数理解析研究所講究録
第1802巻 2012年 27-33



# 支配関係に基づく可変精度ラフ集合アプローチによる事例ベースクラス推定法

大阪大学大学院基礎工学研究科
乾口 雅弘 (Masahiro Inuiguchi)
鶴見 昌代 (Masayo Tsurumi)
Graduate School of Engineering Science, Osaka Univ.

## 1. はじめに

ラフ集合は，クラス分類データをまとめた決定表の解析に有効で，分類に必要な極小な属性集合の算出や極小な条件部をもつ決定ルール抽出が可能である．決定ルールの抽出は未知対象のクラス推定に有用で，これに基づく種々の分類システムが提案されている．しかし，ルール抽出には計算時間が必要なため，データ数が多くなると計算時間が莫大になる．また，クラス推定したい未知対象が少ない場合には，ルール抽出に多くの計算コストをかけることは適切ではない．さらに，クラス分類データが逐次増加する動的な環境では，ルール抽出に基づくクラス推定が有効であるとは言い切れない．このため，ルール抽出を行わずに決定表から直接評価値を推定する事例ベースクラス推定法として，$k$-近傍法，近傍分類法，ラフ集合の概念を導入した方法 [1, 2] などが提案されている．決定属性値が条件属性値に関して単調な場合については，Dembczyński et al. [4] がルール抽出を行わない評価値推定法を提案しているが，文献 [1, 2] と同様に，未知対象に関連するルールをルール抽出法と同等な基準で逐次抽出し，これらのルールに基づいて未知対象のクラスを推定するものである．

本研究は，決定属性値が条件属性値に関して単調な場合において，支配関係に基づく可変精度ラフ集合モデル (VP-DRSA) の精度を用いた事例ベースクラス推定法を提案する．数値実験により，計算時間と推定精度の両面から，DOMLEM [3] に基づいてルール抽出した場合と提案手法を比較し，いずれが有効な手法であるかを明らかにする．

## 2. 可変精度ラフ集合

対象の集合 $U$, 条件属性の集合 $C$，決定属性 $d$ からなるデータ表 $T=(U, C\cup\{d\})$ は，決定表と呼ばれる．$a\in C\cup\{d\}$ の属性値の集合を $V_a$ と表し，いずれの $a\in C\cup\{d\}$ も $V_a$ 上の全順序 $\succeq_a$ をもつ序数属性とする．決定属性 $d$ に対して，一般性を失うことなく，$V_d=\{1,2,\ldots,q\}$, $\succeq_d=\geq$ とする．$u\in U$ の $a\in C\cup\{d\}$ に関する属性値を $a(u)$ と表す．決定属性値 $t\in V_d$ に対して，上側和集合と下側和集合をそれぞれ $Cl_t^{\geq}=\{u\in U: d(u)\geq t\}$, $Cl_t^{\leq}=\{u\in U: t\geq d(u)\}$ で定義する．便宜上，$Cl_{q+1}^{\geq}=Cl_0^{\leq}=\emptyset$ とする．また，条件属性集合 $P\subseteq C$ おける支配関係を $D_P=\{(u_1,u_2)\in U\times U: a(u_1)\succeq_a a(u_2), \forall a\in P\}$ と定義する．対象 $u$ が与えられたとき，$u$ を支配する対象の集合と $u$ に支配される対象の集合は，それぞれ $D_P^+(u)=\{u'\in U:(u',u)\in D_P\}$, $D_P^-(u)=\{u'\in U:(u,u')\in D_P\}$ と定義される．以下では，$\preceq_a$, $\succeq_a$ $(a\in C)$ をそれぞれ単に $\preceq$, $\succeq$ と表す．$P\subseteq C$ のもとで，$u$ が $Cl_t^{\geq}$ に帰属する度合，および $Cl_t^{\leq}$ に帰属する度合は，次で定義される [5].

$$\mu_P\left(u, Cl_t^{\geq}\right)=\frac{|D_P^-(u)\cap Cl_t^{\geq}|}{|D_P^-(u)\cap Cl_t^{\geq}|+|D_P^+(u)\cap Cl_{t-1}^{\leq}|} \tag{1}$$

$$\mu_P\left(u, Cl_t^{\leq}\right)=\frac{|D_P^+(u)\cap Cl_t^{\leq}|}{|D_P^+(u)\cap Cl_t^{\leq}|+|D_P^-(u)\cap Cl_{t+1}^{\geq}|} \tag{2}$$

ここで $|X|$ は集合 $X$ の基数とする．$P \subseteq C$ のもとでの $Cl_t^{\geq}$ と $Cl_t^{\leq}$ の下近似と上近似[3]は，$\underline{P}\left(Cl_t^{\geq}\right) = \{u \in U : \mu_P\left(u, Cl_t^{\geq}\right) = 1\}$, $\overline{P}\left(Cl_t^{\geq}\right) = \{u \in U : \mu_P\left(u, Cl_t^{\geq}\right) > 0\}$, $\underline{P}\left(Cl_t^{\leq}\right) = \{u \in U : \mu_P\left(u, Cl_t^{\leq}\right) = 1\}$, $\overline{P}\left(Cl_t^{\leq}\right) = \{u \in U : \mu_P\left(u, Cl_t^{\leq}\right) > 0\}$ で定義される．

