

カテゴリーへの帰属がグレードで示された 離散データからの知識獲得 ——ラフ集合の縮約計算を拡張する提案

森 典彦
大学院芸術学研究科

Knowledge Acquisition from the Data Consisting of Categories Attended
with Degrees of Conformity
——Proposal of Extended Reduct Calculation in Rough Set Theory

MORI Norihiko
The Postgraduate Media-Art Course
(Received October 12, 2000; Accepted January 19, 2001)

1. 背景と目的

因果関係に関する複雑な離散データがあったとき、その中から意味のある因果関係をルールとして見出そうというデータマイニングや、そのルールを有用な知識として明確化しようという知識獲得の手段として、ラフ集合理論が近年注目され、医療関係その他で実施例が増えつつある^{1,2,3)}。

対象の属性と、対象に接して人が感ずる感性の関係は、前者を原因とし後者を結果とする因果関係であるが、それを離散データとして表わした場合、感性なるがゆえにデータはあいまいかつ複雑になるのが常である。したがって人の感性にからむデータから、何らかの有用な知識を獲得するためにはラフ集合理論の適用がふさわしいのではないかと考えに立つて事例研究がなされるようになった。製品の形態要素と人が感ずるイメージという因果関係を離散データで表わしたものはその代表的なものの一つであり、ラフ集合理論に着目し、その中心的概念である縮約を計算で求めることによってデザインを支援するための有用な知識が得られることを示した事例研究が報告されている^{4,5,6)}。

ラフ集合理論を適用するためにはデータは離散データでなければならない。しかし、感性に関わるデータの場合は原因のデータも結果のデータもあいまいで、一つの観測値が一つのカテゴリーに確実に当てはまるのではなく、あるグレードをもってしか当てはまるといえないという場合が多い。このことから考え、グレードをもつ離

散データにも適用できるようにラフ集合理論を拡張することは今後の自然な要求であろうと思われる。

本報告は最初の試みとして、原因の方のみがグレードをもつ離散データであって結果の方は普通の離散データであるとして、ラフ集合理論の縮約を導出する方法を試案として示し、それによってこの方法が知識獲得のために有用であることを明らかにすることが目的である。また獲得しようとする知識はできるだけ整理された実用的な知識であるように縮約を導出するものとする。

2. 概念、定義、データ

グレードは確率ではない。グレードはある観測値があるカテゴリーに当てはまる度合いを感覚的に言い表したものであるから、ファジイ集合における台集合のある点のメンバシップ値と同義である。したがってグレードを p で表わすとき

$$p = [0, 1]$$

である。またカテゴリー A におけるグレード p は、ファジイ集合の表現では p/A であるが簡単のため単に

$$pA$$

で表わす。そしてグレード p とグレード q の差異 $pA \sim qA$ を

$$pA \sim qA = (p \sim q)A$$

と定義する。差異の度合いを表わしている。

獲得しようとする知識ができるだけ整理された表現であるようにするために、知識の前件部すなわち原因の表現は、あるカテゴリーにおける一つのグレードではなく、

区間をもつグレードで表わすものとする。例えば「0.8A
であれば度合い0.6で」あるカテゴリーが結果されるという知識が得られたとき、その知識は「0.8A 以上 (0.8以上の A) であれば少なくとも度合い0.6で」あるカテゴリーが結果されるという知識が帰結されたものとする。その方が有用な表現である。ここで「0.8以上」が区間をもつグレードである。

データの例として注記欄 4) の文献の36ページ図 2 (表 1 として再録、ただし一部記号変更) を取り上げ、加工することにする。表 1 を表 3 によってグレードつきに加工した表 2 が本報告におけるデータである。

以下縮約導出までの論述においては表 1 からの縮約導出と表 2 からの縮約導出を併記し、離散データの場合とグレードつき離散データの場合とを比較しながら後者の性格をはっきりさせていく。

表 2 において A、C、E、G、J は原因としてのカテゴリーであり、U は対象を表わし、Y は結果としてのカテゴリーである。具体的な世界への投影としては、例えば U はクルマとして

