

形式概念にもとづく質的分析

Qualitative Analysis Based on Formal Concept

長田 博泰

Public opinion polls and most of social researches collect data on discrimination of right and wrong, a feeling of satisfaction and its degree, etc, which are qualitative data obtained from subjective judgments. The data are converted into nominal or ordinal scale data, and the resulting data are analyzed by statistical quantitative methods. This paper presents the application of formal concept analysis, not quantitative analysis, as a method of qualitative comparative analysis. The method focuses on concepts that are definable by conjunctions of properties, and makes it possible not only to classify data based on formal concepts, but also to clarify dependency relations among attributes. In order to demonstrate the usefulness and clarity of the method, several examples are illustrated, and formal concept analysis is compared with cluster analysis and Boolean approach, and discussed in a framework of classifying data.

1 はじめに

人間のさまざまな行為とその組織や社会との相互関係を研究対象とする社会科学の現象は、自然科学の現象とはつぎの意味で異なる。すなわち自然科学の現象は科学者の考察以前に解釈されてはいないし、現象を構成する物質、たとえば原子や分子や電子にとって現象が何かの「意味」をもつということはない。しかし、社会科学の現象は、社会科学者の考察に先立ち、そこで生活している人々によって既に解釈されているのである。したがって社会科学においては「純粋な事実」といったものは存在しない。

社会科学的現実がこのようなものだとするなら、社会科学が経験科学として社会認識の客観性を保つためには、社会現象を把握する何らかの方法が必要となる。その有力な手段が社会調査であり、その中には観察、聞き取り調査あるいは質問紙調査など多様な方法が含まれる。質問紙調査に基づく研究では調査対象を多くし、大数の法則に基づいた分析が可能である。一方、個別的・典型的事例を重視する研究では、観察あるいは聞き取り調査等が用いられる。この場合、データ量も少なく、観察者自身の主観性も入りやすくなる。前者は計量的研究、後者は質的研究といわれるが、両者の方法の背後にはつぎのような相違がある(Ragin, 1987)。以下の記述は鹿又(2001 : pp.3-5)に基づく。

Ragin は、統計的計量的方法を変数指向アプローチとよび、つぎの特徴を指摘する。(1)変数および変数間関係によって仮説を特定すること、そして仮説の検証を重視する。(2)(他の変数と関連をもたない変数を省くという意味で)より節約的な説明(モデル)を選好する傾向をもつ。(3)社会現象の複雑性よりも一般性を重視して、一般化された言明による命題導

出をめざす。

そして、この変数指向アプローチの背後にはつぎの見方があるという。すなわち、社会現象の多様性や因果関係の複雑性を軽視する傾向がある。一般化をめざすため、個別の事例にあらわれる特異な現象を誤差やはずれ値として処理しがちである。因果関係の分析で用いられる統計モデルは、各変数の影響力がそれぞれ独立に従属変数にはたらくという前提のもとにモデル化されたものが多い。

これと対照的なアプローチは、少数の事例を詳細に考察する質的研究(質的調査)である。質的分析はつぎのような特徴をもつ。(1)事例一つ一つを全体としてとらえようとするので、各事例を変数に分解できないものとみる。(2)社会現象の多様性や因果関係の複雑さを重視するため、一般的傾向からずれた事例を例外や誤差として処理しない。(3)質的説明を多用して、各事例に見られる固有な現象の理解と説明をめざす。

このようなアプローチが、逆に、分析結果の一般性や客観性を保証しないという指摘を受けやすい。それは研究者の判断にゆだねられる部分が多いためである。また事例の選択が典型事例や極端な事例にかたよりがちであり、少数の事例、例証にとどまることも少なくない。