## 3. 事例ベースクラス分類法

本論文では，決定属性値が条件属性値に関して単調で，整合性のある決定表において，条件属性値がわかっていて決定属性値がわからない未知対象の決定属性値の推定法，すなわちクラス推定法を複数提案する．

### 3.1. すべての条件属性からの推定

未知対象 $u \notin U$ に対しても，支配関係 $D_P$ を拡張して用いることができ，$D_P^+(u) = \{u' \in U : (u', u) \in D_P\}$, $D_P^-(u) = \{u' \in U : (u, u') \in D_P\}$ とする．未知対象 $u \notin U$ に対して，$D_C^-(u) \neq \emptyset$ または $D_C^+(u) \neq \emptyset$ のときを考える．$\underline{t}$ を，$D_C^-(u) \neq \emptyset$ のとき $\max\{d(u_i) \mid u_i \in D_C^-(u)\}$ で，そうでないとき 1 と定義する．また，$\overline{t}$ を，$D_C^+(u) \neq \emptyset$ のとき $\min\{d(u_i) \mid u_i \in D_C^+(u)\}$ で，そうでないとき $q$ と定義する．このとき，$u$ の決定属性値は，$\underline{t}$ 以上 $\overline{t}$ 以下であると推定される．

### 3.2. 一部の条件属性による推定

上述の推定法で $\underline{t} \neq \overline{t}$ となるとき，既知対象と未知対象を一対比較し，その情報を有効に利用して推定することを考える．

#### 3.2.1. 与えられた既知対象に応じて定まる条件属性集合

既知対象 $\bar{u}_i \in U$ と未知対象 $u \notin U$ に対して，$P_u^{\geq}(\bar{u}_i) = \{a \in C : a(\bar{u}_i) \preceq a(u)\}$, $P_u^{\leq}(\bar{u}_i) = \{a \in C : a(u) \preceq a(\bar{u}_i)\}$ とする．$P_u^{\geq}(\bar{u}_i)$ を用いると $u$ の決定属性値 $d(u)$ が $d(\bar{u}_i)$ 以上であるかどうかを議論でき，$P_u^{\leq}(\bar{u}_i)$ を用いると $d(u)$ が $d(\bar{u}_i)$ 以下であるかどうかを議論できる．

$P_u^{\geq}(\bar{u}_i) \cap P_u^{\leq}(\bar{u}_i) = \{a \in C : a(u) = a(\bar{u}_i)\} \neq \emptyset$ から，$P_u^{\geq}(\bar{u}_i)$, $P_u^{\leq}(\bar{u}_i)$ の代わりに $P_u^{>}(\bar{u}_i) = \{a \in C : a(\bar{u}_i) \prec a(u)\}$, $P_u^{<}(\bar{u}_i) = \{a \in C : a(u) \prec a(\bar{u}_i)\}$ を用いることも考えられる．本論文では，VP-DRSA の精度を用いて，多少矛盾がある場合の支持度も考えるので，$P_u^{>}(\bar{u}_i)$ や $P_u^{<}(\bar{u}_i)$ を用いることも検討する．

#### 3.2.2. 既存対象との一対比較から導かれる指標

$P_u^{\geq}(\bar{u}_i)$ と $P_u^{\leq}(\bar{u}_i)$ を用いると，$u$ の決定属性値が $d(\bar{u}_i)$ 以上である精度と $d(\bar{u}_i)$ 以下である精度は，それぞれ次で得られる．

$$\mu_{P_u^{\geq}(\bar{u}_i)}\left(u, Cl_{d(\bar{u}_i)}^{\geq}\right) = \frac{\left|D^-_{P_u^{\geq}(\bar{u}_i)}(u) \cap Cl^{\geq}_{d(\bar{u}_i)}\right|}{\left|D^-_{P_u^{\geq}(\bar{u}_i)}(u) \cap Cl^{\geq}_{d(\bar{u}_i)}\right| + \left|D^+_{P_u^{\geq}(\bar{u}_i)}(u) \cap Cl^{\leq}_{d(\bar{u}_i)-1}\right|} \tag{3}$$

$$\mu_{P_u^{\leq}(\bar{u}_i)}\left(u, Cl_{d(\bar{u}_i)}^{\leq}\right) = \frac{\left|D^+_{P_u^{\leq}(\bar{u}_i)}(u) \cap Cl^{\leq}_{d(\bar{u}_i)}\right|}{\left|D^+_{P_u^{\leq}(\bar{u}_i)}(u) \cap Cl^{\leq}_{t}\right| + \left|D^-_{P_u^{\leq}(\bar{u}_i)}(u) \cap Cl^{\geq}_{d(\bar{u}_i)+1}\right|} \tag{4}$$

それぞれ，‘$d(u) \geq d(\bar{u}_i)$’ を支持する度合い，‘$d(u) \leq d(\bar{u}_i)$’ を支持する度合いとみなせる．

