条件属性に関して非単調な決定表からの ラフ集合に基づくルール抽出アルゴリズム

学籍番号：90152095 乾口研究室 戸所雄一

1. はじめに

ラフ集合理論は，決定表から最小限必要な属性や極小な決定ルールを抽出する手法として有用である．一般に，すべての極小なルールを求めるには膨大な計算時間が必要となるので，各対象を少なくとも一つのルールにより説明できる範囲で極小なルール群の抽出が行われることが多い．そのようなアルゴリズムの一つとして，LEM2 が提案されている．従来のラフ集合が同値関係の下で定義され，名義的な属性を仮定しているのに対し，支配関係の下でのラフ集合やそれに基づく序数属性を含む決定表の解析手法が，Greco らにより提案されている．

しかし，Greco らの方法では条件属性と決定属性間の単調性を仮定しているため，非単調な，つまり「程よい値が好ましい」といったような条件属性をうまく取り扱うことができない．そこで，本研究では条件属性に関して非単調な決定表に適したラフ集合を考え，これに基づくルール抽出法を考察する．すなわち，LEM2 を拡張した INTLEM を提案する．

2. ラフ集合の一般化

本研究では次式で定義される集合族 $\mathcal{F}_A = \{F_1, F_2, \ldots, F_q\}$ に基づいてラフ集合を定義する．ただし，A は序数条件属性 C の任意の部分集合とする．

$$\mathcal{F}_A = \{\langle V \rangle_A \mid V \subseteq U\} \quad (1)$$

$$\langle V \rangle_A = \{x \in U \mid \min_{y \in V} f(y,a) \leq f(x,a) \leq \max_{y \in V} f(y,a)\}, \exists a \in A \cap C\} \quad (2)$$

このとき，下近似，上近似は次のように定義される．

$$\mathcal{F}_*(X) = \bigcup \{F_i \mid F_i \subseteq X, i = 1,2 \ldots q\} \quad (3)$$

$$\mathcal{F}^*(X) = U - \mathcal{F}_*(U - X) \quad (4)$$

ラフ集合は，対 $(\mathcal{F}_*(X), \mathcal{F}^*(X))$ によって定義される．

3. INTLEM

LEM2 を拡張することにより，以下で述べる INTLEM が構成できる．入力 X は目標クラスの上近似か下近似であり，出力 E はそれに応じた決定ルールの集合である．INTLEM では，条件集合 E のすべての条件を満たす対象の集合が目標クラスに含まれるまである評価に従い条件 e を加えていくことにより条件集合 E を求める．次に，E から冗長な条件を除き決定ルールの条件部とする．この操作を決定表内のすべての対象が説明されるまで繰り返すことにより決定ルールの集合 E を求める．さらに，E から冗長な決定ルールを除くという 3 段階のプロセスを経て，極小な決定ルール群が生成される．下のアルゴリズムでは，Cond は追加する条件部の候補集合で，各条件部は「属性値がある区間に入る」という形式で表されている．条件集合 E に対して [E] は，E のすべての条件を満たす対象の集合を表し，E に追加する条件 e の選択はその条件を追加した場合の条件集合 $\{e\} \cup E$ の評価 evaluate($\{e\} \cup E$) に従ってなされる．

```
Procedure INTLEM
(input: a set of X,output: a set of rules E covering X)
begin
  G := X; E:=∅;
  while (G ≠ ∅) do begin
    E := ∅; S:=G;
    while(E=∅) or ([E] ⊈ X) do begin
      best := ∅
      for each attribute q ∈ C do begin
        Cond := Cond ∪ {f(x,q) ∈ [min(f(x_i,q), f(x_j,q)),
                                  max(f(x_i,q), f(x_j,q))] | x_i, x_j ∈ S}
      end;{for}
      for each e ∈ Cond do begin
        if evaluate({e} ∪ E) is better than
            evaluate({best} ∪ E) then best := e;
      end;{for}
      E := E ∪ {best}; S:=S∩[best];
    end;{while([E] ⊈ X)}
    for each elementary condition e ∈ E do begin
      if [E − {e}] ⊆ X then E := E − {e};
    end;{for}
  E:=E∪{E}
  G := B − ∪_{E∈E}[E];
  end;{while(G ≠ ∅)}
end {function};
```

条件集合 E の評価 evaluate(E) は，E のすべての条件を満たす目標クラス内の対象数 $|X \cap [E]|$ ，あるいは正確度 $|X \cap [E]|/|[E]|$ など複数の評価基準の辞書式順序により評価される．

4. 数値実験

INTLEM の有効性および適切な evaluate(E) を調べるため，数値実験を行った．ここでは，ワインの識別に関するデータを用いた場合の結果を示す．このデータの対象数は 178，序数性を含む条件属性数が 13，決定属性数が 1 で，決定クラスの数は 3 である．5-fold cross validation 法により 5 種類の決定ルール群を求め，決定ルール数，各決定ルールの条件部長の平均，各ルールを支持する対象数の平均，識別率，誤識別率，矛盾の割合，不明の割合を求めた．evaluate(E) として，ここでは次の二つを取り上げて報告する．

- INTLEM1 : evaluate(E) = $(|X \cap [E]|, |X \cap [E]|/|[E]|)$
- INTLEM2 : evaluate(E) = $(|X \cap [E]|/|[E]|, |X \cap [E]|)$

目標クラスとして決定クラスの下近似データを用いた場合の結果を表 1 に示す．表 1 の各数値は 5-fold cross validation 法で評価した値の平均を表している．

表 1: ワインの識別に関するデータの解析結果

	INTLEM1	INTLEM2	LEM2
決定ルール数 (個)	10.8	58.4	87
平均条件部長 (個)	3.27	1.00	1.27
平均支持数 (個)	41.2	3.89	1.75
識別率 (%)	82.5	61.7	16.3
誤識別率 (%)	1.67	9.56	15.2
矛盾の割合 (%)	3.98	15.8	5.60
不明の割合 (%)	11.9	13.0	62.9
計算時間 (秒)	185.8	41.6	1.42

識別率は抽出された決定ルール群により正確に識別された対象の割合，誤識別率は誤って識別された対象の割合を示している．これらの値について INTLEM と LEM2 とを比較することにより，INTLEM の方が良好な結果を得られていることがわかる．また，平均支持数をみると INTLEM2 より INTLEM1 の方がより多くの対象を含む幅の広い決定ルールを抽出していることがわかる．さらに、決定ルール数および識別率に示されるとおり，INTLEM2 より INTLEM1 の方がルール数を抑えられ，正答率も高くなっている．しかし，今回のデータのように条件属性数が多い場合には，評価関数の性質上 INTLEM1 では多くの計算時間がかかってしまう場合があることがわかる．

5. おわりに

本研究では，条件属性に関して非単調な決定表からのルール抽出法として INTLEM を提案し，その有用性を示した．今後の課題としては，新たな評価関数 evaluate(E) の考案と実験による評価，また，アルゴリズムの更なる改善が考えられる．

参考文献

[1] J. W. Grzymala-Buse: “LERS-A system for Learning from Examples based on Rough Sets”, in R. Slowinski (ed), Intelligent Decision Support. Handbook of Application and Advances of the Rough Set Theory, Kluwer Academic Publishers, pp. 3-18, 1992.