F-019

ラフ集合のルール抽出結果に基づく属性間因果関係の推定
Estimation of Causal Relationship for Attributes based on Association Rules Extracted by Rough Set

宮崎 光二†
Koji Miyazaki

1. はじめに

近年，様々な分野でビッグデータが取り扱われるようになり，そこから有益な情報を取り出すデータマイニングが注目されている．ラフ集合は与えられたデータベースからカテゴライズに必要な情報を保ちつつデータ表現の次元数を縮小し，データを簡潔なルール群で表現することにより，ビッグデータから有益な情報を抽出することが期待され，データの特徴を理解する手助けとなる．本稿ではラフ集合の if-then 形式で得られるルール情報を利用し，条件属性間の因果関係を推定する方法を提案し，データの性質や特徴を明らかにすることを試みる．ベイジアンネットワークはノード間で因果関係を持ち，影響度合いを条件付確率で定量化した有向非循環グラフであり，医療診断や機器の故障診断，意思決定など不確実性を含むモデリングに用いられる．ベイジアンネットワークのノード間には親と子の因果関係があり，基本的に人間が適切に設定する必要がある．適切な親子関係の設定がされていれば，ベイジアンネットワークは有効に働くが，機器の故障診断のようにノードの位置関係の設定が難しい場合も多い．本稿ではノードをラフ集合におけるデータの属性とみなし，ノード間の因果関係（親子関係）をラフ集合のルール抽出結果を用いて推定する方法を提案する．

2. ラフ集合

ラフ集合では条件属性と決定属性からなるデータを，決定表と呼ぶ．条件属性の値の組み合わせと決定属性の値に規則性がある場合，それをルールとして抽出することができる．表１はインフルエンザの診断を簡単な決定表で表した例である．条件属性は体温，渇いた咳，頭痛，筋肉痛であり，x_1～x_9 の９人の患者のデータを表している．ラフ集合はデータを正しく分類するために最小限必要となる条件

表１：インフルエンザ診断の決定表

患者	条件属性				決定属性
	体温	渇いた咳	頭痛	筋肉痛	インフル
x_1	正常	なし	なし	なし	なし
x_2	正常	なし	あり	あり	なし
x_3	微熱	なし	あり	あり	あり
x_4	微熱	あり	あり	なし	あり
x_5	微熱	なし	なし	あり	あり
x_6	微熱	なし	なし	あり	なし
x_7	高熱	あり	なし	なし	あり
x_8	高熱	なし	あり	あり	あり
x_9	高熱	あり	あり	あり	あり

† 福山大学

表２：インフルエンザ診断の相対縮約

患者	条件属性		決定属性
	体温	頭痛	インフル
x_1	正常	なし	なし
x_2	正常	あり	なし
x_3	微熱	あり	あり
x_4	微熱	あり	あり
x_5	微熱	なし	あり
x_6	微熱	なし	なし
x_7	高熱	なし	あり
x_8	高熱	あり	あり
x_9	高熱	あり	あり

属性の一部を用いて if-then ルールを作成する．最小限必要な属性の組み合わせは一般的に複数個あり，それを相対縮約と呼ぶ．最も属性数の少ない相対縮約を用いてルール抽出を行うと，簡潔なルールで決定表の規則性を見出すことができる．表２は表１の相対縮約の決定表である．条件属性の「渇いた咳」と「筋肉痛」は条件に入れなくても決定属性による分類分けに影響がないので削除することができる．表２より抽出したルールは

Rule 1：if (体温=正常) then (インフル=なし)
Rule 2：if (体温=高熱) then (インフル=あり)
Rule 3：if (体温=微熱) and (頭痛) then (インフル=あり)

のようになる．

3. 属性間の因果関係

3.1 関連性の強弱

ラフ集合のルール抽出結果により，ある条件属性の組み合わせと決定属性値の因果関係を考える．一般的にラフ集合は条件属性が「原因」で決定属性が「結果」を示す構造をしており，前節では４個の条件属性「体温」「乾いた咳」「頭痛」「筋肉痛」の値の組み合わせを基にインフルエンザかどうかの分類を行っていた．あるルールに含まれている条件属性の個数（種類）が少ない場合は，１個あたりの条件属性から決定属性に作用する因果が大きいことになり，それらの条件属性と決定属性は関連性が強いと考えられる．逆に条件属性の個数が多い場合は関連性が弱いと考えるものとし，この強い・弱いを定量化することで条件属性および決定属性の因果関係の指標を新たに提案する．例えば，Rule1 と Rule2 において条件属性の「体温」が正常または高熱であれば，インフルエンザの有無を診断することができるので，「体温」と「インフルエンザ」の因果の関連性は強い．Rule3 では条件属性の「体温」と「頭痛」の条件が両方ともそろわなければインフルエンザの有無を診断することはできないので，Rule3 における「体温」と「頭痛」

のインフルエンザに対する各々の因果関係は Rule1，Rule2 よりは高くないと判断する．また，抽出された複数のルールにおいて，多くのルールに出現している条件属性も決定属性との因果関係において関連性が強いとみなすことができると考える．