$P_u^{\geq}(\bar{u}_i) = \{a_{g^+(1)}, a_{g^+(2)}, \ldots, a_{g^+(s^+)}\}$ とする．$\bar{u}_i$ からルール $R^+(\bar{u}_i)$ "if $a_{g^+(1)}(u') \succeq a_{g^+(1)}(\bar{u}_i)$ and $a_{g^+(2)}(u') \succeq a_{g^+(2)}(\bar{u}_i)$ and,..., and $a_{g^+(s^+)}(u') \succeq a_{g^+(s^+)}(\bar{u}_i)$ then $d(u') \geq d(\bar{u}_i)$" を導かれるとき，$R^+(\bar{u}_i)$ の正しさ（accuracy）は，次で求められる．

$$\mu_{P_u^{\geq}(\bar{u}_i)}\left(\bar{u}_i, Cl_{d(\bar{u}_i)}^{\geq}\right) = \frac{\left|D^-_{P_u^{\geq}(\bar{u}_i)}(\bar{u}_i) \cap Cl_{d(\bar{u}_i)}^{\geq}\right|}{\left|D^-_{P_u^{\geq}(\bar{u}_i)}(\bar{u}_i) \cap Cl_{d(\bar{u}_i)}^{\geq}\right| + \left|D^+_{P_u^{\geq}(\bar{u}_i)}(\bar{u}_i) \cap Cl_{d(\bar{u}_i)-1}^{\leq}\right|} \tag{5}$$

未知対象 $u$ が $R^+(\bar{u}_i)$ の条件部を満たすのでこのルールから "$d(u) \geq d(\bar{u}_i)$" が得られ，それを支持する度合いが式 (5) の右辺となる．同様に，ルール $R^-(\bar{u}_i)$ の正しさを求めることができ，"$d(u) \leq d(\bar{u}_i)$" を支持する度合いとなる．

$P_u^{\geq}(\bar{u}_i)$ と $P_u^{\leq}(\bar{u}_i)$ を用いて定義される度合いについて議論したが，$P_u^{\geq}(\bar{u}_i)$ と $P_u^{\leq}(\bar{u}_i)$ の代わりに，$P_u^{>}(\bar{u}_i)$ と $P_u^{<}(\bar{u}_i)$ を用いた 4 つの指標を定義できる．すなわち，一つの既知対象 $\bar{u}_i \in U$ について，上側和集合 $Cl_{d(\bar{u}_i)}^{\geq}$ に対して $\mu_{P_u^{\geq}(\bar{u}_i)}(u, Cl_{d(\bar{u}_i)}^{\geq})$, $\mu_{P_u^{\geq}(\bar{u}_i)}(\bar{u}_i, Cl_{d(\bar{u}_i)}^{\geq})$, $\mu_{P_u^{>}(\bar{u}_i)}(u, Cl_{d(\bar{u}_i)}^{\geq})$, $\mu_{P_u^{>}(\bar{u}_i)}(\bar{u}_i, Cl_{d(\bar{u}_i)}^{\geq})$ の 4 指標が得られ，下側和集合 $Cl_{d(\bar{u}_i)}^{\leq}$ に対して $\mu_{P_u^{\leq}(\bar{u}_i)}(u, Cl_{d(\bar{u}_i)}^{\leq})$, $\mu_{P_u^{\leq}(\bar{u}_i)}(\bar{u}_i, Cl_{d(\bar{u}_i)}^{\leq})$, $\mu_{P_u^{<}(\bar{u}_i)}(u, Cl_{d(\bar{u}_i)}^{\leq})$, $\mu_{P_u^{<}(\bar{u}_i)}(\bar{u}_i, Cl_{d(\bar{u}_i)}^{\leq})$ の 4 指標が得られる．これらは，各既知対象 $\bar{u}_i \in U$ ごとに定まるので，$u$ が各上側和集合や下側和集合に含まれる度合いを考えるためには，指標を統合する必要がある．この際には $Cl_{d(\bar{u}_i)}^{\geq}$ と $Cl_{d(\bar{u}_i)}^{\leq}$ のそれぞれについて，すべての $\bar{u}_i \in U$ に対して同じ種類の指標を用いる．

### 3.2.3. 指標の統合

以上のように，$\bar{u}_i \in U$ ごとに $Cl_{d(\bar{u}_i)}^{\geq}$ と $Cl_{d(\bar{u}_i)}^{\leq}$ の支持度が 4 種類ずつ得られる．$s_i(Cl_{d(\bar{u}_i)}^{\geq})$ を $Cl_{d(\bar{u}_i)}^{\geq}$ の支持度，$s_i(Cl_{d(\bar{u}_i)}^{\leq})$ を $Cl_{d(\bar{u}_i)}^{\leq}$ の支持度とし，$\bar{u}_i \in U$ ごとに得られるこれらの支持度を統合するための 4 つの方法を考える．

最初に，次のような統合法が考えられる．

$$\mu^T\left(u, Cl_t^{\geq}\right) = \frac{\sum_{\substack{\bar{u}_i \in U \\ t \leq d(\bar{u}_i)}} s_i\left(Cl_{d(\bar{u}_i)}^{\geq}\right)}{\sum_{\substack{\bar{u}_i \in U \\ t \leq d(\bar{u}_i)}} s_i\left(Cl_{d(\bar{u}_i)}^{\geq}\right) + \sum_{\substack{\bar{u}_i \in U \\ d(\bar{u}_i) \leq t-1}} s_i\left(Cl_{d(\bar{u}_i)}^{\leq}\right)} \tag{6}$$

$$\mu^T\left(u, Cl_t^{\leq}\right) = \frac{\sum_{\substack{\bar{u}_i \in U \\ t \geq d(\bar{u}_i)}} s_i\left(Cl_{d(\bar{u}_i)}^{\leq}\right)}{\sum_{\substack{\bar{u}_i \in U \\ t \geq d(\bar{u}_i)}} s_i\left(Cl_{d(\bar{u}_i)}^{\leq}\right) + \sum_{\substack{\bar{u}_i \in U \\ d(\bar{u}_i) \geq t+1}} s_i\left(Cl_{d(\bar{u}_i)}^{\geq}\right)} \tag{7}$$