表 1 離散データ

	アイテム					Y
	1	2	3	4	5	
U1	A	C	E	G	J	1
U2	B	C	E	H	J	2
U3	A	C	F	H	K	2
U4	A	D	F	G	K	2

表 2 グレードつき離散データ

	カテゴリー					Y
	A	C	E	G	J	
U1	0.7A	0.8C	0.6E	0.9G	1.0J	1
U2	0.3A	0.8C	0.7E	0.2G	1.0J	2
U3	0.6A	0.8C	0.2E	0.2G	0.0J	2
U4	0.7A	0.4C	0.2E	0.7G	0.0J	2

表 3 加工のための対比表

表 1	→	表 2
A	→	[0.5, 1] A
B	→	[0, 0.4] A
C	→	[0.5, 1] C
D	→	[0, 0.4] C
E	→	[0.5, 1] E
F	→	[0, 0.4] E
G	→	[0.5, 1] G
H	→	[0, 0.4] G
J	→	[0.5, 1] J
K	→	[0, 0.4] J

A：個性的な形のヘッドランプ

C：台形度の強いシルエット

E：大きくて目立つタイヤ

G：低い車高

J：派手なボデーカラー

Y=1：スポーティ

Y=2：スポーティでない

である。

逆に現実世界の観測データからグレードつき離散データへの変換のしかたについて触れる。

原因としての対象の属性が例えば「全体の形の丸さ」であって

非常に丸い／やや丸い／中くらい／やや角張っている／非常に角張っている

の 5 個のカテゴリーに分けられるとし、それぞれに

0.9／0.7／0.5／0.3／0.1

のグレードが割り当てられるとすれば、「全体の形の丸さ」はグレードを間隔尺度とみなした 1 つのグレードつきカテゴリーで表わせる。

しかし属性が「色の鮮やかさ」であって

鮮やかな色／渋い色

の 2 つのカテゴリーがあるとしたとき、1 つのグレードつきカテゴリーで表わすことはできないかもしれない。深みのあるブルーを見たとき、[鮮やかな色]のグレード 0.9、[渋い色]のグレード 0.5 というように、グレードを足しても 1 にならない 2 つのカテゴリーで表わしたい場合があるからである。

3. 識別行列

つぎにデータから識別行列を結果としての Y ごとに作る。離散データの場合の、ラフ集合理論における Shan/Ziarko のアルゴリズム⁷⁾に使われる識別行列に相当するものである。

まず離散データ表 1 の識別行列を表 4 として示しておく⁸⁾。

表 4 離散データの識別行列

Y=1				Y=2	
U	2	3	4	U	1
1	A	E	C	2	B
	G	G	E		H
		J	J	3	F
					H
				4	K
					D
					F
					K

表5 グレードつき識別行列

Y=1				Y=2	
U	2	3	4	U	1
1	0.4A	0.1A	0.4C	2	0.4E [~]
	0.1E [~]	0.4E	0.4E		0.1E
	0.7G	0.7G	0.2G		0.7G [~]
		1.0J	1.0J	3	0.1A [~]
					0.4E [~]
					0.7G [~]
				4	1.0J [~]
					0.4C [~]
					0.4E [~]
					0.2G [~]
					1.0J [~]

グレードつき離散データにおける識別行列の作り方を述べる。Y=1の識別行列は、Y=1に該当するUを行に、該当しないUを列に置き、行のUと列のUをデータで比較したとき、たとえばカテゴリーAに差異があったとき、行のUにおけるカテゴリーAのグレード値から列のUにおけるカテゴリーAのグレード値を引き、正負を区別したものを行列の要素として記入する。それは、正の値ならば行のカテゴリーAのグレード値が「そのグレード値以上」を意味し、負の値ならば「そのグレード値以下」を意味するとみなすことによって縮約を区間で求めるという目的を達するためである。

引き算が正ならばそのカテゴリーたとえばAならばAとし、負ならばA[~]と書くことにする。

Y=2も同様である。

データ表2に対する識別行列は表5のように得られる。その要素はグレードの差分であり、差異の度合いを表わす。

グレードつきの場合は離散データの場合と違って識別行列の要素が行のUのグレードつきカテゴリーの上に立つ差分であることに注意する。たとえばY=1における1行1列の0.4Aというのは、U1のAが0.7A 以上 (0.7Aを含む)であればU2のAに対し少なくとも0.4の差分をもつこと、また2行1列の0.1E[~]というのはU1のEが0.6E 以下 (0.6Eを含む)であればU2のEに対し少なくとも0.1の差分をもつことを意味する。