3.2 因果関係

関連性の強弱の他に，ある２個の属性間の関係について「原因」と「結果」の因果の設定も重要である．表１の決定表では，インフルエンザの有無が決定属性となっており，「体温が高いからインフルエンザである」と主張できる一方で，「インフルエンザだから体温が高い」のように原因と結果を逆に考えることができる．表１の決定表において条件属性と決定属性の種別を無くした５個の属性を素属性と呼ぶことにする．素属性から決定属性を１個選び，他のの素属性を条件属性としてルール抽出をすると，決定属性の素属性を「結果」，他の素属性を「原因」とする因果関係を調べることができる．すべての素属性を決定属性としてそれぞれにルール抽出を行い，定量化した因果関係を統合して，すべての素属性間の因果関係を推定する．

4. 実験

4.1 因果関係の定量化

表３はあるデータを数値に置き換えた決定表である．この決定表を用いて，素属性の因果関係を算出する．表３の決定表から得られる相対縮約のうち，最も属性数の少ない相対縮約は c1，c4，c5，c9 である．この縮約でルールを抽出すると，図１のように８個のルールが抽出される．図１の「*」は任意の値を意味し，例えば，ルール[8]は c4=1

表３：決定表

U	c1	c2	c3	c4	c5	c6	c7	c8	c9	class
x1	0	6	1	1	1	2	2	0	1	1
x2	0	6	0	1	1	1	2	1	1	1
x3	0	6	0	1	1	2	2	1	1	1
x4	0	4	1	1	1	2	2	1	0	2
x5	0	6	0	1	1	1	1	1	1	1
x6	0	6	0	2	1	1	1	0	2	0
x7	0	6	0	1	1	1	2	1	2	0
x8	1	4	0	2	0	2	0	1	0	2
x9	0	4	0	2	0	2	0	1	1	1
x10	0	4	0	2	0	2	1	0	1	1
x11	1	4	0	1	0	2	0	1	0	2
x12	1	4	0	1	1	1	1	1	1	2
x13	0	4	0	2	1	1	1	1	1	1
x14	1	4	1	1	0	2	2	1	1	2
x15	1	4	0	2	0	1	0	1	1	2
x16	0	4	1	1	1	1	2	1	1	1
x17	0	6	0	1	1	1	2	0	1	1
x18	0	4	0	1	1	1	2	0	1	1
x19	1	4	0	1	0	2	1	1	1	2
x20	0	4	0	1	0	2	1	1	1	2
x21	0	4	0	2	0	2	1	1	1	1

```
                c1  c4  c5  c9     class
-------------------------------------------
ルール [ 1] :  0   *   1   1  |     1
ルール [ 2] :  *   *   *   0  |     2
ルール [ 3] :  *   *   *   2  |     0
ルール [ 4] :  1   *   *   *  |     2
ルール [ 5] :  0   2   *   1  |     1
ルール [ 6] :  0   2   0   *  |     1
ルール [ 7] :  *   2   1   1  |     1
ルール [ 8] :  *   1   0   *  |     2
```

図１：ルール抽出の実行結果

かつ c5=0 ならば class=1 であり，c1 と c9 はどの値でも構わない．因果関係を表す値は，縮約で得られた条件属性の個数を N_C，ルール数を N_R，ルール[i]（i=1, 2, ･･, N_R）の中に含まれる「*」以外の条件属性値の個数を N_i，ルール[i]における条件属性 cj(j=1, 2, ･･, N_C)のデータを $d_{i,j}$ とすると，class を因果関係の「結果」とした条件属性 cj(j=1, 2, ･･, N_C)を「原因」とする関連性を表す値 $V_{cj,class}$ (j=1, 2, ･･, N_C)を次式で計算する．

$$V_{cj,class} = \sum_{j=1}^{N_R} \frac{a}{N_i}, \left(a = \begin{cases} 0 & if\ d_{i,j} = * \\ 1 & if\ d_{i,j} \neq * \end{cases} \right) \qquad (1)$$

すべての素属性を順番に決定属性にした決定表を作成し，式(1)により $V_{cj,class}$ を計算した結果を合計する．表３の決定表では，素属性が 10 個なので式(1)による計算が 10 回行われることになる．

4.2 実行結果

以下に，表３の決定表を用いた実行結果の一部を示す．右側の数値は式(1)で計算した値を合計したもので，その値を基に降順に並び替えている．

```
1st: [c4] --> [c7] = 5.250000
2nd: [class] --> [c7] = 4.333333
3rd: [c5] --> [c6] = 4.083333
4th: [c9] --> [c7] = 3.833333
5th: [c8] --> [c7] = 3.666667
6th: [c7] --> [c4] = 3.666667
7th: [c7] --> [c8] = 3.666667
8th: [class] --> [c9] = 3.666667
                :
```

式(1)で計算した値の合計が大きいものが，因果関係が強い．1st は属性 c4 が原因で属性 c7 が結果の因果関係が強く表れていることを意味する．

5. おわりに

データベースの属性間には何かしらの関連性が存在するが，原因と結果が入れ替わることも多く，その因果関係や関連の強さを調査することは容易ではない．本稿では，ラフ集合の if-then 形式で得られるルール情報を用いて，条件属性間の因果関係を推定する方法を提案した．提案手法を用いることにより，ベイジアンネットワークの親・子を設定するための一助になると考えられる．ラフ集合で得られたルールの有効性を測る指標でサポートや C.I.値などがあり，それらを用いた方法も今後検討する．