以上の問題点を解決する方法として、1980年代前半に提案された形式概念による分析を採用する(Wille, 1982)。長田(2004)は、この分析方法をもちいて価値判断過程の比較記述しHasse図を援用したが、社会調査データにも適用可能であり、その比較分析に有望だと考えられるからである。すなわち、Raginが提案したブール代数的方法(Ragin, 1984)を包含し、さらにより多様な分析を可能にする方法であるためである。それはつぎの点である。(1)ブール代数的方法のように変数値は二値に限らず、任意の名義尺度、順序尺度を扱うことが可能である。(2)社会現象の多様性をそのまま扱うことができる。(3)分析手続きが客観的である。(4)より節約的なモデルを想定しない(結果としてより節約的なモデルが導出される可能性はある)。

本稿の目的はつぎのとおりである。ひとつは、管見する限り、日本で人文科学、社会科学的数据分析に形式概念分析を適用した例が見当たらないので、具体的事例に適用しその有効性を示すことである。また、Hasse図に依らずとも形式概念分析から得られるデータを解析する方法を提示することである。最後に、本方法と統計的数量的方法を具体例で比較し、それぞれの長短を指摘し、両者の併用の有効性を示すことである。

以下、2節で形式概念分析の展開に必要な定義と分析方法を説明し、3節では質的データで記述される具体例として車のデザイン分析に適用し、方法の特徴と問題点を指摘する統計的数量的方法との併用および比較を検討するため、4節では数量的データへ形式概念分析を適用する。5節では形式概念分析とクラスター分析、ブール代数的方法との長短を比較考察する。

2 形式概念分析

順序は数学だけでなく、日常生活のあらゆる面に浸透している。一番、二番、…はもちろん、大きいー小さい、よいーわるい、満足ー不満など。これらにその程度を表す「まあ」あるいは「やや」などの形容詞をつけることも可能である。

順序や順序構造を扱う数学の一分野は束論である。とくにデータの集合の複雑な構造を解析、分類する強力な手法が束論の枠組みの中で比較的新しく研究されてきた、すなわち形式概念分析(formal concept analysis)がそれである(Wille, 1982)。管見する限り、形式概念分析を詳細に説明した日本語文献もなく、ましてや社会科学(者)のために紹介したものもないようなので、まず Ganter & Wille(1999), Davey & Priestley(2002)に基づきながら、以下の展開に必要な限りでその要点を解説することから始めよう。

2.1 形式概念

必要最小限の定義をやや形式的に述べるが、定義の後に直感的な説明を加えておく。

定義1：形式文脈(formal context) $K=(G, M, I)$ は2つの集合 G と M 、および G と M の間の関係 I からなる。 G の要素を対象、 M の要素を文脈の属性という。対象 g が属性 m と I の関係にあることを gIm あるいは $(g, m) \in I$ と書き、“対象 g は属性 m を有する”と読む。 (G, M) はそれぞれドイツ語 Gegenstande, Merkmale に由来する)

小さな(形式)文脈は図1のクロス表で表現することができる。すなわち、各行に対象、各列に属性を割り当て、行 g と列 m の交叉した位置が対象 g は属性 m を有することを意味する。上の定義はそれを数学的に述べただけである。

	m	
	...	
gx	

属性

対象

図1 形式文脈

つぎに形式概念を定義するため、記法をひとつ導入する。

定義2：対象の集合 $A \subseteq G$ に対しつぎの集合を定義する：

$$A' = \{m \in M \mid \text{すべての } g \in A \text{ に対し } gIm\}$$

すなわち、 A の対象に共通する属性の集合である。同様に、属性の集合 $B \subseteq M$ に対しつぎの集合を定義する：

$$B' = \{g \in G \mid \text{すべての } m \in B \text{ に対し } gIm\}$$

すなわち、 B のすべての属性を有する対象の集合である。

定義 3 : 形式文脈 (G, M, I) の形式概念(formal concept)とは対 (A, B) である。ここで、 $A \subseteq G, B \subseteq M, A' = B, B' = A$ である。概念 (A, B) の A を外延, B を内包と呼ぶ。