表3の対比表においてすべての[0.5, 1]を1に、すべての[0, 0.4]を0にとれば表5は表4に帰着する。したがってグレードつき識別行列はShan/Ziarkoの識別行列を特別な場合として含んでおり、拡張したものといえる。

4. 縮約の導出と知識表現

ラフ集合理論における極小縮約とは与えられたデータ

の中で目的とする結果が他から識別されるための、原因としての属性に関する極小の十分条件である⁹⁾。設計支援などを目的とする知識獲得においては、データを最大限に活用したいので知識は少なくともそのデータの範囲内では確実なもの、つまり必然性あるものとしたい。このことから知識獲得の方法としては、可能性を与える縮約ではなく、必然性を与える縮約すなわち極小縮約を求めることにする。

極小縮約は識別行列において、各列につき行間をor結合したものを列間でand結合し、ブール演算することによって求められる。以下極小縮約を単に縮約という。

ブール演算則は、or結合を+、and結合を*で表わせば

$$A * A = A, A + A = A, A + AB = A(1 + B) = A$$

である。

まず離散データの識別行列表4のY=1にブール演算を実行すると

$$(A+G)(E+G+J)(C+E+J) \\ = AE + AJ + GC + GE + GJ \dots\dots\dots (1)$$

となってAE、AJ、GC、GE、GJの5個の縮約が得られる。

つぎにグレードつき離散データにおける識別行列の要素に対するブール演算則を考える。

識別行列の要素、すなわちグレードつきカテゴリーの差異の度合いに関する2項演算は論理演算であり、+は∨(大きい方をとる)、*は∧(小さい方をとる)であることに注意すると、同じカテゴリーAの演算は差異の度合いをsおよびtとして

$$\left. \begin{array}{l} sA + tA = (s \vee t)A \\ sA * tA = (s \wedge t)A \end{array} \right\} \dots\dots\dots (2)$$

また $sAA = sA$

ただし AA^{\sim} はそのままであってAではない。

さらに異なるカテゴリーの演算は

$s \geq t$ のとき

$$sA + tAB = sA \dots\dots\dots (3)$$

なぜなら $s \geq t$ のとき $sA + tA = (s \vee t)A = sA$ だから離散データのブール演算の $A + AB = A$ を拡張すると

$$sA + tAB = sA + tA + tAB = sA + tA = sA \\ (s < t \text{ のとき不成立})$$

4-1. 目的に該当する対象が1個の場合の縮約導出

表2よりY=1に該当するUはU1のみである。グレー

ドつき識別行列表5のY=1に(2)(3)を適用する。

$$\begin{aligned} & (0.4A+0.1E^{\sim}+0.7G)(0.1A+0.4E+0.7G+1.0J) \\ & \cdot (0.4C+0.4E+0.2G+1.0J) \\ & = 0.1AC+0.4GC+0.4AE+0.4EG+0.2G \\ & + 0.4AJ+0.7GJ+0.1E^{\sim}J+0.1EE^{\sim} \cdots (4) \end{aligned}$$

(4)は求める解、すなわちY=1を識別する9個の縮約である(導出の計算過程は末尾参考欄に示す)。各項の数値は識別する度合いを示している。因みに0.4以上に限ると(1)の5個の縮約と同じになっている。

グレードつきの場合は離散データの場合と違って各縮約のカテゴリーは、該当するもとのUのグレードつきカテゴリーを指定していることに注意する。この例ではY=1はU1だけであるからデータ表2のU1におけるグレードを指定している。たとえば(4)の第1項の0.1ACのACは0.7A * 0.8Cを指定している。このことを表示するため、各縮約にもとのUのグレードを次のように書き加えると便利である。

$$\begin{aligned} & 0.1/0.7A*0.8C+0.4/0.9G*0.8C \\ & + 0.4/0.7A*0.6E+0.4/0.6E*0.9G \\ & + 0.2/0.9G+0.4/0.7A*1.0J+0.7/0.9G*1.0J \\ & + 0.1/0.6E^{\sim}*1.0J+0.1/0.6E*0.6E^{\sim} \cdots (5) \end{aligned}$$

4-2. 目的に該当する対象が1個の場合の知識表現

(5)の知識としての表現は次のようになる。

「0.7A以上かつ0.8C以上であれば少なくとも0.1の度合いで、あるいは0.9G以上かつ0.8C以上であれば少なくとも0.4の度合いで、……、あるいは0.6E以下かつ1.0J(以上)であれば0.1の度合いで、あるいは0.6E以上かつ0.6E以下すなわち0.6Eであれば少なくとも0.1の度合いで、いずれもY=1を他から識別する」

