知 能 と 複 雑 系 128−4
（２００２．５．２３）

ラフ集合とグラフィカルモデリングによるルール表示法の開発

津本周作
島根医科大学医学部医療情報学講座

概要: ルール生成法は一般に頻度情報からパターンを生成し，項目間の独立性を検討するため，項目間の独立性が高い場合，ルールの記述長が長くなり，専門家にとって，奇異な印象を与えるルールが生成されることはまれではない．本論文では，ルール生成において，グラフィカルモデリングの基本的な考え方である条件付独立の概念を導入する．ルールを生成する属性の集合はグラフィカルモデリングでは，完全グラフ (クリーク) に相当し，属性の 2 項関係をもとに，属性間の完全グラフを探索しなければならない．完全グラフ内の属性を利用することにより，ルールに必要な属性数が抑制できる上，属性間の従属度が明らかになる．本稿では，完全グラフ探索のために，逐次的な属性選択と統計学的な方法の二つを導入，それらの基本的な考え方を示した．

Visualization of Rules based on Rough Sets and Graphical Modeling

Shusaku Tsumoto
Department of Medical Informatics, Shimane Medical University

Abstract: Rule induction methods have been introduced since 1980's and many applications show that they are very useful to acquire simple patterns from large databases. However, when a database is very large, the methods generate too many rules, which makes domain experts interpret all the rules. Moreover, since rules only shows the relations between attribute-value pairs, it is very difficult to capture the relations between concepts or among induced rules. In order to solve this problem, many kinds of visualization has been introduced. Rough set theory has a technique on conflict analysis with qualitative distance obtained from attributes, which gives graphical relations between class or rules. On the other hand, statistical methods have a graphical model method, which gives graphical relations between attributes by using partial coefficients or other indices. In this paper, we introduce a new approach which combines conflict analysis and graphical modeling. The results show that the combination of these two methods gives the other type of visualization of rules, which gives also a formal mathematical model for rule visualization.

1 はじめに

ルール生成法は一般に頻度情報から計算された条件つき確率を指標として，その指標がある一定以上であれば，項目間に相関があるとみなして，パターンを生成する [5].

例えば，表 1 で示された分割表では，accuracy は 100/120(=0.833)，coverage は 100/140(=0.714) であり，閾値の選択にもよるが，$[a = 0] \rightarrow [b = 0]$ なるルールが生成される．一方，この分割表では，χ^2 検定量は 0 となり，二つの項目には統計的に独立である．分割表を見れば，列においても行においても，それぞれの列間と行間に比例的関係があることが見いだせる．しかも，

$$p([a=0])p([b=0]) = \frac{120}{168} \times \frac{140}{168} = \frac{100}{168} = p([a=0],[b=0])$$

となり，この二つは確率からは独立である．これらの結果から，必ずしも，accuracy と coverage はそれぞれ頻度が高いという事実だけで，それぞれの項目の関連性を示すものでないことに注意しなければならず，項目の関連性を探索するためには，かならずその独立性を調べなければならない．

Table 1: 独立性の高い分割表

	a=0	a=1	
b=0	100	40	140
b=1	20	8	28
	120	48	168

以上をまとめれば，たとえ accuracy と coverage が高くても，独立性の検定で棄却されれば，そのような項目はルールとして，選択してはならない．た

だし，次の節で示すように，accuracy が 1.0 に近ければ，独立性が成り立つ可能性は極めて低い．本論文では，ルール生成において，グラフィカルモデリングの基本的な考え方である条件付独立の概念を導入する．ルールを生成する属性の集合はグラフィカルモデリングでは，完全グラフ (クリーク) に相当し，属性の 2 項関係をもとに，属性間の完全グラフを探索しなければならない．完全グラフ内の属性を利用することにより，ルールに必要な属性数が抑制できる上，属性間の従属度が明らかになる．本稿では，完全グラフ探索のために，逐次的な属性選択と統計学的な方法の二つを導入，それらの基本的な考え方を示した．