(6) と (7) から，$t < s$ に対して，$\mu^T\left(u, Cl_t^{\geq}\right) \geq \mu^T\left(u, Cl_s^{\geq}\right)$，$\mu^T\left(u, Cl_t^{\leq}\right) \leq \mu^T\left(u, Cl_s^{\leq}\right)$ と $\mu^T\left(u, Cl_t^{\geq}\right) + \mu^T\left(u, Cl_{t-1}^{\leq}\right) = 1$ を得る．このとき，$\mu^T\left(u, Cl_t^{\geq}\right) > \alpha$ $(\alpha \geq 0.5)$ ならば $u \in Cl_t^{\geq}$ と結論づけることができ，同様に，$\mu^T\left(u, Cl_s^{\leq}\right) > \alpha$ ならば，$u \in Cl_s^{\leq}$ と結論づけることができる．つまり，$\hat{t} = \sup\{t \in V_d \mid \mu^T\left(u, Cl_t^{\geq}\right) > \alpha\}$, $\hat{s} = \inf\{s \in V_d \mid \mu^T\left(u, Cl_s^{\leq}\right) > \alpha\}$ とすると，$u$ の決定属性値 $d(u)$ は区間 $[\hat{t}, \hat{s}]$ に含まれると推定できる．

$\mu^T$ において和を取る際に重複して足される $\bar{u}_i \in U$ が存在することがある．このような対象の影響は他の対象より相対的に大きすぎると考えられるので，重複をなくすように和集合をとる指標 $\mu^U$ も2番目の指標として考えられる．$\mu^U$ に基づく $d(u)$ の推定は，$\mu^T$ に基づく推定と同様に行うことができる．3番目の統合指標として，次を定義する．

$$v_i^u\left(Cl_{d(\bar{u}_i)}^{\geq}\right) = \begin{cases} 1, & s_i(Cl_{d(\bar{u}_i)}^{\geq}) > \alpha \\ 0, & s_i(Cl_{d(\bar{u}_i)}^{\geq}) \leq \alpha \end{cases} \tag{8}$$

$$v_i^u\left(Cl_{d(\bar{u}_i)}^{\leq}\right) = \begin{cases} 1, & s_i(Cl_{d(\bar{u}_i)}^{\leq}) > \alpha \\ 0, & s_i(Cl_{d(\bar{u}_i)}^{\leq}) \leq \alpha \end{cases} \tag{9}$$

ただし，$\alpha \geq 0.5$ とする．$V(u, Cl_t) = \sum_{d(\bar{u}_i) \leq t} v_i^u(Cl_{d(\bar{u}_i)}^{\geq}) + \sum_{d(\bar{u}_i) \geq t} v_i^u(Cl_{d(\bar{u}_i)}^{\leq})$ と定義すると，$u$ の決定属性値 $d(u)$ は，$V(u, Cl_t)$ が最大となる $t$ で決定される．このような $t$ が一意ではないとき，$d(u)$ を一つに定めることができないことに注意する．

3つ目の統合指標において，値が小さい $d(\bar{u}_i)$ に対して $s_i(Cl_{d(\bar{u}_i)}^{\geq}) > \alpha$ が成り立つ場合，一つの $\bar{u}_i \in U$ が多数の決定クラスを支持していることになり，逆に，値が大きい $d(\bar{u}_i)$ に対して $s_i(Cl_{d(\bar{u}_i)}^{\geq}) > \alpha$ が成り立つ場合は，少数の決定クラスのみ支持していることになる．このことで，既知対象間に影響の大きさに違いが生じていると考えられるため，この違いを軽減するために，次のように $v_i^u$ を $va_i^u$ で置き換えることが考えられる．

$$va_i^u\left(Cl_{d(\bar{u}_i)}^{\geq}\right) = \begin{cases} \dfrac{1}{q - d(\bar{u}_i) + 1}, & s_i(Cl_{d(\bar{u}_i)}^{\geq}) > \alpha \\ 0, & s_i(Cl_{d(\bar{u}_i)}^{\geq}) \leq \alpha \end{cases} \tag{10}$$

$$va_i^u\left(Cl_{d(\bar{u}_i)}^{\leq}\right) = \begin{cases} \dfrac{1}{d(\bar{u}_i)}, & s_i(Cl_{d(\bar{u}_i)}^{\leq}) > \alpha \\ 0, & s_i(Cl_{d(\bar{u}_i)}^{\leq}) \leq \alpha \end{cases} \tag{11}$$

$Cl_t$ を支持する度合の定義と $d(u)$ の推定は，3番目の指標と同様に行うことができる．

### 3.2.4. $k$-近傍法の導入

上述の統合法においては，すべての対象 $\bar{u}_i \in U$ に対する $s_i(Cl_{d(\bar{u}_i)}^{\geq})$ と $s_i(Cl_{d(\bar{u}_i)}^{\leq})$ を利用することを考えていた．したがって，既知対象の数が多いと計算効率が悪化する．そこで，$k$-近傍法を利用することを考える．すなわち，条件属性値が未知対象 $u$ のものと似ている $k$ 個の既知対象を選択し，それらのみを用いて上述の推定法を用いることを考える．具体的には，文献 [6] に従い，$k = \sqrt{n}$（$n$ は既知対象数）とし，各条件属性に関して，$v_1 \preceq v_2 \preceq \cdots$ のとき，$v_1, v_2, \ldots$ を $1, 2, \ldots$ と数量化し，L1 ノルムを使って距離を考える．

