条件属性による類別を用いた相対縮約の近似計算手法について

条件属性による類別を用いた相対縮約の近似計算手法について

その他（別言語等）のタイトル	On an Approximate Calculation Method of a Relative Reduct Based on Classification by Condition Attributes
著者	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/10258/2200

doi: info:doi/10.14864/fss.25.0.106.0

条件属性による類別を用いた相対縮約の近似計算手法について

On an Approximate Calculation Method of a Relative Reduct Based on Classification by Condition Attributes

工藤 康生　　村井 哲也
Yasuo Kudo　　Tetsuya Murai
室蘭工業大学　　北海道大学
Mutoran Institute of Technology　Hokkaido University

Abstract: In this paper, we introduce an approximate calculation method of a relative reduct based on an evaluation method of classification ability of condition attributes with respect to decision classes proposed by the authors. Because it has been proved that computational complexity of calculating all relative reducts of a given decision table is NP-hard, many algorithms have been proposed to calculate relative reducts approximately. By our proposal, we intend to calculate a relative reduct with as better evaluation by the evaluation method of relative reducts as possible. We applied the proposed method to Zoo data set in UCI machine learning repository, and obtained the best relative reduct among 33 relative reducts of Zoo data set.

1　はじめに

近年，ラフ集合理論 [7, 8] はカテゴリカルなデータに対するデータマイニング手法として注目されており，特に，データを正しく分類するために最小限必要となる項目の集合 (相対縮約) およびデータに含まれる if-then 形式のルール (決定ルール) の抽出について，理論と応用の両面から幅広く研究が進められている (詳細は例えば [5])．分析対象のデータからすべての相対縮約を抽出する手法として，識別行列 [10] を用いる手法が知られている．しかし，すべての相対縮約を求める計算は NP 困難であることが証明されているため [10]，大規模データに対してすべての相対縮約を計算することは現実的ではない．そのため，相対縮約の近似計算手法が多数研究されている [1, 2, 4, 9, 11, 12, 13].

本研究では，著者らが提案した条件属性による類別に基づく相対縮約の評価手法 [6] に基づいて，この評価手法においてできるだけ良い評価を得る相対縮約を 1 個だけ求める手法を提案する．

2　ラフ集合

本節ではラフ集合の概要について簡略に説明する．なお，本節の内容は文献 [3, 5] に基づく．

2.1　決定表と識別不能関係

ラフ集合で扱うデータは一般的に，以下で定義される決定表で表される:

$$(U, C \cup D, V, \rho).$$

ここで，U は対象の空でない有限集合，C は条件属性の空でない有限集合，D は決定属性の空でない有限集合であり，$C \cap D = \emptyset$ とする．すべての属性の集合を $AT \stackrel{\text{def}}{=} C \cup D$ と表す．V は各属性 $a \in AT$ の値の集合，$\rho : U \times AT \to V$ は対象 x の属性 a での値 $\rho(x, a) \in V$ を表す関数である．

属性の任意の部分集合 $A \subseteq AT$ に対して，U 上の識別不能関係 R_A を次式で定義する:

$$xR_Ay \stackrel{\text{def}}{\Longleftrightarrow} \rho(x, a) = \rho(y, a),\ \forall a \in A. \qquad (1)$$

関係 R_A が同値関係となることは容易に確かめられる．特に，決定属性集合 D に基づく識別不能関係は対象の全体集合の分割 $U/D = \{D_1, \cdots, D_m\}$ を与え，各 D_i は決定クラスと呼ばれる．

各決定クラス D_i に対して，識別不能関係 R_A による下近似 $\underline{A}(D_i)$ を次式で定義する:

$$\underline{A}(D_i) \stackrel{\text{def}}{=} \{x \in U \mid [x]_A \subseteq D_i\}. \qquad (2)$$

識別不能関係 R_A の定義より，D_i の下近似 $\underline{A}(D_i)$ は，A に含まれる属性の値によって，確実に D_i に分類される対象の集合となる．