概念 (A, B) の外延 A と内包 B は関係 I によって結合されている, すなわち, $B' = A$ と $A' = B$ であるから, 概念のおのおのは他を決定づけ, それによって概念を決定する。概念 (A, B) はクロス表で図 2 のように表すことができる。

		内包			
		B			
					外 延
A		×	×	×	
		×	×	×	
		×	×	×	

図 2 形式概念

形式文脈が与えられたとき, その形式概念の間に順序, すなわち, その外延間の包含関係がある場合がある。これを正確に定義するとつぎのようになる。

定義 4 : (A_1, B_1) と (A_2, B_2) を文脈の概念とする。 $A_1 \subseteq A_2$ (同じことであるが, $B_2 \subseteq B_1$) なら (A_1, B_1) を (A_2, B_2) の下位概念, (A_2, B_2) を (A_1, B_1) の上位概念といい, $(A_1, B_1) \leq (A_2, B_2)$ であらわす。関係 \leq を概念の階層的順序という。このように順序づけられた (G, M, I) のすべての形式概念の集合を $B(G, M, I)$ で表し, 文脈 (G, M, I) の概念束という。

このように定義された概念束は, 完備束とよばれる束になることが証明できるが, これについて詳述する必要がないので省略する。

2.2 形式概念の階層化と図示

このままでは概念束の構造を直感的にとらえることができず, 応用もしにくいので, 概念束を図示化することを考える。

アルゴリズム 1: 外延の決定

ステップ 1 : 文脈 (G, M, I) の概念のすべての外延を見つける。

- (1.1) 属性と外延という 2 つの列からなる表(以下, $[m | m']$ 表という)を作る。属性列の最初のセルを空にし, 外延列の最初のセルに G を書く。

- (1.2) クロス表から属性の極大な外延を見つける (残っている属性の中で外延が最も大きいものを選ぶ), これを m' とする。
- (1.2.1) m' がまだ外延列に現れていないなら, $[m | m']$ 表に行 $[m | m']$ を付け加える。集合 m' と外延列の既にあるすべての外延と共通集合を求める。この共通集合を外延列に(リストになければ)付け加え, 属性列のセルを空にする。
- (1.2.2) m' がすでに外延列に現れているなら, m' がある行の属性セルに m を追加する。
- (1.3) クロス表から属性 m の列を削除する。
- (1.4) 最後の列が削除されたなら, 終了。さもないと, (1. 2) に戻る。
- ステップ 2 : $[m | m']$ 表からダイアグラム (Hasse 図) を描く

対象, 属性の数が少ない場合はアルゴリズム 1 で得られた集合から直接, Hasse 図を描くことも不可能ではない。その数が多い場合, コンピュータで処理する必要がある。そのようなプログラムも開発され, 公開されているものもある (例えば, Freese)。しかし, Hasse 図を描いても, 外延が多数の場合, 必ずしも見通しのよいものではない。

人手であれ, コンピュータであれいずれの場合でも, アルゴリズム 1 の外延集合を階層化しておくこと全体像の把握が容易である。外延集合を階層化する方法を示す。

アルゴリズム 2: 外延の階層化

ステップ 0 : アルゴリズム 1 の $[m | m']$ 表の各外延に上から順に通し番号を付ける。

ステップ 1 : 外延間のすべての包含関係をみつける。

ステップ 2 : 以下の手順で外延を階層化する :

- (2.1) すべての外延の通し番号の集合を S とする。
- (2.2) S の各外延が含む外延を見つけ, その通し番号の和集合 W をつくる。
- (2.3) S と W の差集合を R とし, R を出力する。
- (2.4) R が空集合でなければ, W を S とし, (2.2) に戻る。
空であれば, 終了。

この手順によって, 属性の組合せ個数, すなわち, $1, 2, \dots, n$ 個の順に外延集合が確定する。最初の外延の通し番号の集合は独立した属性を示し, 最後の外延の通し番号の集合が対象の最終的分類である。以上の記述だけではわかり難いと思われるので, 例を掲げる。