### 3.3. 下近似の導入

上述の統合法においては，推定に用いる条件属性が少なくなり，すべての条件属性から得た精度に比べて代替精度の信頼性が低くなると考えられるので，下近似に含まれる既知対象のみを推定に用いることで信頼性を高めることが考えられる．具体的には，たとえば，次の精度を用いることになる．

$$\underline{\mu}_P\left(u, Cl_t^{\geq}\right) = \frac{\left|D_P^-(u) \cap \underline{P}\left(Cl_t^{\geq}\right)\right|}{\left|D_P^-(u) \cap \underline{P}\left(Cl_t^{\geq}\right)\right| + \left|D_P^+(u) \cap \underline{P}\left(Cl_{t-1}^{\leq}\right)\right|}$$

この指標を統合した結果は，$\mu_P\left(u, Cl_t^{\geq}\right) = 1$ を満たすものだけを統合した結果と等しい．

表 1: 16 手法による正答率

| $\|U\|$ | TW$u$ | TW$\bar{u}_i$ | TS$u$ | TS$\bar{u}_i$ | VW$u$ | VW$\bar{u}_i$ | VS$u$ | VS$\bar{u}_i$ | DOM |
|---|---|---|---|---|---|---|---|---|---|
| 500 | $63.4_{xx}$ | 72.2 | $59.6_{xx}$ | $55.8_{xx}$ | $66.6_{xx}$ | 71.2 | $66.7_{xx}$ | $59.3_{xx}$ | 73.8 |
| 400 | $60.4_{xx}$ | 68.8 | $56.1_{xx}$ | $52.6_{xx}$ | $63.1_{xx}$ | 68.0 | $63.0_{xx}$ | $55.1_{xx}$ | 68.7 |
| 300 | $56.8_{xx}$ | 65.8 | $52.1_{xx}$ | $47.1_{xx}$ | $59.9_{xx}$ | 65.8 | $59.5_{x}$ | $51.9_{xx}$ | 64.2 |
| 200 | $51.9_{x}$ | 61.7** | $47.8_{xx}$ | $43.0_{xx}$ | 56.7 | 63.3** | 56.3 | $47.8_{xx}$ | 56.7 |
| 150 | 48.9 | 61.1** | $44.6_{xx}$ | $38.8_{xx}$ | 56.1** | 62.1** | 56.8** | $44.3_{xx}$ | 49.7 |
| 100 | 44.3* | 56.0** | 39.8 | $33.9_{xx}$ | 53.4** | 61.4** | 53.7** | 42.5 | 41.6 |
| 50 | 38.2 | 51.6** | 34.5 | 32.7 | 50.4** | 58.0** | 49.5** | 39.9 | 37.4 |
| $\|U\|$ | AW$u$ | AW$\bar{u}_i$ | AS$u$ | AS$\bar{u}_i$ | UW$u$ | UW$\bar{u}_i$ | US$u$ | US$\bar{u}_i$ | all $C$ |
| 500 | $61.3_{xx}$ | $65.1_{xx}$ | $60.9_{xx}$ | $60.8_{xx}$ | $62.9_{xx}$ | $52.3_{xx}$ | $59.0_{xx}$ | $50.6_{xx}$ | 45.0 |
| 400 | $57.6_{xx}$ | $61.8_{xx}$ | $57.1_{xx}$ | $56.9_{xx}$ | $59.1_{xx}$ | $47.6_{xx}$ | $55.0_{xx}$ | $46.5_{xx}$ | 39.4 |
| 300 | $54.4_{xx}$ | $59.1_{xx}$ | $53.7_{xx}$ | $54.2_{xx}$ | $56.0_{xx}$ | $43.0_{xx}$ | $51.0_{xx}$ | $41.9_{xx}$ | 33.5 |
| 200 | $51.4_{xx}$ | 56.6 | $50.7_{xx}$ | $50.9_{xx}$ | $52.9_{xx}$ | $37.8_{xx}$ | $48.1_{xx}$ | $37.4_{xx}$ | 27.3 |
| 150 | 49.3 | 56.2** | 48.6 | 49.9 | 51.0 | $35.4_{xx}$ | $45.9_{xx}$ | $34.6_{xx}$ | 23.3 |
| 100 | 48.0** | 55.1** | 47.2* | 48.7** | 49.6** | $31.0_{xx}$ | 44.9* | $28.6_{xx}$ | 17.6 |
| 50 | 45.6** | 54.6** | 46.1** | 45.4** | 48.3** | 30.7 | 44.5** | $22.2_{xx}$ | 12.0 |

