\mathrm{c5})\end{gathered}$$

これにより，相対縮約として {c1,c5},{c3,c5},{c5,c6}, {c1,c2,c3},{c1,c2,c4},{c1,c2,c6},{c2,c4,c5} の 7 個が得られる．

2.3 同値類による分割を用いた相対縮約の評価手法

同値類による分割を用いた相対縮約の評価手法 [2] について概要を説明する．

条件属性集合 $B \subseteq C$，決定属性集合 D，対象 $x \in U$ から得られる決定ルールを

$$(B, x) \rightarrow (D, x)$$

と表し，その評価指標として，決定ルールの一般性を表す指標であり，次式で定義される被覆度を用いる．

$$\mathrm{Cov}((B, x) \rightarrow (D, x)) = \frac{|[x]_B \cap D_i|}{|D_i|}. \tag{7}$$

ここで，D_i は $x \in D_i$ となる決定クラスである．

評価の方針として,「粗い」分割を与える相対縮約ほど高く評価する．その際，被覆度の平均値を用いる．任意の空でない条件属性集合 $B \subseteq C$ および決定属性集合 D に対して，属性集合 B から得られるすべての決定ルール $(B, x) \rightarrow (D, x)$ の被覆度の平均値 $\mathrm{ACov}(B)$ は，

$$\mathrm{ACov}(B) = \frac{|U/R_D|}{\mathrm{N_{rule}}} \tag{8}$$

表 3: 表 1 の相対縮約から得られる Eval(B) と ACov(B)

相対縮約	Eval(B)	ACov(B)
{c1,c5}	10.5	0.75
{c3,c5}	12	0.75
{c5,c6}	12.5	1
{c1,c2,c3}	8.67	0.5
{c1,c2,c4}	7.33	0.6
{c1,c2,c6}	9	0.6
{c2,c4,c5}	9	0.6

で求められる．ここで，$\mathrm{N_{rule}}$ は条件属性集合 B および決定属性集合 D から得られる決定ルールの総数である．条件属性集合として相対縮約を用いると，被覆度の平均値は，相対縮約を用いて構成された分割に含まれる同値類の個数が少ない (分割が「粗い」) ほど値が高くなる．

例として，表 1 の相対縮約 B={c1,c5} について，ACov(B) を求める．相対縮約 B および $x_i \in U$ $(1 \leq i \leq 6)$ から以下の 4 個の決定ルールが得られる．

- $(\mathrm{c1}=1) \wedge (\mathrm{c5}=1) \rightarrow (\mathrm{d}=1)$, 被覆度 $\frac{1}{3}$
- $(\mathrm{c1}=2) \wedge (\mathrm{c5}=1) \rightarrow (\mathrm{d}=1)$, 被覆度 $\frac{2}{3}$
- $(\mathrm{c1}=2) \wedge (\mathrm{c5}=2) \rightarrow (\mathrm{d}=2)$, 被覆度 1
- $(\mathrm{c1}=1) \wedge (\mathrm{c5}=2) \rightarrow (\mathrm{d}=3)$, 被覆度 1

被覆度の平均値を実際に計算すると $\frac{3}{4}$ であり，式 (8) で求めた値と一致する．すべての相対縮約について，ACov(B) を求めた結果を表 3 に示す．この結果より，相対縮約 B={c5,c6} が最も「粗い」分割を与える相対縮約とみなされる．実際に，{c5,c6} から得られる分割は，決定属性集合 D が生成する同値関係 R_D による分割 U/R_D と一致する．

3 条件属性による類別に基づく相対縮約の評価

2.3 節の手法は、決定ルールの被覆度の平均値で相対縮約の評価を行うが、その評価の際に相対縮約から対象の同値類を作る必要がある．また，この手法では個々の条件属性の評価を行っていない．同値類を作らず，個々の条件属性を行うという観点から，本節では，条件属性による類別に基づく相対縮約の評価手法を提案する．

評価の方針として，異なる決定クラスに属する対象を正確に区別し，かつ同じ決定クラスに属する対象をできるだけ区別しない相対縮約を高く評価する．そこで，相対縮約を構成する各条件属性の類別能力に基づいて，相対縮約を評価する．まず，各条件属性 $a \in C$ に対して，以下の 2 つの集合を定義する．

$$\mathrm{Dis}(a) = \left\{ (x_i, x_j) \,\middle|\, \begin{array}{l} \rho(x_i,a) \neq \rho(x_j,a), \\ \rho(x_i,d) \neq \rho(x_j,d), \\ \exists d \in D, i > j \end{array} \right\}, \quad (9)$$

$$\mathrm{Indis}(a) = \left\{ (x_i, x_j) \,\middle|\, \begin{array}{l} \rho(x_i,c) = \rho(x_j,c), \\ \rho(x_i,d) = \rho(x_j,d), \\ \forall d \in D, i > j \end{array} \right\}. \quad (10)$$

Dis(a) は異なる決定クラスに属し，かつ条件属性 a での値が異なる対象 x_i と x_j の対の集合，Indis(a) は同じ決定クラスに属し，かつ条件属性 a での値が等しい対象 x_i と x_j の対の集合である．

条件属性 a の類別能力の評価する指標として，

$$\mathrm{Eval}(a) = |\mathrm{Dis}(a)| + |\mathrm{Indis}(a)| \quad (11)$$

を用いる．式 (9) より，$|\mathrm{Dis}(a)|$ は識別行列の下三角部分における a の出現回数に等しい．また，$|\mathrm{Indis}(a)|$ は識別行列を作成する際に計算可能である．この指標は，決定クラスが異なる対象を区別しない，または同じ決定クラスの対象を区別すると低くなり，決定クラスが異なる対象を区別し，かつ同じ決定クラスの対象を区別しないと高くなる．