2.3 例

上述の形式概念分析のアルゴリズムを簡単な例を用いて説明する。太陽系を記述した属性を表 1 に示す (Davey & Priestley, 2002 : p.65)。

表1 太陽系クロス表

	大きさ			太陽からの距離		衛星の有無	
	1 小	2 中	3 大	4 近い	5 遠い	6 有	7 無
水星	×			×			×
金星	×			×			×
地球	×			×		×	
火星	×			×		×	
木星			×		×	×	
土星			×		×	×	
天王星		×			×	×	
海王星		×			×	×	
冥王星	×				×	×	

9個の惑星と7個の属性(3+2+2)に対し、アルゴリズム1を適用する。まず[m | m']表のm欄を空にし、m'欄に全対象を記入する。つぎに各属性の外延の中で最も個数の多い属性、(衛星)有(属性6)を選択し、m欄に6、m'欄にその外延を記入し、クロス表から属性6を削除する。まだ属性が残っているので、同様の操作を続ける。すなわち、属性の外延の中で最も広いものを選択する。属性1と5が候補になりうるが、どちらを選んでもよい。ここでは属性5を選択することにする。属性5の外延は既に記入済みの外延と異なるので、[m | m']表にそれぞれ5、5の外延を記入する。属性5の外延と記入済みの属性6の外延の共通集合は属性5の外延と同じなので、属性5の処理を終え、クロス表から属性5を削除する。属性1の外延は明らかに属性6、5の外延と異なるので、[m | m']表にそれぞれ1、1の外延を記入する。属性1と属性6の外延の共通集合をとると、既に記入済みの集合とは異なる集合が得られるので、[m | m']表にそれぞれ(1, 6)、共通集合を記入する。以上の操作をクロス表の属性がすべて削除されるまで実行する。その結果を表2に示す。

表2 表1に対する[m | m']表

通番	m(属性)	m' (外延)
0		[冥王星, 土星, 地球, 天王星, 木星, 水星, 海王星, 火星, 金星]
1	6	[冥王星, 土星, 地球, 天王星, 木星, 海王星, 火星]
2	5	[冥王星, 土星, 天王星, 木星, 海王星]
3	1	[冥王星, 地球, 水星, 火星, 金星]
4	1, 6	[冥王星, 地球, 火星]
5	1, 6, 5	[冥王星]
6	4	[地球, 水星, 火星, 金星]
7	4, 6	[地球, 火星]
8	4, 6, 5	[] (空集合)
9	2	[天王星, 海王星]
10	3	[土星, 木星]
11	7	[水星, 金星]

得られた外延にアルゴリズム2を適用しよう。ステップ0にしたがい、各外延に0~11の番号をつける(各属性も番号1~7で示したので、両者を混同しないこと)。ステップ1の外延の包含関係を見出した後、直接の包含関係だけを取り出す。結果は以下のとおりである。

外延の直接包含関係：

- 1 \supseteq 2 1 \supseteq 4
- 2 \supseteq 5 2 \supseteq 9 2 \supseteq 10
- 3 \supseteq 4 3 \supseteq 6
- 4 \supseteq 5 4 \supseteq 7
- 6 \supseteq 7 6 \supseteq 11

通番0(全対象)と8(空集合)を除く外延の中で包含関係の左辺に出現しない外延、すなわち、5, 7, 9, 10および11は対象の分類結果である。ステップ2を適用し、外延を階層化すると、外延間の関係が一層明確になる。結果はつぎのとおりである。

ノードの階層：

- 深さ 0: [0]
- 深さ 1: [1, 3]
- 深さ 2: [2, 4, 6]
- 深さ 3: [5, 7, 9, 10, 11]
- 深さ 4: [8]