2 確率から見た事象の独立性

付録に示したように，ある属性空間 A で記述された表形式のデータ集合 (U, A) を考える．R を属性の組み合わせからなる記述子，R_A を R をみたす例の集合，D をあるクラスに属する例の集合とする．この時，accuracy と coverage は以下のように定義できる．

$$\begin{aligned}\alpha_R(D) &= \frac{|R_A \cap D|}{|R_A|}(= P(D|R)), \text{ および}\\ \kappa_R(D) &= \frac{|R_A \cap D|}{|D|}(= P(R|D)),\end{aligned}$$

ここで，P(R) と P(D) とを

$$P(R) = \frac{|R_A|}{|U|} \quad \text{および} \quad P(D) = \frac{|D|}{|U|}$$

と定義すれば，条件付き確率の性質から，

$$\begin{aligned}\alpha_R(D) &= \frac{|R_A \cap D|}{|R_A|}\\ &= \frac{P(R \cap D)}{P(R)} = \frac{P(R, D)}{P(R)}\\ &= \frac{P(D)}{P(R)}\kappa_R(D),\end{aligned}$$

であることが示せる．これらの定義を使えば，独立性に関する指標 $\varsigma_R(D)$ が次のように定義できる:

$$\varsigma_R(D) = \frac{|R_A \cap D|}{|R_A||D|} = \frac{P(R, D)}{P(R)P(D)},$$

この指標から，独立性について，

$$\varsigma_R(D) = \begin{cases} > 1.0 & \textit{Dependent}\\ = 1.0 & \textit{Statistical Independent}\\ < 1.0 & \textit{Independent}\end{cases}$$

が言え，$\varsigma_R(D)$ と accuracy, coverage との間には:

$$\varsigma_R(D) = \frac{\alpha_R(D)}{P(D)} = \frac{\kappa_R(D)}{P(R)}$$

が成り立つ．

ここで，$\alpha_R(D) = 1.0$ なる式 R を用いれば，

$$\varsigma_R(D) = 1/P(D) > 1.0$$

となり，$P(D)$ が低い疾患であればあるほど，R と D の従属性は高くなる．また，$\alpha_R(D) > P(D)$ となる accuracy をもつ式 R を採用すれば，$\varsigma_R(D) > 1.0$ となる．ただし，上記の指標はノイズの影響を簡単に受けてしまうので，頑健性は低いので，ノイズが多い実ドメインでは，次節で説明する独立性の検定が必要になる．

3 独立性の検定

3.1 分割表

ここでは，議論を簡単にするため，2×2 分割表に限定して議論する．式 R_1 and R_2 を $F(B, V)$ 上の式で，それぞれ 2 値であると仮定する．分割表は，次の式を満たす例の数で構成され，表 2 の形にまとめられる: $|[R_1 = 0]_A|$,$|[R_1 = 1]_A|$, $|[R_2 = 0]_A|$,$|[R_2 = 1]_A|$, $|[R_1 = 0 \wedge R_2 = 0]_A|$,$|[R_1 = 0 \wedge R_2 = 1]_A|$, $|[R_1 = 1 \wedge R_2 = 0]_A|$,$|[R_1 = 1 \wedge R_2 = 1]_A|$, $|[R_1 = 0 \vee R_1 = 1]_A|(= |U|)$. ここで，$a_{1\cdot}$, $a_{\cdot 1}$ はそれぞれ 1 行目と 1 列目の集計を表し，$a_{\cdot\cdot}$ は全体の集計，つまり標本数を示す．

Table 2: Two way Contingency Table

	$R_1 = 0$	$R_1 = 1$	
$R_2 = 0$	a_{11}	a_{12}	$a_{1\cdot}$
$R_2 = 1$	a_{21}	a_{22}	$a_{2\cdot}$
	$a_{\cdot 1}$	$a_{\cdot 2}$	$a_{\cdot\cdot}$ $(= \|U\| = N)$

この表から，$[R_1 = 0] \rightarrow [R_2 = 0]$ に関する accuracy and coverage は次のように定義できる:

$$\begin{aligned}\alpha_{[R_1=0]}([R_2 = 0]) &= \frac{|[R_1=0 \wedge R_2=0]_A|}{|[R_1=0]_A|} = \frac{a_{11}}{a_{\cdot 1}},\\ \kappa_{[R_1=0]}([R_2 = 0]) &= \frac{|[R_1=0 \wedge R_2=0]_A|}{|[R_2=0]_A|} = \frac{a_{11}}{a_{1\cdot}}.\end{aligned}$$