表 2: 16 手法による誤答率

| $\|U\|$ | TW$u$ | TW$\bar{u}_i$ | TS$u$ | TS$\bar{u}_i$ | VW$u$ | VW$\bar{u}_i$ | VS$u$ | VS$\bar{u}_i$ | DOM |
|---|---|---|---|---|---|---|---|---|---|
| 500 | $36.7_{xx}$ | $27.8_{xx}$ | $40.4_{xx}$ | $44.3_{xx}$ | $33.4_{xx}$ | $28.6_{xx}$ | $33.2_{xx}$ | $38.5_{xx}$ | 19.2 |
| 400 | $39.7_{xx}$ | $31.2_{xx}$ | $44.0_{xx}$ | $47.5_{xx}$ | $36.9_{xx}$ | $31.2_{xx}$ | $36.5_{xx}$ | $42.3_{xx}$ | 22.4 |
| 300 | $43.3_{xx}$ | $34.3_{xx}$ | $48.0_{xx}$ | $52.9_{xx}$ | $39.6_{xx}$ | $33.5_{xx}$ | $39.6_{xx}$ | $43.7_{xx}$ | 26.0 |
| 200 | $48.1_{xx}$ | $38.3_{xx}$ | $52.3_{xx}$ | $57.0_{xx}$ | $42.6_{xx}$ | $35.6_{x}$ | $43.0_{xx}$ | $46.6_{xx}$ | 31.3 |
| 150 | $51.2_{xx}$ | 39.0 | $55.5_{xx}$ | $61.2_{xx}$ | $43.2_{x}$ | 36.7 | 42.7 | $47.0_{xx}$ | 38.6 |
| 100 | $55.8_{xx}$ | 44.0 | $60.3_{xx}$ | $66.1_{xx}$ | 45.6 | 36.7** | 45.0 | 47.3 | 43.7 |
| 50 | $61.6_{xx}$ | 48.5 | $65.6_{xx}$ | $67.4_{xx}$ | 47.3 | 37.9** | 46.1 | 47.8 | 49.0 |
| $\|U\|$ | AW$u$ | AW$\bar{u}_i$ | AS$u$ | AS$\bar{u}_i$ | UW$u$ | UW$\bar{u}_i$ | US$u$ | US$\bar{u}_i$ | all $C$ |
| 500 | $38.7_{xx}$ | $34.9_{xx}$ | $39.2_{xx}$ | $38.6_{xx}$ | $37.1_{xx}$ | $47.7_{xx}$ | $41.0_{xx}$ | $47.6_{xx}$ | 0 |
| 400 | $42.4_{xx}$ | $38.2_{xx}$ | $42.9_{xx}$ | $42.2_{xx}$ | $40.9_{xx}$ | $52.5_{xx}$ | $45.1_{xx}$ | $51.4_{xx}$ | 0 |
| 300 | $45.6_{xx}$ | $40.9_{xx}$ | $46.3_{xx}$ | $44.3_{xx}$ | $44.0_{xx}$ | $57.1_{xx}$ | $49.0_{xx}$ | $55.2_{xx}$ | 0 |
| 200 | $48.6_{xx}$ | $43.4_{xx}$ | $49.4_{xx}$ | $47.1_{xx}$ | $47.1_{xx}$ | $62.1_{xx}$ | $52.0_{xx}$ | $58.3_{xx}$ | 0 |
| 150 | $50.7_{xx}$ | $43.9_{x}$ | $51.4_{xx}$ | $46.8_{xx}$ | $49.1_{xx}$ | $64.5_{xx}$ | $54.1_{xx}$ | $60.3_{xx}$ | 0 |
| 100 | $51.9_{xx}$ | 44.9 | $52.8_{xx}$ | 48.0 | $50.4_{xx}$ | $68.8_{xx}$ | $55.1_{xx}$ | $63.4_{xx}$ | 0 |
| 50 | $54.4_{xx}$ | 45.1 | $53.8_{x}$ | 49.2 | 51.8 | $69.4_{xx}$ | $55.6_{xx}$ | $71.2_{xx}$ | 0 |

## 4. 数値実験

### 4.1. 人工データ作成

実験用データは，決定属性が条件属性に関して単調増加となるように，乱数を用いて，データ生成の元となる if-then ルール群を次のように生成した．決定属性値を 1 から $q$，条件属性値を 1 から $z$ までの整数値とする．ある決定属性値 $t$ 以上のルールの条件部を構成する条件属性 $a_i$ の下限値 $l_{a_i}(t)$ は，初期値を $t=q$ として，次の (i) ～ (vi) で生成する．(i) 一様乱数 $r_1 \in [0,z)$ を生成する．(ii) $t \geq q/2$ ならば，一様乱数 $r_2 \in [0, tz/q - z/2)$ を生成する．(iii) $t < q/2$ ならば，一様乱数 $r_2 \in [tz/q - z/2, 0)$ を生成する．(iv) $l_{a_i}(t) = [r_1 + r_2 + 0.5]$ とし，$l_{a_i}(t) \notin [1,z]$ ならば (i) へ戻る．ここで，$[r]$ は $r$ より小さい最大の整数とする．(v) すべての条件属性に対して $l_{a_i}(t)$ が生成されるまで，(i) から (iv) を繰り返す．生成された条件をもつルールがそれまでに生成されたルールと矛盾しなければ，生成されたルールを加え，矛盾すれば，(i) に戻り，ルールを生成しなおす．(vi) 各 $t$ に関して一定数までルールを生成し，$t=1$ ならば終了，そうでなければ，$t=t-1$ として (i) に戻る．

### 4.2. 数値実験概要

正答率 (%)，誤答率 (%)，計算時間 (ms) を数値実験で比較する．唯一のクラスが推定され，それが正しい場合を正答，推定された決定属性値の候補に正しいものが含まれない場合を誤答とし，それぞれの総数を未知対象数で割ったものを，正答率，誤答率とする．CPU: Intel(R) Core(TM) i7-870 (2.93GHz), メモリ: 2.00 GB, OS:windows 7 professional