以上の情報が得られると、ダイアグラムを容易に描くことができる。仮に描けなくても

属性間の関係および属性による対象の分類に関して十分な情報が得られている。表 1 の Hasse 図を図 3 に示す。

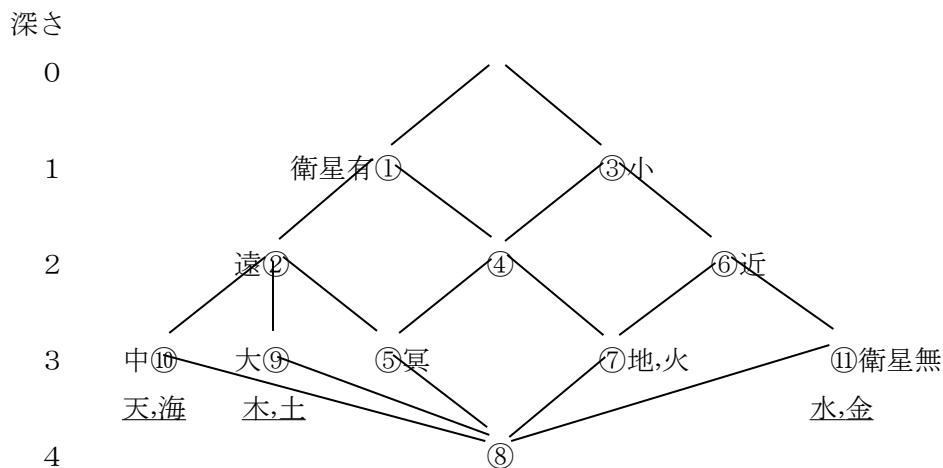


図 3 太陽系の Hasse 図

3 適用例一車のデザイン特性分析

形式概念分析をもちいて対象の質的比較を行うには、対象の属性値が順序尺度であるか、あるいは、何らかの名義尺度でなければならない。前者は、たとえば「重視する」から「重視しない」、あるいは「満足している」から「不満である」のように、自然で明確な順序関係からなっていることを意味する。後者は、はいいいえや帰属を表す場合のように名義的な値である。

ここでは、自動車のフロント部分のデザイン特性に形式概念分析を適用する。データとして森他(2004, pp. 26-33)を借用する。森は、誰もがデザインとして認知していると思われる7箇所を選定し、国内外のメーカーの60車種について3段階の属性値を設定した。そのうち国産38車種の結果をクロス表として表3に掲げる。

表3に形式概念分析したところ、全ノード数236、分類個数が27、深さ9となった。各車種を属性記述し、分類することができたが、全体的あるいは特徴的な事実が見えてこない。そこで、メーカーごとに形式概念分析を適用し、特徴的な点がないか調べてみたところ、メーカーごとに特徴らしきものが浮かび上がってきた。その中から2つのメーカーのHasse図を図4、図5に例示する。

日産車のHasse図から、深さ5(トップとボトムを除く)、対象記述に用いられた属性数10、一グループ(2車種)を除き、他の車種はすべて5つの属性で記述される。例えば、Cima, SunnyはB2・C2・D2・F2, CedricはA2・B1・C3・D2・F2である。一方、トヨタのそれは、深さ6、対象記述に用いられた属性数11、4属性記述のもの1車種、5属性2車種、6属性4車種となっている。