この表現を2章で述べた現実世界への投影の例を使って表現すると次のようになる。度合いの大きいものからいくつかをピックアップする。

「(第7項より)非常に低い車高で非常に派手なボデーカラーでさえあれば度合い0.7で、あるいは(第2項より)非常に低い車高で台形度の強いシルエットでさえあれば度合い0.4で、あるいは(第3項より)かなり個性的なヘッドランプをもちいくらかでも大きく目立つタイヤでさえあれば度合い0.4で、スポーティなクルマが他から完全に識別される(特定される)」

4-3. 目的に該当する対象が複数個の場合の縮約導出

表2よりY=2に該当するUはU2、U3、U4である。グレードつき識別行列表5のY=2に(2)(3)を適用する。Uが複数のとき演算は各U毎に行い、Uの間をまたがってはいけないことに注意する。差異をもたらす元の

Uが違うのだから差分どうし演算しても意味がないからである。(離散データでは差分イコール元のカテゴリーで差分が同じなら元のUにかかわらず解釈できたがグレードつきの場合は区別しなければならない)

Y=2の縮約は

$$\left. \begin{array}{l} \text{U2より } 0.4A^{\sim}+0.1E+0.7G^{\sim} \\ \text{U3より } 0.1A^{\sim}+0.4E^{\sim}+0.7G^{\sim}+1.0J^{\sim} \\ \text{U4より } 0.4C^{\sim}+0.4E^{\sim}+0.2G^{\sim}+1.0J^{\sim} \end{array} \right\} \cdots (6)$$

という11個の縮約の和集合である。(6)はこれ以上単純化できない。

(5)のように表2に戻って(6)にもとのUのグレードを書き加えて次の11個の縮約を得る。

$$\begin{aligned} & 0.4/0.3A^{\sim}+0.1/0.7E+0.7/0.2G^{\sim}+0.1/0.6A^{\sim} \\ & + 0.4/0.2E^{\sim}+0.7/0.2G^{\sim}+1.0/0.0J^{\sim} \\ & + 0.4/0.4C^{\sim}+0.4/0.2E^{\sim}+0.2/0.7G^{\sim} \\ & + 1.0/0.0J^{\sim} \cdots (7) \end{aligned}$$

4-4. 目的に該当する対象が複数個の場合の知識表現

目的に該当する対象が複数個の場合に同じカテゴリーの縮約はグレードと度合いによっては統合して表現できることがあることに注意する。

たとえば $p \geq q$ として

$$s/pA+t/qA \cdots (8)$$

(8)は「pA以上ならば度合いsで、あるいはqA以上ならば度合いtで」を表している。

ここでさらに $p \geq q$ かつ $s \leq t$ ならば

$s/pA+t/qA=t/qA$ と統合できて「qA以上ならば度合いtで」と表現を単純化することができる。しかし $p \geq q$ かつ $s > t$ のときは統合できない。

$s/pA^{\sim}+t/qA^{\sim}$ のときはこの逆であって $p \geq q$ かつ $s \leq t$ ならば

$s/pA^{\sim}+t/qA^{\sim}=s/pA^{\sim}$ と統合できて「pA以下ならば度合いsで」と表現を単純化することができる。しかし

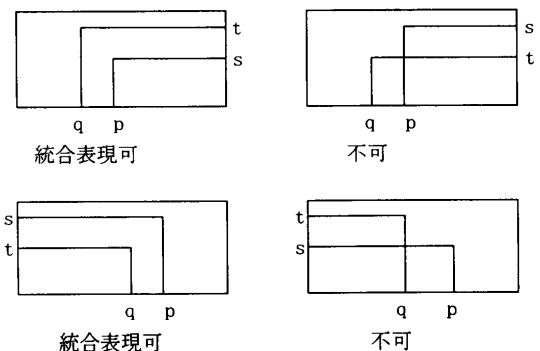


図1 同じカテゴリーのグレードと度合いの関係

$p \geq q$ かつ $s < t$ のときは統合できない。図 1 にこれらの事情を模式的に示す。

またもし度合いを全く問題にしなくてよいならば上記の $p \geq q$ のとき (8) は「 qA 以上ならば」と表現できる。 s と t の大小関係を問題にしないからである。