表 3: $k$-近傍法と下近似で制限したときの推定法の正答率

| $\|U\|$ | K-TW$u$ | K$\beta$-TW$u$ | K-TW$\bar{u}_i$ | K$\beta$-TW$\bar{u}_i$ | K-VW$u$ | K$\beta$-VW$u$ | K-VW$\bar{u}_i$ | K$\beta$-VW$\bar{u}_i$ | DOM |
|---|---|---|---|---|---|---|---|---|---|
| 500 | 74.5 | 76.3* | 75.6 | 76.5** | 72.4 | 76.8** | 72.2 | 77.5** | 73.8 |
| 400 | 72.5** | 73.0** | 73.5** | 73.8** | 69.6 | 73.8** | 69.8 | 74.1** | 68.7 |
| 300 | 69.3** | 70.3** | 70.4** | 70.3** | 66.7 | 70.8** | 67.3* | 70.7** | 64.2 |
| 200 | 66.9** | 65.4** | 67.4** | 65.7** | 62.9** | 66.3** | 63.8** | 66.5** | 56.7 |
| 150 | 63.9** | 63.1** | 65.7** | 63.4** | 60.5** | 64.0** | 61.7** | 64.3** | 49.7 |
| 100 | 61.2** | 59.9** | 62.7** | 58.7** | 58.1** | 61.2** | 59.0** | 59.7** | 41.6 |
| 50 | 55.8** | 53.8** | 58.7** | 53.1** | 51.7** | 55.1** | 52.8** | 54.0** | 37.4 |

表 4: $k$-近傍法と下近似で制限したときの推定法の誤答率

| $\|U\|$ | K-TW$u$ | K$\beta$-TW$u$ | K-TW$\bar{u}_i$ | K$\beta$-TW$\bar{u}_i$ | K-VW$u$ | K$\beta$-VW$u$ | K-VW$\bar{u}_i$ | K$\beta$-VW$\bar{u}_i$ | DOM |
|---|---|---|---|---|---|---|---|---|---|
| 500 | $25.5_{xx}$ | 17.3 | $24.4_{xx}$ | 13.0** | $24.5_{xx}$ | 16.3** | $23.7_{xx}$ | 12.2** | 19.2 |
| 400 | $27.6_{xx}$ | 19.7** | $26.5_{xx}$ | 16.1** | $26.7_{xx}$ | 18.8** | $25.9_{xx}$ | 15.0** | 22.4 |
| 300 | $30.8_{xx}$ | 21.7** | $29.6_{x}$ | 17.9** | $28.6_{x}$ | 20.9** | 27.3 | 17.0** | 26.0 |
| 200 | 33.1 | 24.6** | 32.7 | 22.1** | 30.6 | 23.4** | 28.9 | 20.9** | 31.3 |
| 150 | 35.7* | 26.3** | 34.3** | 24.0** | 31.8** | 25.1** | 29.9** | 22.5** | 38.6 |
| 100 | 38.0** | 29.7** | 37.1** | 26.8** | 33.0** | 27.5** | 31.7** | 25.4** | 43.7 |
| 50 | 43.7 | 32.7** | 40.9* | 31.0** | 33.2** | 31.2** | 33.4** | 29.9** | 49.0 |

表 5: $k$-近傍法と下近似で制限したときの推定法の計算時間

| $\|U\|$ | K-TW$u$ | K$\beta$-TW$u$ | K-TW$\bar{u}_i$ | K$\beta$-TW$\bar{u}_i$ | K-VW$u$ | K$\beta$-VW$u$ | K-VW$\bar{u}_i$ | K$\beta$-VW$\bar{u}_i$ | DOM |
|---|---|---|---|---|---|---|---|---|---|
| 500 | 212.6** | 532.9** | 212.6** | 331.3** | 215.8** | 542.2** | 217.3** | 326.7** | 1133.7 |
| 400 | 168.8** | 395.5** | 162.6** | 250.1** | 173.6** | 403.2** | 168.9** | 257.9** | 731.7 |
| 300 | 118.8** | 256.4** | 115.8** | 170.4** | 117.3** | 257.9** | 122.0** | 178.2** | 444.1 |
| 200 | 68.9** | 143.7** | 70.4** | 101.5** | 72.0** | 143.7** | 72.0** | 101.6** | 214.3 |
| 150 | 51.7** | 93.8** | 51.6** | 67.2** | 51.7** | 96.8** | 62.6** | 75.0** | 139.2 |
| 100 | 29.8** | 54.7** | 32.9** | 42.2** | 32.9** | 54.7** | 34.5** | 45.3** | 73.5 |
| 50 | 15.7* | 20.3 | 14.1* | 18.8 | 15.6* | 21.9 | 14.0* | 18.8* | 26.6 |

のパーソナルコンピュータ上の 'JavaSE-1.6'(P)(jre1.6) で実験した.

条件属性数が 6 で，各条件属性値集合および決定属性値集合が $\{1,2,3,4,5\}$ となる 5 種類の人工のデータセットを上述の方法で生成した．既知対象数に関する各推定法の性能の違いをみるため，既知対象数が 50, 100, 150, 200, 300, 400, 500 となる決定表をランダムに 10 ずつ生成した．各決定表に対して，200 個の未知対象の決定属性値を推定し，正答率，誤答率，計算時間を求めた．各対象数ごとに，正答率，誤答率，計算時間の平均と分散を求め，DOMLEM を用いた場合を含めて手法間の優位性を検定した.