表3 自動車のフロントデザインと車種別クロス表																					
車種	A 造形			B センター			C グリル			D ランプ			E 表情			F バンパ-			G 縦横		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Micra		x				x		x			x			x		x			x		
Cube		x		x				x			x			x			x			x	
Almera		x				x		x				x			x			x			x
Sunny			x		x			x			x			x			x			x	
Tino		x				x		x			x			x			x			x	
Primera		x				x		x			x			x			x			x	
Presage		x			x				x		x			x			x			x	
Cedric		x		x					x		x			x			x			x	
Cima			x		x			x			x			x			x			x	
WillVi	x			x				x			x			x			x			x	
Yaris	x				x		x				x				x			x		x	
Starlet		x		x				x			x			x			x			x	
Carolla		x			x			x			x			x			x			x	
Celica	x				x		x				x				x			x		x	
Gaia		x			x			x			x			x			x			x	
Camry		x		x				x			x			x			x			x	
Progres			x	x			x			x				x		x			x		
Supra	x			x			x				x			x			x			x	
Civic		x			x			x			x			x			x			x	
Integra	x			x			x			x			x			x			x		
Accord		x			x			x			x				x			x		x	
Prelude		x		x				x		x				x			x			x	
Legend			x		x			x			x			x			x			x	
Capa		x			x			x			x			x			x			x	
HRV		x		x				x			x			x			x			x	
Lancer		x			x			x			x			x			x				x
Mirage		x		x				x			x			x			x				x
Dingo		x		x				x				x			x	x					x
Carisma		x			x				x		x			x			x				x
Galant			x		x				x		x				x		x				x
GTO		x		x			x				x			x				x	x		
SpaceWG		x			x				x		x			x			x				x
MZDemio		x			x			x			x		x				x				x
MZ626		x			x			x			x			x			x				x
MZEunos		x			x			x			x				x		x				x
Impreza		x		x				x			x			x			x				x
Legacy		x			x			x			x			x		x					x
Forester		x		x				x			x			x		x					x

(森典彦, 他著『ラフ集合と感性』 p.29)