これより (7) の知識表現はつぎのようになる。いくつかの縮約のみ取り上げる。

第 1 項と第 4 項より「 $0.6A$ 以下ならば少なくとも度合い 0.1 でそのうち $0.3A$ 以下ならば少なくとも度合い 0.4 で識別できる」であるが度合いを問題にしなければ単に「 $0.6A$ 以下ならば識別できる」である。

また第 2 項と第 5 項より「 $0.7E$ 以上ならば少なくとも度合い 0.1 で、 $0.2E$ 以下ならば少なくとも度合い 0.4 で識別できる」であるがここでも度合いを問題にしなければ「 $0.7E$ 以上かつ $0.2E$ 以下ならば識別できる」である。

度合いを問題にしないというのはすべての度合いを取り上げることであるが、実用的には度合いにしきい値を設けて (0.4 以上など) しきい値未満は切り捨て、しきい値以上は度合いに関せずとすれば統合が増えて理解しやすいものとなろう。

5. 一般の大規模なデータの場合

一般にデータが大規模で表 8 のように識別行列が行・列とも複数、すなわちある目的の U もその他の U も複数あるときも同様にして縮約を導出することができる。

表 8 大規模なデータの識別行列
(G_{nm} は U_n が U_m と異なるカテゴリーの差分の和集合)

U	$n+1$	$n+2$...	m
1	$G_1(n+1)$	$G_1(n+2)$...	G_{1m}
2	$G_2(n+1)$	$G_2(n+2)$...	G_{2m}
...
n	$G_n(n+1)$	$G_n(n+2)$...	G_{nm}

表 8 の縮約は

U_1 より $G_1(n+1) * G_1(n+2) * \dots * G_{1m}$

U_2 より $G_2(n+1) * G_2(n+2) * \dots * G_{2m}$

...

のそれぞれより 4 章の方法によって計算される。それらの和集合が縮約の全体である。

6. 結論と今後の課題

本報告はカテゴリーに当てはまる度合いがグレード値 $[0, 1]$ で与えられたデータに対するラフ集合の極小縮約の計算法を提案した。

実用上は度合いの小さな縮約は無視すべきであること

が多い。現実世界におけるグレード値のあいまいさはたとえば 0.6 と 0.7 の区別をほとんど無意味にしているとみられる。そうすると度合いが 0.1 の縮約などは意味がない。おそらく 0.4 程度の値をしきい値としてそれ未満を切り捨てるとともにそれ以上は度合いを同一視して取り上げる (つまり度合いを 2 値化する) のが実用上よいかもしれない。これによって縮約の統合もすみ、理解が容易になる。

またグレード値を 10 段階で判定することが無理なこともあると思われる。5 段階とか 3 段階の方が実態に近いことが多い。あるいはカテゴリーごとのファジイ数、たとえば 5 段階である U のアイテム A が「かなり大きい A 」であったときそれを表すファジイ数として

$$0.3/0.6A + 1.0/0.7A + 0.6/0.8A$$

などとしてブール演算して縮約を求めることも複雑にはなるが考えられる。

参 考

4-1 節の (4) を演算則 (2) (3) を使って導出する計算過程を示す。便宜的な書き方である。

与えられた式は

$$(0.4A + 0.1E^{\sim} + 0.7G)(0.1A + 0.4E + 0.7G + 1.0J) \\ (0.4C + 0.4E + 0.2G + 1.0J)$$

という 3 因数からなる式である。

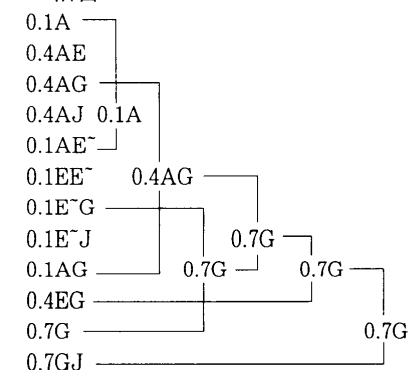
まず第 1、第 2 因数の積を and 結合の演算則で計算する。2 因数を行列の行と列に置いて行列の要素に計算結果を示すとこれらは or 結合の関係にある。