3.2 χ^2 検定

χ^2 検定は，独立性検定で最も良く使われる検定であり，表 2 の要素から得られる式:

$$\chi^2 = \sum_{i,j=1}^{2,2} \frac{(x_{ij} - a_{i\cdot}a_{\cdot j}/N)^2}{a_{i\cdot}a_{\cdot j}/a_{\cdot\cdot}}$$

が漸近的に，$(n-1)(m-1)$ の自由度をもつ χ^2 分布にしたがうことが利用される．χ^2 検定統計量は式を見てわかるように，各セルと観察された度数と

の残差を評価しようとするものである．特に，2×2 分割表の場合，簡単な計算で，

$$\chi^2 = \frac{a_{..}(a_{11}a_{22} - a_{12}a_{21})^2}{a_{1\cdot}a_{2\cdot}a_{\cdot 1}a_{\cdot 2}}$$

であることが示され，分子に分割表を行列としてみなした時の行列式の値が示されている．つまり，χ^2 統計量はベクトルの独立性に相当している．一般に分割表は正方性を保つものではないが，正方行列に相当する分割表では，どのような関係を導き出すことができる．

この検定量の問題点は，

(1) 標本数が 20 以下であると，χ^2 分布への近似が悪くなる．
(2) セルに 0 が多くなると，χ^2 分布への近似が悪くなる．
(3) 自由度が大きくなると，統計量の値が大きくなる．

などの問題があり，これらの場合に注意が必要である．

3.3 クラメールの連関係数

χ^2 検定の三番目の問題点に対処するために，分割表の行数を m，列数を n とした時に，

$$V = \sqrt{\frac{\chi^2}{a_{..}\{\min(m,n) - 1\}}}$$

で定義される指標を用いる場合がある．この指標をクラメールの連関係数と呼ぶ．

3.4 Fisher's Exact Test

χ^2 検定量の一番めと二番目の問題に対しては，Fisher の正確検定を適用するのが良い [2, 6]．2×2 の分割表の場合，

$$p(a_{11}, a_{12}, a_{21}, a_{22} | a_{1\cdot}, a_{2\cdot}, a_{\cdot 1}, a_{\cdot 2}) = \frac{{}_{a_{\cdot 1}}C_{a_{11}} \times {}_{a_{\cdot 2}}C_{a_{12}}}{{}_{a_{..}}C_{a_{1\cdot}}}$$

が超幾何分布をなし，この値は，ある境界分布が得られた時，各セルの値が与えられる分割表が観察される確率を表す．したがって，この値自身が χ^2 統計量から得られる独立性の検定の p 値に相当することになる．例えば，表 1 の例では，

$$\begin{aligned} & p(a_{11}, a_{12}, a_{21}, a_{22} | a_{1\cdot}, a_{2\cdot}, a_{\cdot 1}, a_{\cdot 2}) \\ &= \frac{{}_{120}C_{100} \times {}_{48}C_{40}}{{}_{168}C_{140}} \\ &= 0.181 \end{aligned}$$

4 条件付き独立性とルール生成

4.1 条件付き独立とは

例えば，表 3 を考えてみる．この表から，χ^2 統計量は 4.1667，自由度 1 での p 値は 0.0412 で，危険率が 0.05 であれば，この分割表は従属性の高いものとみなせる．(クラメールの連関係数は 0.167) ここ

Table 3: 従属性の高い分割表

	a=0	a=1	
b=0	100	20	120
b=1	20	10	30
	120	30	150

で，このデータに c なる因子があったとして，c で層別することで，次の二つの分割表に分解できたと仮定する．

Table 4: 分解後の分割表 (c=0)

	a=0	a=1	
b=0	30	15	45
b=1	15	9	24
	45	24	69

Table 5: 従属性の低い分割表 (c=1)

	a=0	a=1	
b=0	70	5	75
b=1	5	1	6
	75	6	81

表 4 では，検定統計量は 0.12，p 値は 0.729 (クラメールの連関係数は 0.028)，表 5 では，検定統計量は 0.81，p 値は 0.361 (クラメールの連関係数は 0.073) となり，いずれも独立であることが否定できない．以上から，次のことがわかる:

(1) c と a，c と b とは相関がある．
(2) a と b との関係は c を固定すると独立．
(3) c を背景因子とすることで，a と b が相関があるように見える

これを Simpson パラドックス [4] と呼び，背景の交絡因子によって，見かけ上の相関関係が得られることになる．以上のことから，a と b とは c に関して条件付き独立といい，確率からは

$$p(a, b|c) = p(a|c)p(b|c)$$

と表せる [1].

4.2 Accuracy と条件付き独立

上記の式を今までの記号で書き直せば，

$$p(R_1, R_2|D) = p(R_1 \wedge R_2|D) = p(R_1|D)p(R_2|D)$$

と表される．これは coverage に対する関係と見なせるが，ここで，$p(D|R_1, R_2)$ と $p(D|R_1)$ との関係を調べてみよう．

$$\begin{aligned}
&p(D|R_1, R_2) - p(D|R_1) \\
&= \frac{p(R_1, R_2, D)}{p(R_1, R_2)} - \frac{p(R_1, D)}{P(R_1)} \\
&= \frac{p(R_1, R_2|D)p(D)}{p(R_1, R_2)} - \frac{p(R_1|D)p(D)}{p(R_1)} \\
&= \frac{p(R_1, R_2|D)p(D)}{p(R_2|R_1)P(R_1)} - \frac{p(R_1|D)p(D)}{p(R_1)} \\
&= \frac{p(D)}{p(R_1)} \times \left(\frac{p(R_1, R_2|D)}{p(R_2|R_1)} - p(R_1|D) \right)
\end{aligned}$$

ここで，R_1 と R_2 とが D について条件つき独立でなく，$P(R_1, R_2|D) > P(R_1|D)P(R_2|D)$ を満たしていれば，

$$\begin{aligned}
&p(D|R_1, R_2) - p(D|R_1) \\
&= \frac{p(D)}{p(R_1)} \times \left(\frac{p(R_1, R_2|D)}{p(R_2|R_1)} - p(R_1|D) \right) \\
&> \frac{p(D)}{p(R_1)} \times \left(\frac{p(R_1|D)p(R_2|D)}{p(R_2|R_1)} - p(R_1|D) \right) \\
&= \frac{P(D)p(R_1|D)}{p(R_1)} \left(\frac{p(R_2|D)}{p(R_2|R_1)} - 1 \right) \\
&= P(D|R_1) \left(\frac{p(R_2|D)}{p(R_2|R_1)} - 1 \right)
\end{aligned}$$

上記の式は，$P(R_1, R_2|D) > P(R_1|D)P(R_2|D)$ を満たしている R_1 と R_2 を join させる方が，条件つき独立である属性を join させるように，良い accuracy が得られることを示している．また，従属している式のうちで，$p(R_2|D) >> p(R_2|R_1)$ である関係から順に選ぶべきであることを示唆している．

4.3 ルール生成

上記の議論から，条件付き独立でない属性の集合を考えることで，accuracy が高く，記述長の短いルールが生成されることが示された．このような属性の集合はグラフィカルモデリングでは，完全グラフに対応し，クリーク (clique) と呼ばれる [1]．つまり，表形式のデータを記述する属性全体の集合から，Target Concept に関わるクリークを探索することで，ルールを記述する属性の集合が求められる．

5 例

グラフィカルモデリングによるルールの生成を表 6 に適用してみる．ここでは，例数が少ないことから，独立性の検定として，Fisher's Exact Test を用いる．

5.1 分割表の作成

まず，class と各属性との独立性を調べるために，分割表を作成すれば，表 7 から 12 のようになる．

Table 7: Contingency Table for *age*

	40-49	50-59	
m.c.h.	2	1	3
migra	2	0	2
psycho	0	1	1
	4	2	6

Table 8: Contingency Table for *location*

	occular	whole	lateral	
m.c.h.	1	2	0	3
migra	0	1	1	2
psycho	0	1	0	1
	1	4	1	6

Table 9: Contingency Table for *nature*

	persistent	throbbing	radiating	
m.c.h.	2	0	1	3
migra	0	2	0	2
psycho	1	0	0	1
	3	2	1	6