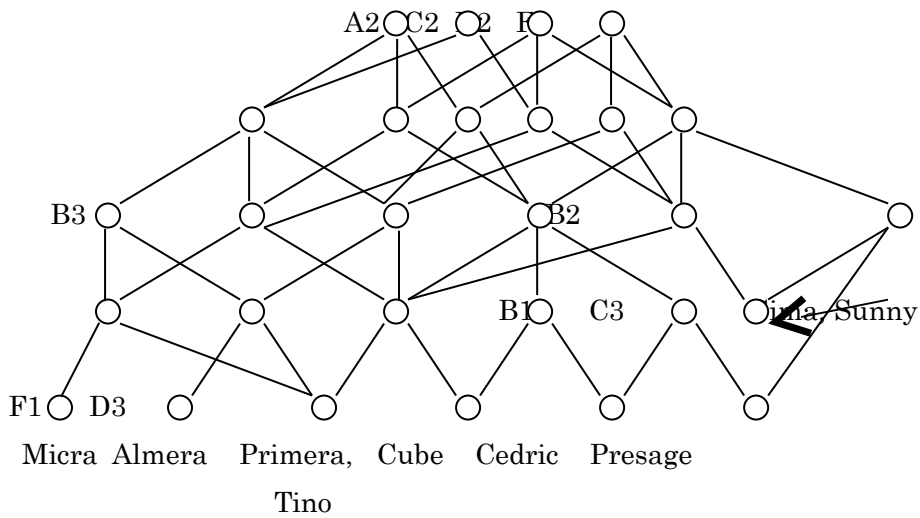


図4 日産車の属性 Hasse 図

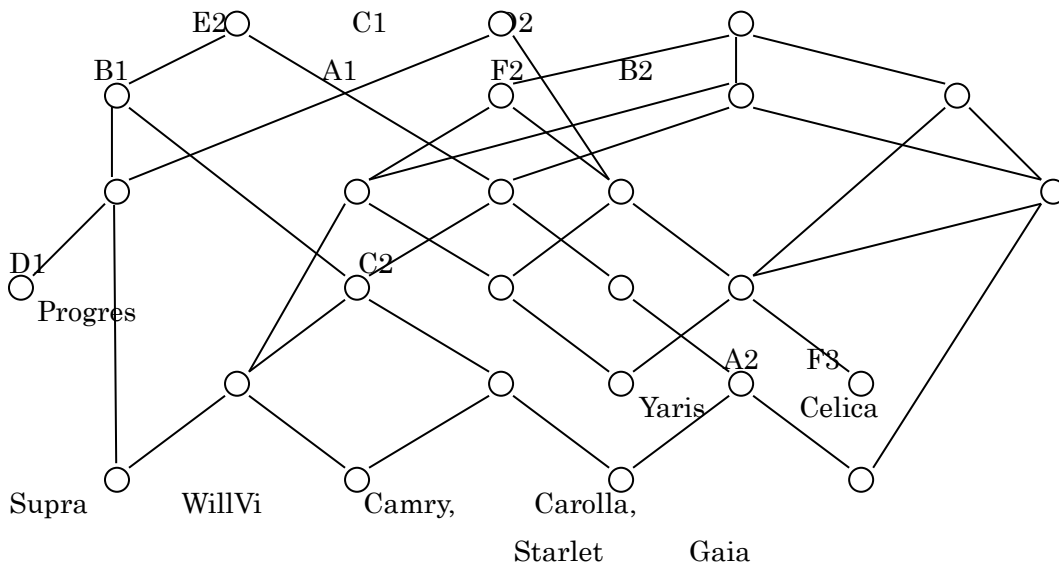


図5 トヨタ車の属性 Hasse 図

この事実はずぎのように解釈することができる。日産車は全車種に共通する属性が比較的多く、これにそれぞれの車種の特徴を表す属性が附加されている。これに対し、トヨタ車は、日産車に比べ車種共通属性が少なく、これに特徴的属性を附加しながら広がりをもった車種を構成している。

メーカーごとのデザイン特性が見えてきたので、表3にメーカーを示す属性(日産、トヨタ、ホンダ、三菱、その他メーカーにそれぞれ、1, 2, 3, 4, 5の属性値を与えた)を追加

し分析した。結果は全ノード数 323, 分類個数 34, 深さ 10 となった。仮に Hasse 図を描いても特徴を読み取れそうにない。

ところが, 下記の深さ 1 の属性に興味深い事実, すなわちメーカーを表す属性のうちトヨタと三菱が出現していることに気づいた。

D2, F2, E2, A2, C2, G2, B1, G1, Y2(トヨタ), Y4(三菱), E3, F1, A3, A1

これはトヨタ, 三菱の車種が全体として固有の特性を有することを意味するが, 車種ごとに特性がないことを意味しない。それは, 表 4 に示すようにグループを成す [Cima, Sunny], [Primera, Tino] (以上, 日産), [Carolla, Gaia], [Camry, Starlet] (以上, トヨタ) も含め, 30 車種それぞれが, 異なる属性の組合せ, 個性を有していることからわかる。

表 4 メーカー属性を含む分類

[Capa]	[Dingo]	[Legacy]
[Supra]	[Lancer]	[Forester]
[GTO]	[Mirage]	[Progres]
[Prelude]	[Civic]	[Legend]
[Presage]	[HRV]	[WillVi]
[Cima, Sunny]	[Integra]	[Primera, Tino]
[Cube]	[MZ626]	[Carisma]
[Micra]	[Impreza]	[Celica]
[Carolla, Gaia]	[Galant]	[Almera]
[Camry, Starlet]	[Accord]	[MZDemio]
[Yaris]	[MZEunos]	34 個の分類
[SpaceWG]	[Cedric]	

これは対象記述の特性である以上に, 分析手続き上の方法に重要な示唆を与える。すなわち, メーカー属性を日産, ホンダ, その他メーカーからなるグループ, トヨタおよび三菱に三分し, それぞれのグループを独立に形式概念分析してよいということである。これは, さらに Hasse 図を全対象から描く必要はなく, 三分されたグループごとの Hasse 図を結合し, 全体の Hasse 図を描いてよいことを意味する(長田, 2004 : p.91)。したがって, 全体の Hasse 図を図 6 に示すように描くことができる。