どのデータセットも類似した結果であったので，この論文では，ある一つのデータセットの実験結果を示す．提案法はアルファベットの組み合わせで表す．'T', 'U', 'V', 'A' は，それぞれ $\mu^T$, $\mu^U$, $V$, $Va$ を統合の際に用いたことを表す．'S' は $P_u^{\geq}(\bar{u}_i)$ と $P_u^{\leq}(\bar{u}_i)$ を用いたことを表し，'W' は $P_u^{\geq}(\bar{u}_i)$ と $P_u^{\leq}(\bar{u}_i)$ を用いたことを表す．指標 $\mu_{P_u^{\geq}(\bar{u}_i)}$, $\mu_{P_u^{>}(\bar{u}_i)}$, $\mu_{P_u^{\leq}(\bar{u}_i)}$, $\mu_{P_u^{<}(\bar{u}_i)}$ における第一引数に未知対象 $u$ を用いる場合を '$u$', 既知対象 $\bar{u}_i$ を用いる場合を '$\bar{u}_i$' と記すこととした．'K-' は，$k$-近傍法を用いたことを示し，'K$\beta$-' は，$k$-近傍法と下近似による制限の両方を用いたことを示す.

### 4.3. 16 提案法の結果

$\alpha = 0.5$ とし，$k$-近傍法も下近似による制限も行わなかった場合の 16 提案法の結果を表 1 と 2 に示す．'$|U|$', 'DOM', 'all $C$' は，それぞれ，対象数，DOMLEM [3] で導かれたルールを用いた推定，すべての条件属性からの推定のみによる推定を表す．DOM においては，矛盾した結果は誤答として誤答率に加えている．上つきの ** (*) と下つきの xx (xx) は，それぞれ，有意水準 1% (5%) で提案法が DOM より優れていると言えるか，悪くなっていると言えるかということを表す.

表 1 と 2から，$|U|$ が小さいときにいくつかの提案法が正答率で DOM より優れていたが，残念ながらすべての提案法が非常に優れているとは言えない．これは，支持の度合いが小さい既知対象も含めて，すべての既知対象から得られる指標を加算してしまっていることに起因していると考えられる．そこで，条件属性値の点で未知対象に近い既知対象のみを用いることや，支持度の高いものだけを推定に用いる方がよいと考え，次で，$k$-近傍法と下近似による制限を用いた実験を行う．

提案法の中で結果を比較すると，'W' の方が 'S' より優れていることがわかる．さらに，'U' や 'Va' は 'T' や 'V' よりも優れているようには見えない．そこで，提案法のうち 'TW' や 'VW' について $k$-近傍法や下近似で制限した手法について数値実験を行う．

### 4.4. $k$-近傍法と下近似による制限

表 3 〜 5 は，'TW' や 'VW' に $k$-近傍法や下近似を適用した方法の結果である．これらの表から，いくつかの提案法が優れていることがわかる．$k$-近傍法と下近似による制限の導入を行った方が結果が優れている．特に，K$\beta$-TW$u$ と K$\beta$-VW$u$ が，多くのケースにおいてすべての基準で DOM より有意に優れている．K$\beta$-TW$u$, K$\beta$-VW$u$, K$\beta$-TW$\bar{u}_i$ K$\beta$-VW$\bar{u}_i$ では，$|U|$ が大きい場合にも DOM よりも優れている．これらの 4 手法のうちの一つを用いることが応用に有効であろう．

## 5. おわりに

支配関係に基づく精度を用いた事例ベースクラス推定法を提案した．数値実験により，すべての既知対象を用いた提案法は必ずしもいいとは言えなかったが，$k$-近傍法と下近似による制限を用いた提案法はルール抽出アルゴリズム DOMLEM による推定法よりも優れていることがわかった．この研究では，決定表が整合していることを仮定していたので，整合しないデータを含む場合に関する数値実験を行うことも一つの課題である．さらに，下近似の条件を緩和した場合についての議論も今後の課題である．

**謝辞** 山本俊介氏の数値実験における協力に謝意を表する．

### 参考文献

[1] J.G. Bazan, "Discovery of decision rules by matching new objects against data tables," in *Proceedings of RSCTC 1998*, 1998, pp.521–528.

[2] H.S. Nguyen, "Scalable classification method based on rough set," in *Proceedings of RSCTC 2002*, LNAI 2475, Springer, Berlin, 2002, pp.433–440.

[3] S. Greco, B. Matarazzo, R. Słowiński, "Rough approximation by dominance relations," *Int. J. Intel. Syst.*, vol.17, no.2, pp.153–171, 2002.

[4] K. Dembczyński, R. Pindur, R. Susmaga, "Dominance-based rough set classifier without induction of decision rules," *Electronic Notes in Theoretical Comput. Sci.*, vol.82, no.4, pp.87–98, 2003.

[5] M. Inuiguchi, Y. Yoshioka, Y. Kusunoki, "Variable-precision dominance-based rough set approach and attribute reduction," *Int. J. of Approx. Reason.*, vol.50, no.8, pp.1199–1214, 2009.

[6] D.O. Loftgaarden, C.P. Queensbrerry, "A nonparametric estimate of a multivariate density function," *Annals of Math. Stat.*, vol.36, pp.1049–1051, 1965.