★ and 結合

	$0.1A$	$0.4E$	$0.7G$	$1.0J$
$0.4A$	$0.1A$	$0.4AE$	$0.4AG$	$0.4AJ$
$0.1E^{\sim}$	$0.1AE^{\sim}$	$0.1EE^{\sim}$	$0.1E^{\sim}G$	$0.1E^{\sim}J$
$0.7G$	$0.1AG$	$0.4EG$	$0.7G$	$0.7GJ$

つぎにこの 12 個を縦に並べ、or 結合の演算則を適用すると整理されていく。その過程を示す。

★ or 結合



残った項は

$$0.4AE + 0.4AJ + 0.1EE^{\sim} + 0.1E^{\sim}J + 0.1A + 0.7G$$

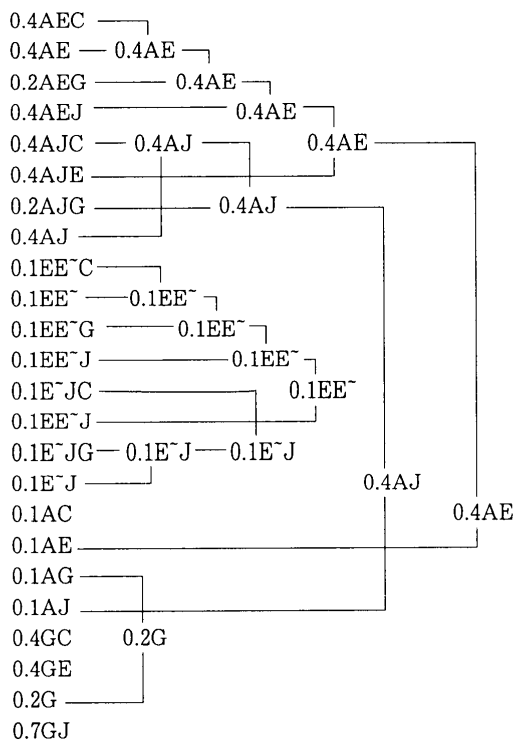
の6項である。

つぎにこれらと与えられた式の3番目の因数との積を上記と同様に計算し、そして縦に並べ、上記と同様に整理していく。

★ and 結合

	0.4C	0.4E	0.2G	1.0J
0.4AE	0.4AEC	0.4AE	0.2AEG	0.4AEJ
0.4AJ	0.4AJC	0.4AJE	0.2AJG	0.4AJ
0.1EE [~]	0.1EE [~] C	0.1EE [~]	0.1EE [~] G	0.1EE [~] J
0.1E [~] J	0.1E [~] JC	0.1E [~] J	0.1E [~] JG	0.4E [~] J
0.1A	0.1AC	0.1AE	0.1AG	0.1AJ
0.7G	0.4GC	0.4GE	0.2G	0.7GJ

★ or 結合



残った項は

$$0.1AC + 0.4GC + 0.4AE + 0.4GE + 0.2G + 0.4AJ + 0.7GJ + 0.1E^{\sim}J + 0.1EE^{\sim}$$

の9項であってこれが'(4)である。

注

- 1) 横山茂樹、津本周作、松岡喜美子、福並正剛：細菌検査データベースからのデータマイニング、第16回ファジイシステムシンポジウム講演論文集、p475-478、2000
- 2) 森典彦、高梨令、桑田起男、熊丸健一：属性の縮約を使って多様な商品選好を少数の新開発商品に集約する方法、第16回ファジイシステムシンポジウム講演論文集、p305-306、2000
- 3) 高梨令、森典彦、讀井純一郎、田慕玲：大規模データの縮約の実際と考察-大規模データからのラフ集合の縮約(2)、第2回日本感性工学会大会予稿集、p49、2000
- 4) 森典彦、高梨令：ラフ集合の概念による推論を用いた設計支援、東京工芸大学芸術学部紀要 Vol. 3、p35-38、1997
- 5) 森典彦、高梨令：製品の感性評価における判断過程についての一考察、第1回日本感性工学会大会予稿集、p48、1999
- 6) 井上勝雄、広川美津雄：認知部位と評価用語の関連分析、第2回日本感性工学会大会予稿集、p62、2000
- 7) 中村昭他、ラフ集合-その理論と応用(第6回)、数理科学12月号、1994
- 8) 森典彦、高梨令：ラフ集合の概念による推論を用いた設計支援、東京工芸大学芸術学部紀要 Vol. 3、p36、1997
- 9) ibid., p37