相対縮約 B の評価指標を次式で定義する．

$$\mathrm{Eval}(B) = \frac{1}{|B|} \sum_{a_i \in B} \mathrm{Eval}(a_i). \quad (12)$$

この値が大きいほど類別能力が高い相対縮約とみなす．

例として，表 1 の相対縮約 B={c1,c5} について，Eval(B) を求める．

Dis(c1)={(x1,x3),(x1,x4),(x2,x6),(x3,x6),(x4,x6),(x5,x6)},
Indis(c1)={(x2,x5),(x3,x4)},
Eval(c1)=6+2=8.
Dis(c5)={(x1,x3),(x1,x4),(x1,x6),(x2,x3),(x2,x4),(x2,x6),(x3,x5),(x4,x5),(x5,x6)},
Indis(c5)={(x1,x2),(x1,x5),(x2,x5),(x3,x4)},
Eval(c5)=9+4=13.
よって，

$$\mathrm{Eval}(B) = \frac{8+13}{2} = \frac{21}{2} = 10.5$$

となる．

すべての相対縮約について Eval(B) を求めた結果を表 3 に示す．この結果より，{c5,c6} が最も「類別能力が高い」相対縮約とみなされる．前節で述べた通り，{c5,c6} からは決定属性集合 D が生成する同値関係 R_D による

表 4: 評価結果の比較

相対縮約	Eval(B)	ACov(B)	順位 (Eval(B))	順位 (ACov(B))
{eggs,milk,aquatic,toothed,legs}	3542.8	0.318	1	3
{eggs,aquatic,toothed,legs,catsize}	3384	0.259	2	17
{eggs,aquatic,toothed,breath,legs}	3376.8	0.304	3	6
{milk,aquatic,backbone,fins,legs}	3123.2	0.35	18	1
{eggs,milk,aquatic,backbone,legs}	3371.2	0.333	4	2
{eggs,milk,aquatic,toothed,legs}	3542.8	0.318	1	3
{milk,aquatic,toothed,fins,legs}	3294.8	0.318	6	3
{egg,aquatic,toothed,backbone,fins,legs}	3176.167	0.318	14	3

分割 U/R_D が得られることから，{c5,c6} は実際に，決定クラスに対して高い類別能力を有することがわかる．

なお，Yamaguchi [6] はこの評価手法と同様の方針に基づいて，Pawlak の attribute dependency の改良を試みているが，提案手法とは定義および使用方法が異なる．

4 実験

前節で提案した手法を，サンプル数 101，属性数 17 である UCI Machine Learning Repository の Zoo データ [1] に適用した．17 個の属性のうちの 1 つである”type” を決定属性，それ以外を条件属性とした．相対縮約は全部で 33 個存在した．提案手法を用いて得られた相対縮約を評価し，その結果を同値類による分割に基づく相対縮約の評価手法 [2] における評価結果と比較した．

まず，33 個の相対縮約に対する，提案手法による評価値と同値類による分割に基づく相対縮約の評価手法による評価値との関連性を調べるため，スピアマンの順位相関係数を求めた．その結果，順位相関係数は 0.627 であり，両者の間には中程度の正の相関が見られた．このことから，提案手法による評価と分割に基づく評価手法は，ある程度似た傾向を示すことがわかる．

表 4 は，それぞれの評価手法による評価値での上位 3 番目までの相対縮約を抜き出したものである．この表から Eval(B) と ACov(B) とでは，評価値が最も高い相対縮約が異なっていることがわかる．また，評価値の全体的な傾向は，両者とも相対縮約を構成する条件属性の数が少ないものほど高くなっている．また，ACov(B) での評価が 3 番目となる相対縮約のように，ACov(B) の値が同じでも Eval(B) の値が異なる場合が複数見られた．このことから，提案手法ではより細かく評価できる可能性があると考えられる．

5 まとめ

本研究では，条件属性の類別能力の評価を用いて，相対縮約を評価する手法を提案した．そして，提案手法により得られた評価結果を同値類による分割に基づく相対縮約の評価手法 [2] における評価結果と比較した．

今後の課題として，提案手法のより詳細な検討，提案手法により得られた評価結果の妥当性の検証，さらに複数のデータでの実験が挙げられる．

参考文献

[1] Forsyth, R.: *Zoo Database*, UCI Machine Learning Repository, 1990.

[2] 工藤 康生, 村井 哲也: 同値類による分割を用いた相対縮約の評価について, 感性フォーラム札幌 2008 講演論文集, 2008.

[3] 森 典彦, 田中 英夫, 井上 勝雄 (共編): 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.

[4] Pawlak, Z.: *Rough Sets: Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers, 1991.

[5] Skowron, A. and Rauszer, C.M.: *The discernibility matrix and functions in information systems, R.Slowinski(ed.), Intelligent Decision Support: Handbook of Application and Advances of Rough Set Theory*, Kluwer Academic Publishers, pp.331-361, 1992.

[6] Yamaguchi, D.: On the Improvement of Pawlak’s Attribute Dependency Model, *Proc. of the 2nd International Conference on Kansei Engineering and Affective Systems*, pp.83–88, JSKE, 2008.

連絡先

工藤 康生
〒 050-8585 北海道室蘭市水元町 27-1
室蘭工業大学工学部情報工学科
TEL 0143-46-5469 FAX 0143-46-5499
E-mail: kudo@csse.muroran-it.ac.jp