決定表の例を表 1 に示す．表 1 は議論の対象となる要素の集合 $U = \{x_1, \cdots, x_6\}$, 条件属性集合 $C = \{c_1, \cdots, c_6\}$ 決定属性集合 $D = \{d\}$ などで構成され，$\rho(x_i, d) = i$ となる要素の集合を決定クラス D_i とすると，3 個の決定クラス $D_1 = \{x_1, x_2, x_5\}$ および $D_2 = \{x_3, x_4\}$，$D_3 = \{x_6\}$ が得られる．

表 1: 決定表の例

U	c_1	c_2	c_3	c_4	c_5	c_6	d
x_1	1	0	0	0	0	1	1
x_2	0	1	0	0	0	1	1
x_3	0	2	1	0	1	0	2
x_4	0	1	1	1	0	0	2
x_5	0	1	2	0	0	1	1
x_6	0	1	0	0	1	1	3

2.2 相対縮約

データから規則性を見出す観点から，できるだけ少ない属性数で，条件属性 C をすべて用いた識別不能関係 R_C による分類と同等な分類を与え，すべての決定クラスを近似できることが望ましい．そのような性質を満たす条件属性の集合 $A \subseteq C$ を相対縮約と呼ぶ．形式的には，分割 U/D の C に関する相対縮約とは，すべての決定クラス $D_i \in U/D$ $(i = 1, \cdots, m)$ に対して以下の 2 条件を満たす条件属性の部分集合 $A \subseteq C$ である:

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

ここで，$B \subset A$ は集合 B が集合 A の真部分集合であることを意味する．相対縮約は複数個存在することがあり，すべての相対縮約に現れる条件属性の集合をコアと呼ぶ．例として，表 1 には以下の 3 種類の相対縮約が存在する: $\{c_3, c_5\}$, $\{c_5, c_6\}$, $\{c_2, c_4, c_5\}$．これらの相対縮約のコアは $\{c_5\}$ である．

2.3 識別行列による相対縮約の計算

相対縮約を具体的に計算する手法として，識別行列 [10] を用いた手法が知られている．決定表 $(U, C \cup D, V, \rho)$ が与えられたとき，決定属性集合 D に関する識別行列は，以下で定義する i 行 j 列目の成分 δ_{ij} を持つ $|U| \times |U|$ 行列である:

$$\delta_{ij} = \begin{cases} \{a \in C \mid \rho(x_i, a) \neq \rho(x_j, a)\}, \\ \qquad \exists d \in D,\ \rho(x_i, d) \neq \rho(x_j, d), \\ \emptyset, \quad \text{その他}. \end{cases} \tag{3}$$

ここで，$|U|$ は集合 U の要素数を表す．定義より明らかに，任意の $i, j \in \{1, \cdots, |U|\}$ に対して $\delta_{ij} = \delta_{ji}$ かつ $\delta_{ii} = \emptyset$ となるので，実際に識別行列を計算する場合は，行列の上三角部分または下三角部分のみで十分である．

$\delta_{ij} \neq \emptyset$ である i 行 j 列の成分 δ_{ij} は，決定クラスが異なる対象 x_i と x_j に対して，δ_{ij} に含まれるいずれかの属性を比較することで x_i と x_j を区別できることを表している．よって，すべての δ_{ij} に対して，$\delta_{ij} \neq \emptyset$ ならば $\delta_{ij} \cap A \neq \emptyset$ となり，かつ包含関係について極小となるような条件属性の部分集合 $A \subseteq C$ が相対縮約となる．識別行列を用いることで，与えられた決定表におけるすべての相対縮約を求めることが可能である．しかし，すべての相対縮約を求める計算は NP 困難であることが証明されている．

3 条件属性による類別を用いた相対縮約の評価

本節では，著者ら [6] が提案した条件属性による類別を用いた相対縮約の評価手法について概要を説明する．

本手法では，相対縮約に出現する各条件属性の決定クラスに対する類別能力を評価することにより，相対縮約の評価を行う．具体的には，以下の 2 条件を満たす条件属性を類別能力が高いと見なす:

- 異なる決定クラスの対象をできるだけ区別する．
- 同じ決定クラスの対象をできるだけ区別しない．

この評価方針に基づいて各条件属性 $c \in C$ を評価するために，以下の 2 種類の集合を導入する:

$$Dis(c) = \left\{ (x_i, x_j) \;\middle|\; \begin{array}{l} \rho(x_i, a) \neq \rho(x_j, a), \\ \rho(x_i, d) \neq \rho(x_j, d), \\ \exists d \in D, i > j \end{array} \right\}, \tag{4}$$

$$Indis(c) = \left\{ (x_i, x_j) \;\middle|\; \begin{array}{l} \rho(x_i, c) = \rho(x_j, c), \\ \rho(x_i, d) = \rho(x_j, d), \\ \forall d \in D, i > j \end{array} \right\}. \tag{5}$$

集合 $Dis(c)$ は，条件属性 c によって識別可能であり，かつ属する決定クラスが異なる対象 x_i と x_j $(i > j)$ の対の集合である．一方，集合 $Indis(c)$ は c によって識別不能であり，かつ属する決定クラスが等しい対象の対の集合である．

これらの集合を用いて，条件属性 c に対する評価指標 $Eval(c)$ を次式で定義し，この値が大きい属性ほど決定クラスに対する類別能力が高いと見なす:

$$Eval(c) \stackrel{\text{def}}{=} |Dis(c)| + |Indis(c)|. \tag{6}$$

集合 $Dis(c)$ の定義より，その要素数 $|Dis(c)|$ は識別行列の下三角部分における属性 c の出現回数に等しい．また，$Indis(c)$ も識別行列の構成と並行して構成することができるため，識別行列を構成する際に，すべての条件属性 $c \in C$ の評価値 $Eval(c)$ を求めることが可能である．

各条件属性に対する評価値を用いて，相対縮約 A の評価値 $Eval(A)$ を，その相対縮約に含まれる属性の評

表 2: 表 1 の決定表に対する識別行列

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	$\emptyset$					
x_2	$\emptyset$	$\emptyset$				
x_3	$\{c_1, c_2, c_3, c_5, c_6\}$	$\{c_2, c_3, c_5, c_6\}$	$\emptyset$			
x_4	$\{c_1, c_2, c_3, c_4, c_6\}$	$\{c_3, c_4, c_6\}$	$\emptyset$	$\emptyset$		
x_5	$\emptyset$	$\emptyset$	$\{c_2, c_3, c_5, c_6\}$	$\{c_3, c_4, c_6\}$	$\emptyset$	
x_6	$\{c_1, c_2, c_5\}$	$\{c_5\}$	$\{c_2, c_3, c_6\}$	$\{c_3, c_4, c_5, c_6\}$	$\{c_3, c_5\}$	$\emptyset$

価値の平均として次式で定義する:

$$Eval(A) \stackrel{\text{def}}{=} \frac{1}{|A|} \sum_{c_i \in A} Eval(c_i). \tag{7}$$

相対縮約の評価値が高いほど，類別能力が高い相対縮約であると見なす.

なお，Yamaguchi [14] 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，この手法とは定義および使用方法が異なる.

4 条件属性による類別を用いた相対縮約の近似計算手法

本節では，前節で述べた相対縮約の評価手法によって，できるだけ高い評価を得る相対縮約を 1 個求めるための近似計算手法を提案する.

提案手法は，現時点で最も評価値 $Eval(c)$ が高い条件属性 $c \in C$ を現在の相対縮約の候補に追加し，評価値を更新することを繰り返して，できるだけ少ない条件属性で相対縮約の候補を作成する．評価値を更新する際に，$|Dis(c)|$ の更新には識別行列を用いる．一方，$|Indis(c)|$ を更新するために，以下で定義する同値行列を導入する.

