2003年日本オペレーションズ・リサーチ学会
春季研究発表会

ラフセットにおける近似リダクトについて

02992020　東京工業大学　*郭　天放　KUO Tien-Fang
01703730　東京工業大学　矢島　安敏　YAJIMA Yasutoshi
01601360　慶応義塾大学　森　雅夫　MORI Masao

1　はじめに

ラフセット理論は、デシジョンテーブルから知識発見を行うための手法の一つである。デシジョンテーブルからリダクトと呼ばれる属性の部分集合を探し、決定ルールを生成できる。しかし、リダクトの計算には高速なアルゴリズムがないこと、また、ノイズが数多く入ったデシジョンテーブルに対しては分類能力が弱く、適応できないといった問題点がある。

本研究では近似的なリダクトを求めるために、幾つかの指標を導入する。これらの指標を少ない計算量で求めるために分割表を導入する。この分割表は、各クラスの対象数を基に作られたものである。得られた近似的なリダクトは、ノイズが入ったデシジョンテーブルに対しても代表性が強いものである

2　ラフセットとリダクト

デシジョンテーブルは対象の集合 (U)、条件属性の集合 (C)、と決定属性の集合 (D) から構成されて、$(U, C \cup D)$ で表す。C の各属性のとる値の組（ベクトル）により U が分割できる。その分割を $C^* = \{X_{C_1}, X_{C_2}, \cdots, X_{C_m}\}$ とおく。D に関しては、その決定値により U を分割する。その分割を $D^* = \{Y_{D_1}, Y_{D_2}, \cdots, Y_{D_n}\}$ $(n = |D|)$ とおく。リダクトとは C の部分集合であって、C と同じ分類能力を持って、条件属性の本質的な部分である。リダクトから決定ルールを生成できて、パターンの分類と予測で使われる。

しかし、ノイズが数多く入ったデシジョンテーブルに対して、リダクトの分類能力が弱く、適応できないといった問題点がある。それに現在まで提案された近似的なリダクトは、時間計算量が大きいといった問題がある。

3　分割表

高速のアルゴリズムを作るために、本研究は分割表を導入する。この分割表は、各クラスの対象数を基に作られたものである。もとのデシジョンテーブルは C に応じて、分割表 $CT(C)$ に変換する。

$$CT(C) = \begin{array}{c} \\ Y_{D_1} \\ Y_{D_2} \\ \vdots \\ Y_{D_n} \end{array} \begin{array}{c} \begin{array}{cccc} X_{C_1} & X_{C_2} & \cdots & X_{C_m} \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \end{array}.$$

ここで、列と行の数それぞれは $|C^*|$ と $|D^*|$ である。各成分 x_{ij} はそれぞれのクラス $X_{C_i} \cap Y_{D_j}$ の対象数である。また、分割表を生成する時間計算量は $O(|C||U|log|U|)$ である。

後述の指標を計算するために、分割表から、必要なパラメーターを定義する。

$$Row_i(C) = [x_{i1}, \cdots, x_{im}]^T, i = \{1, \cdots, n\}$$
$$Col_i(C) = [x_{1i}, \cdots, x_{ni}]^T, i = \{1, \cdots, m\}$$
$$NCol_i(C) = Col_i(C)^T \cdot e$$
$$p_i(C) = \frac{\max_j x_{ji}}{NCol_i(C)}, i = \{1, \cdots, m\}$$

C の部分集合も同じように生成できる。また、C のどんな部分集合に対して分割表を作っても、各行の対象数は常に同じであるから、ND_i で表す。

4　行ベクトルと列ベクトルの指標

デシジョンテーブルのある属性の集合の分類能力を推測するために、ラフセット理論と決定木からの代替的な指標を導入して、分割表でその値を計算する。列のベクトルから、条件属性で判別できるペア数が得ら

