結論部の含意関係を考慮したルール抽出法に関する考察

学籍番号：90176018　乾口研究室　井上 正則

1. はじめに

決定表に内在するルールを抽出する方法として，ラフ集合理論に基づいた種々の方法が提案されている．決定属性に序数性が存在する場合には，直接，決定属性値が v_d であることを導くルールを抽出するよりも，「v_d 以上」および「v_d 以下」であることを導くルールを抽出するほうがルールの条件部が簡潔になり，有用性が高くなる．しかし，通常，複数の決定属性値に対して独立にルール群が抽出されるため，決定属性の序数性により結論部に含意関係が成立していたとしても，条件部にも含意関係が成立するとは限らないとう不自然な結果が得られることがある．ルールの結論部の含意関係を反映した条件部をもつルール群を抽出する三つのアプローチが既に議論されている [2].

本研究では，これらのアプローチについてさらに考察する．ルール抽出法を改良するとともに，従来，「v_d 以上」を導くルール抽出しか考察されていなかったのに対し，「v_d 以下」を導くルール抽出も併用し，決定属性値推定における有用性を数値実験により検討する．

2. 結論部の含意関係を考慮したルール抽出アプローチ

複数のルールの結論部間の含意関係を条件部に反映したルール群を抽出する方法として，制限法，緩和法，優先順位法の三つが考えられている [2].

制限法は，緩い結論部をもつルールを先に抽出し，得られたルールの条件部をさらに強化することにより，より強い結論部をもつルールを抽出していく方法である．緩和法は，制限法とは逆に，強い結論部をもつルールを先に抽出し，得られたルールの条件部を緩和することにより，より緩い結論部をもつルールを抽出していく方法である．優先順位法は，ルール抽出アルゴリズム MLEM2[1] におけるルールの条件選択基準に，決定属性値の序数性を考慮したものである．ある結論部をもつルールの条件を選ぶ際，最も強い結論部に関する基準から現在の結論部の基準までを辞書式に適用することにより条件選択を行う方法である．

制限法と緩和法は，ルール抽出アルゴリズムに依存しないアプローチであり，必ず緩い結論部をもつルールの条件部が強い結論部をもつものよりも弱くなる．一方，優先順位法は MLEM2 の考え方に基づくアプローチであり，必ずしも緩い結論部をもつルールの条件部が強い結論部をもつものよりも弱くなるとは限らない．MLEM2 のアルゴリズムを一部修正することにより，制限法と緩和法とが実現できる．MLEM2 に基づく制限法と緩和法と比べると，優先順位法は計算時間が短いという特徴がある [2].

本研究では，従来法に不要なルールを取り除く機能を追加した MLEM2 に基づいた制限法，緩和法および優先順位法を用いる．「v_d 以上」を導くルールと「v_d 以下」を導くルールを抽出する際，表 1 の 5 通りの方法を考える．

抽出されたルールに基づく未知データの識別は，次の方法による．(1) 各上下累積集合についての支持度を求める．(2) この支持度に基づき，対立する二つの補完的な累積集合間の勝者を求める．(3) 各クラスの評価値を勝者集合から定め，最も高い評価値を与えるクラスとして識別する．

表 1: 五つの方法

手法	v_d 以上	v_d 以下
(1) 制限-制限法	制限法	制限法
(2) 制限-緩和法	制限法	緩和法
(3) 緩和-制限法	緩和法	制限法
(4) 緩和-緩和法	緩和法	緩和法
(5) 優先順位法	優先順位法	優先順位法

3. 数値実験

五つの方法を比較するため，数値実験を行った．n 次元実数空間上に，入れ子状になった複数の超直方体 H_i, $i = 1, 2, \ldots, k$ ($H_1 \supseteq H_2 \supseteq \cdots \supseteq H_k$) を数組発生させ，各組を "if $x \in H_i$ and $x \notin H_{i-1}$ then the decision attribute value of x is v_i" というルールと見做した．各属性は整数値を取るとし，生成したルールがカバーするすべてのパターンのデータを収集した．なお，複数の組のルールに適合した対象 x の決定属性値は各ルールから得られる v_i の最も大きな値とした．また，無関係な条件属性による影響をみるため，作為的に，ランダムな値をとる条件属性をデータに追加した．収集したデータの 1%, 2%, ..., 9%, 10%, 20%, ..., 90% を訓練用データとしてサンプリングし，五つの方法でルール抽出を行い，残りのデータで検証する実験を，各割合について 100 回行った．

表 2 に条件属性数が 5，条件属性値数が 6，クラス数が 4 で三つの入れ子状の超直方体に基づくデータに適用した結果を示す．紙面の制約により，2%, 4%, 6%, 8%, 10% のみの正答率の平均値と標準偏差を示している．手法 (0) は従来の MLEM2 を，手法 (1)〜(5) は表 1 の手法を示している．手法 (1)〜(5) については，正答率の平均値が手法 (0) と変わらないという帰無仮説のもとでの対応のある t-検定も行い，有意水準 5%で帰無仮説が棄却されなかったものには，$*$ を付している．表 2 に示すとおり，制限-緩和法が最も良く，t-検定でも有意な差が現れている．逆に，緩和-制限法はあまり芳しくなく，MLEM2 に劣っている．他は MLEM2 と大差ないか，それ以下である．本実験のデータ生成法を考えると，峰状や谷状の階層的ルールを抽出するには，制限法が良く，逆に，平面から峰や谷が抜けた階層的ルールを抽出するには，緩和法が良いようである．

最後に，無関係な条件属性を加えた場合の結果を大まかに述べると，正答率は低下するが，40%以上のデータが得られていれば，提案したルール抽出法によりほぼ完全にデータを再現することができることがわかった．

表 2: 数値実験結果

割合 (%)	2	4	6	8	10
手法 (0)	71.06 ± 5.14	87.46 ± 3.49	94.28 ± 2.47	96.8 ± 1.73	98.22 ± 1.03
手法 (1)	67.09 ± 6.92	85.58* ± 4.29	92.92 ± 3.48	96.19 ± 1.89	97.86 ± 1.12
手法 (2)	73.91 ± 6.06	91.18 ± 3.64	96.6 ± 2.16	98.15 ± 1.22	98.9 ± 0.64
手法 (3)	62.5* ± 5.94	82.71* ± 5.21	91.6 ± 3.53	95.45 ± 2.39	97.38 ± 1.48
手法 (4)	67.64 ± 5.68	87.73 ± 4.86	95.04 ± 2.99	97.39 ± 1.62	98.47 ± 1.12
手法 (5)	70.41 ± 5.65	87.89 ± 3.88	94.89 ± 2.48	97.26 ± 1.53	98.5 ± 0.92

4. 結論

結論部の含意関係が条件部にも成立するようなルール抽出法について数値実験を行い，各手法の有用性を検討した．

参考文献

1. J.W.Grzymala-Busse: "MLEM2-Discretization During Rule Induction", *International Conference on Intelligent Information Processing and WEB Mining Systems*, Springer-Verlag, pp.499-508, 2003.
2. 松本圭祐: 序数決定属性をもつ決定表からのルール抽出，平成 18 年度大阪大学基礎工学部基礎工学研究科修士論文 (2007)