5.2 Fisher's Exact Test(1)

これらの表から，Fisher's Exact Test は次のようになる．

$$\begin{aligned}
&p(age|4, 2, 3, 2, 1) \\
&= \frac{{}_4C_2 \times {}_2C_1 \times {}_4C_2 \times {}_2C_0}{{}_6C_3 \times {}_6C_2} \\
&= \frac{6 \times 2 \times 6 \times 1}{20 \times 15} = \frac{72}{300} = 0.24
\end{aligned}$$

Table 6: データ集合の例

No.	age	location	nature	prodrome	nausea	M1	class
1	50-59	occular	persistent	no	no	yes	m.c.h.
2	40-49	whole	persistent	no	no	yes	m.c.h.
3	40-49	lateral	throbbing	no	yes	no	migra
4	40-49	whole	throbbing	yes	yes	no	migra
5	40-49	whole	radiating	no	no	yes	m.c.h.
6	50-59	whole	persistent	no	yes	yes	psycho

DEFINITIONS. M1: tenderness of M1, m.c.h.: muscle contraction headache, migra: migraine, psycho: psychological pain.

Table 10: Contingency Table for *prodrome*

	yes	no	
m.c.h.	0	3	3
migra	1	1	2
psycho	0	1	1
	1	5	6

Table 11: Contingency Table for *nausea*

	yes	no	
m.c.h.	0	3	3
migra	2	0	2
psycho	1	0	1
	3	3	6

Table 12: Contingency Table for $M1$

	M1=yes	M1=no	
m.c.h.	3	0	3
migra	0	2	2
psycho	1	0	1
	4	2	6

$$p(location|1,4,1,3,2,1) = \frac{{}_1C_1 \times {}_4C_2 \times {}_1C_1 \times {}_1C_1 \times {}_4C_1 \times {}_1C_1}{{}_6C_3 \times {}_6C_2} = \frac{1 \times 6 \times 1 \times 1 \times 4 \times 1}{20 \times 15} = \frac{24}{300} = 0.08$$

$$p(nature|3,2,1,3,2,1) = \frac{{}_3C_2 \times {}_2C_0 \times {}_1C_0 \times {}_3C_0 \times {}_2C_2 \times {}_1C_0}{{}_6C_3 \times {}_6C_2} = \frac{1 \times 3 \times 1 \times 1 \times 1 \times 1}{20 \times 15} = \frac{1}{300} = 0.0033$$

$$p(prodrome|1,5,3,2,1) = \frac{{}_1C_0 \times {}_5C_3 \times {}_1C_1 \times {}_5C_1}{{}_6C_3 \times {}_6C_2} = \frac{1 \times 10 \times 1 \times 5}{20 \times 15} = \frac{50}{300} = 0.16667$$

$$p(nausea|3,3,3,2,1) = \frac{{}_3C_0 \times {}_3C_3 \times {}_3C_2 \times {}_3C_0}{{}_6C_3 \times {}_6C_2} = \frac{1 \times 1 \times 3 \times 1}{20 \times 15} = \frac{3}{300} = 0.001$$

$$p(M1|4,2,3,2,1) = \frac{{}_4C_3 \times {}_2C_0 \times {}_4C_0 \times {}_2C_2}{{}_6C_3 \times {}_6C_2} = \frac{1 \times 4 \times 1 \times 1}{20 \times 15} = \frac{4}{300} = 0.00133$$

これらから，棄却率を 0.05 とすれば，*nature*, *nausea*, $M1$ が選択される．

5.3 Fisher's Exact Test(2)

次に，*nature*, *nausea*, $M1$ の間で，*class* の下での条件付き独立を調べる．表 13 から 15 より，それぞれの p 値はほぼ 1.0 で，各項目は条件付き独立を満たしている．したがって，これら 3 つの属性は，class に関して条件付き独立である．以上から，完全グラフは，このデータ集合では存在せず，*nature*, *nausea*, $M1$ のそれぞれ一つの属性で *class* に関するルールを生成することで十分である．