れる。指標 α と β それぞれは判別できるペア数の平均と正規化したペア数の平均である。行のベクトルから、C^* の各クラスの正解率が得られる。指標 γ と δ それぞれは Gini index と cross-entropy の式を応用する。各指標は以下の式で表す。

$$\alpha(C) = \sum_{1\le i<j\le n}\left(\frac{ND_i\cdot ND_j}{\sum_{1\le k<l\le n}(ND_k\cdot ND_l)}\right)\cdot\frac{(Row_i(C)^T\ Row_j(C))}{ND_i\cdot ND_j}$$

$$\beta(C) = \sum_{1\le i<j\le n}\left(\frac{ND_i\cdot ND_j}{\sum_{1\le k<l\le n}(ND_k\cdot ND_l)}\right)\cdot\frac{(Row_i(C)^T Row_j(C))}{|Row_i(C)|\cdot|Row_j(C)|}$$

$$\gamma(C) = \sum_{i=1}^{n}\left(\frac{NCol_i(C)}{|U|}\right)p_i(1-p_i)$$

$$\delta(C) = -\sum_{i=1}^{n}\left(\frac{NCol_i(C)}{|U|}\right)p_i\log p_i$$

得られた指標は単調の関数を使って、$[0,1]$ の間に基準化する。変換した指標を利用して、近似的なリダクトを定義する。例えば、$C''\subset C'$ に対して、変換した値が $\alpha(C')\ge 0.9$、かつ、$\alpha(C'')<0.9$ ならば、C' は C の $\alpha_{0.9}$-リダクトと呼ぶことにする。また、指標 γ と δ は、2 クラスの分類問題に限定する。

指標 I の単調性とは、$C^{(n)}\subset C^{(n-1)}\subset\cdots\subset C'\subset C$ ならば、$I(C^{(n)})\ge I(C^{(n-1)})\ge\cdots\ge I(C')\ge I(C)$。が成り立つことを言う。

補題 指標 α、γ、と δ は単調性を持つ。

証明. $C'\subset C$ ならば、$CT(C')$ の各列は、$CT(C)$ のいくつかの列の和である。2 クラスの分類問題にで、二つの列ベクトルの和から、指標関数の性質（凹関数と減少関数）を使って、この補題を証明できる。例えば、$CT(C)$ の列ベクトルは $(a,b)^T$、$p_1=\max(a,c)/(a+c)$ と $(c,d)^T$、$p_2=\max(b,d)/(b+d)$ で、$CT(C')$ の列ベクトルは $(a+c,b+d)^T$、$(a+b+c+d)$ は u と設定して、$p_3=\max(a+b,c+d)/u$ とすると、

$$\begin{aligned}
&\delta(C)\\
=\ & -\frac{a+c}{|U|}p_1\log p_1-\frac{b+d}{|U|}p_2\log p_2\\
=\ & \frac{u}{|U|}\frac{a+c}{u}(-p_1\log p_1)+\frac{u}{|U|}\frac{b+d}{u}(-p_2\log p_2)\\
\le\ & \frac{u}{|U|}\left(-\frac{\max(a,c)+\max(b,d)}{u}\log\frac{\max(a,c)+\max(b,d)}{u}\right)\\
\le\ & \frac{u}{|U|}\left(-\frac{\max(a+b,c+d)}{u}\log\frac{\max(a+b,c+d)}{u}\right)\\
=\ & -\frac{u}{|U|}p_3\log p_3=\delta(C')
\end{aligned}$$

以上の一番目の不等号は凹関数の性質、二番目の不等号は減少関数の性質から得られた結果である。

5 結論

本研究で導入した分割表は、各クラスの対象数を基に作られたものである。この分割表の列ベクトルと行ベクトルを使って、種々の指標を作れる。閾値を決め、作った指標と比較することより、近似的なリダクトを定義できる。

この分割表を使って、より高速で近似的なリダクトを求める。得られた近似的なリダクトは、ノイズが入ったデシジョンテーブルに対しても分類能力が強いものである。数値実験の結果については発表のときに紹介させていただく。

参考文献

[1] K. Cios, W. Pedrycz, R. Swiniarski, Rough sets, Data mining methods for knowledge discovery, Kluwer Academic Publishers, 27-72.

[2] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning, Springer, 270-272.

[3] M. Beynon, Reducts within the variable precision rough sets model: a further investigation, European journal of operational research, 134(2001), 592-605.

[4] S. Vinterbo, A. Ohrn, Minimal approximate hitting sets and rule templates, International Journal of approximate reasoning, 25(2000), 123-143.