定義 1 決定表 $(U, C \cup D, V, \rho)$ が与えられたとき，決定属性集合 D に関する同値行列とは，以下で定義する i 行 j 列目の成分 ϵ_{ij} を持つ $|U| \times |U|$ 行列である:

$$\epsilon_{ij} = \begin{cases} \{a \in C \mid \rho(x_i, a) = \rho(x_j, a)\}, \\ \qquad \forall d \in D,\ \rho(x_i, d) = \rho(x_j, d), \\ \emptyset, \quad \text{その他.} \end{cases} \tag{8}$$

この定義より明らかに，任意の $i, j \in \{1, \cdots, |U|\}$ に対して $\epsilon_{ij} = \epsilon_{ji}$ かつ $\epsilon_{ii} = C$ となるので，実際に同値行列を計算する場合は，行列の上三角部分または下三角部分のみで十分である．識別行列とは逆に，同値行列の i 行 j 列目の要素は，同じ決定クラスに属する要素 x_i と x_j において，値が等しい条件属性の集合である．例として，表 1 の同値行列を表 3 に示す.

識別行列 DM および同値行列 EM を用いると，式 (4) および式 (5) は以下のように表すことができる:

$$Dis(c) = \{(x_i, x_j) \mid \delta_{ij} \in DM, i > j\}, \tag{9}$$

$$Indis(c) = \{(x_i, x_j) \mid \epsilon_{ij} \in EM, i > j\}. \tag{10}$$

よって，識別行列および同値行列を用いることで，条件属性の評価値 $Eval(c)$ を求めることができる.

提案手法のアルゴリズムを以下に示す．なお，このアルゴリズムにおいて，表現 $X := Y$ は集合 X を集合 Y で置き換えることを意味する.

条件属性の評価値を用いた相対縮約の近似計算法
入力: 条件属性集合 C，識別行列 DM，同値行列 EM.
出力: 相対縮約の候補となる条件属性集合 S.

1. $S := \bigcup\{\delta \in DM \mid |\delta| = 1\}$.
2. すべての $\delta_{ij} \in DM$ について，$\delta_{ij} \neq \emptyset$ ならば $S \cap \delta_{ij} \neq \emptyset$ であれば，S を出力して終了する.
3. それぞれの $\delta_{ij} \in DM$ について，$S \cap \delta_{ij} \neq \emptyset$ であれば，$\delta_{ij} := \emptyset$ とする．同様に，それぞれの $\epsilon_{ij} \in EM$ について，$S \cap \epsilon_{ij} \neq \emptyset$ であれば，$\epsilon_{ij} := \emptyset$ とする.
4. DM および EM を用いて，それぞれの条件属性 $c \in C$ の評価値 $Eval(c) = |Dis(c)| + |Indis(c)|$ を求める.
5. 最も評価値が高い属性 c_h を S に追加する; $S := S \cup \{c_h\}$.
6. C から c_h を除去する; $C := C - \{c_h\}$.
7. 2. に戻る.

このアルゴリズムにおいて，1. では，条件属性 $c \in C$ が相対縮約のコアに含まれることの必要十分条件は，$\delta_{ij} = \{c\}$ となる識別行列の要素 δ_{ij} が存在することである [10] ことを用いて，相対縮約のコアを求めている．また，3. および 4. では，相対縮約の候補 S と非空共通部分を持つ要素 δ_{ij} および ϵ_{ij} を空集合に置き換えることで，評価値 $Eval(c)$ を更新する.