6 おわりに

ルールとグラフィカルモデリング (条件付き独立のグラフ表示の方法) を組み合わせる方法について概説した．ここで紹介したグラフィカルモデリングの方法はまだナイーブなものであり，今後，対数線形モデルや共分散選択を含めたグラフィカルモデリングとルール生成との組み合わせを検討していく予定である．

Table 13: Class, nature, nausea に関する分割表

	m.c.h.		migra		psycho	
	nausea=yes	nausea=no	nausea=yes	nausea=no	nausea=yes	nausea=no
persistent	0	2	0	0	1	0
nature throbbing	0	0	2	0	0	0
radiating	0	1	0	0	0	0

Table 14: Class, nature, M1 に関する分割表

	m.c.h.		migra		psycho	
	M1=yes	M1=no	M1=yes	M1=no	M1=yes	M1=no
persistent	2	0	0	0	1	0
nature throbbing	0	0	0	2	0	0
radiating	1	0	0	0	0	0

Table 15: Class, nausea, M1 に関する分割表

	m.c.h.		migra		psycho	
	M1=yes	M1=no	M1=yes	M1=no	M1=yes	M1=no
nausea yes	0	0	0	2	1	0
no	3	0	0	0	0	0

謝辞

本研究の一部は、文部科学省科学研究費補助金 (特定領域研究 (B)(No.759))、「情報洪水時代におけるアクティブマイニングの実現」の助成による。

References

[1] Edwards, D. *Introduction to Graphical Modelling* (2nd Ed.), Springer, New York, 2000.

[2] Fisher, R.A. *Statistical Methods for Research Workers* (5th Ed.), Oliver&Boyd, Edinburgh, 1934.

[3] Skowron, A. and Grzymala-Busse, J. From rough set theory to evidence theory. In: Yager, R., Fedrizzi, M. and Kacprzyk, J.(eds.) *Advances in the Dempster-Shafer Theory of Evidence*, pp.193-236, John Wiley & Sons, New York, 1994.

[4] Simpson, C.H. The interpretation of interaction in contingency tables. *Journal of Royal Statistical Society Series B*, **13**, 238-241, 1951.

[5] Tsumoto, S. Automated Discovery of Positive and Negative Knowledge in Clinical Databases based on Rough Set Model., *IEEE EMB Magazine*, 56-62, 2000.

[6] Tsumoto, S. Statistical Extension of Rough Set Rule Induction *Proceedings of SPIE: Data Mining and Knowledge Discovery: Theory, Tools, and Technology III*, 2001.

A　データ集合に関する記法

データ全体を示す有限集合 U を考え，U を観察して得られた属性の集合を A とする．この時，ある属性 $a \in A$ は，V_a を a の値域とすれば，$a : U \rightarrow V_a$ を満たす写像とみなせる．決定表は，決定属性を d とすれば，$A = (U, A \cup \{d\})$ で表現できる [3].

ある原始式は $B \subseteq A \cup \{d\}$ and V の上で，$a \in B$ および $v \in V_a$ を用いて，$[a = v]$ として表現でき，B における記述子と呼ぶ．

B 上の式全体の集合 $F(B, V)$ は B 上の記述子をすべて含み，選言，連言，否定に関して閉じた最小の集合として定義できる．各式 $f \in F(B, V)$ において，f_A は f in A の**意味**と定義され，f という性質をもつ U 上のすべての例を含む集合で，帰納的に次のように定義できる:

(1) もし f が $[a = v]$ という形なら，$f_A = \{s \in U | a(s) = v\}$ と表される

(2) $(f \wedge g)_A = f_A \cap g_A$; $(f \vee g)_A = f_A \vee g_A$; $(\neg f)_A = U - f_a$ を満たす．

例えば，表 6 であれば，$[age = 40-49]$ は記述子であり，属性全体の集合を A とすれば，$[age = 40-49]$ の意味は $[age = 40 - 49]_A = \{2, 3, 4, 5\}$ で与えられ，$[age = 40 - 49] \wedge [location = whole]$ の意味は $\{2, 4, 5\}$ で与えられる．