表 3: 表 1 の決定表に対する同値行列

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	C					
x_2	$\{c_3, c_4, c_5, c_6\}$	C				
x_3	$\emptyset$	$\emptyset$	C			
x_4	$\emptyset$	$\emptyset$	$\{c_1, c_3, c_6\}$	C		
x_5	$\{c_4, c_5, c_6\}$	$\{c_1, c_2, c_4, c_5, c_6\}$	$\emptyset$	$\emptyset$	C	
x_6	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	C

5 実験

提案した近似計算手法を，UCI Machine Learning Repository の Zoo データ [15] に対して用いた．Zoo データは 101 個の対象と 17 個の属性で構成されている．属性 “type”を決定属性，それ以外を条件属性とする決定表を作成し，提案手法を用いて相対縮約の候補を 1 個求めた．また，比較のために決定表におけるすべての相対縮約を求め，得られた 33 個の相対縮約に対して，条件属性による類別を用いた相対縮約の評価手法 [6] を用いて評価を行った．

実験の結果，出力された条件属性集合は {eggs, milk, aquatic, toothed, legs} であった．この集合は 33 個存在する相対縮約の中の 1 つであり，かつ文献 [6] で述べられている通り，上述の評価手法では 33 個中最も評価が高い相対縮約である．

6 おわりに

本研究では，条件属性による類別に基づく相対縮約の評価手法において，できるだけ良い評価を得る相対縮約を 1 個だけ求める近似計算手法を提案した．また，実験として提案手法を Zoo データに対して用い，決定表から得られた相対縮約の中で，上述の評価手法で最も評価が高い相対縮約が得られることを示した．

今後の課題として，提案手法の理論的性質，特に計算量に関する考察および他のデータでの検証，他の近似計算手法との比較などが挙げられる．

文献

[1] Hedar, A. H., Wang, J. and Fukushima, M.: Tabu search for attribute reduction in rough set theory, *Soft Couping*, Vol. 12, No. 9, pp. 909–918, Springer, 2008.

[2] Hu, F., Wang, G. and Feng, L.: Fast Knowledge Reduction Algorithms Based on Quick Sort, *Rough Sets and Knowledge Technology*, LNAI 5009, Springer, pp.72–79, 2008.

[3] 工藤 康生, 村井 哲也: ラフ集合によるルール生成, 第 16 回あいまいと感性研究部会ワークショップ講演論文集, pp.18–23, 2006.

[4] Kudo, Y. and Murai, T.: A Heuristic Algorithm for Selective Calculation of a Better Relative Reduct in Rough Set Theory, *New Advances in Intelligent Decision Technologies*, Nakamatsu, K. et al. (eds.), SCI199, pp.555-564, Springer, 2009.

[5] 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.

[6] 中浦 孝仁, 工藤 康生, 村井 哲也: 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, 第 25 回ファジィシステムシンポジウム講演論文集, 2009.

[7] Pawlak, Z.: Rough Sets, *International Journal of Computer and Information Science*, Vol. 11, pp.341–356, 1982.

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

[9] Pawlak, Z. and Słowiński, R.: Rough Set Approach to Multi-Attribute Decision Analysis, *European Journal of Operation Research*, Vol. 74, pp.443–459, 1994.

[10] Skowron, A. and Rauszer, C. M.: The discernibility matrix and functions in information systems, *Intelligent Decision Support: Handbook of Application and Advance of the Rough Set Theory*, Słowiński, R. (ed.), Kluwer Academic Publishers, pp.331–362, 1992.

[11] Ślęzak, D.: Approximate Entropy Reducts, *Fundamenta Informaticae*, Vol. 53, No. 3–4, pp.365–387, 2002.

[12] Xu, J. and Sun, L.: New Reduction Algorithm Based on Decision Power of Decision Table, *Rough Sets and Knowledge Technology*, LNAI 5009, Springer, pp.180–188, 2008.

[13] Xu, Z., Zhang, C., Zhang, S., Song, W. and Yang, B.: Efficient Attribute Reduction Based on Discernibility Matrix, *Rough Sets and Knowledge Technology*, LNAI 4481, Springer, pp.13–21, 2007.

[14] 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.

[15] http://archive.ics.uci.edu/ml/datasets/Zoo

連絡先
工藤 康生
〒 050-8585 北海道室蘭市水元町 27-1
室蘭工業大学しくみ情報系領域
Tel: 0143–46–5469, Fax: 0143–46–5499
E-mail: kudo@csse.muroran-it.ac.jp