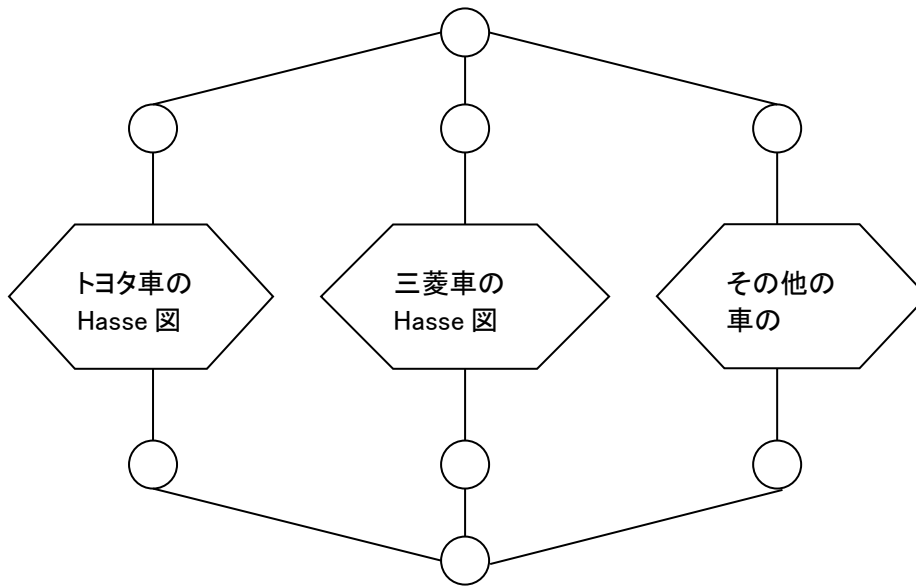


図6 全車種の Hasse 図

このように形式概念分析によって対象を区分する有意味な属性を見つけられることは、データ解析、とくに社会調査データの解析にとってきわめて有効である。なぜなら、それがわかれば、まず区分する属性に注目した分析に専念し、その上で帰属の社会学的意味等を問い、必要なら、さらに調査・研究を深化させることができるからである。

以上を要するに形式概念分析は以下の特徴を有する：

- 1) 与えられたデータを加工せずに対象を属性によって記述する。
- 2) 対象の分類に必要な属性を特定することができる。残余の属性は対象分類に与らない。これは調査項目が説明変数として有効かどうかを判定する際に有効である。
- 3) Hasse 図を描かなくても、ノードの深さおよび分類結果を調べることによって全体構造の概略を把握することができる。

Hasse 図を機械的に描くことは可能であるが、一般にノード数が多くなれば見通しのよいものを描くことは至難である。しかし、上例のようにノード数が多い場合、対象を区分する属性を見出すことができるなら、各分類を独立に分析し考察することができる。

- 4) 深さ n のノードは n 個の属性の論理積を表している。したがって、深さおよびノード数は対象を規定する論理積による論理式の複雑さに対応する。

同じ属性の異なる対象を分析すればわかるように、属性の組合せの数によって Hasse 図の深さは異なる、また、簡単な組合せで対象を規定できれば、ノード数も少なくなる。

- 5) 名義尺度を導入することによって性別、職業、学歴等の社会学的変数も属性として分析することができる。これによって属性の性質等を意識せずに分析対象とすることができ、その結果から属性の意味づけを考察することができる。

4 量的データと形式概念分析

上述のように形式概念分析は質的データを対象とする。しかし、質的データであっても数量の多いアンケート調査等ではそのまま形式概念分析を行うことは必ずしも適当ではない。なぜなら、対象が多過ぎて分類しても見通しのよい結果が得られるとは限らないからである。そのような場合、量的データを質的データに変換し形式概念分析を適用すると、その威力を発揮し、また適用範囲が一層広がる可能性がある。

ここでは、そのような例を2つ採りあげる。一つは、質的データを数量化したものに形式概念分析を適用する場合である。いまひとつは、既に数量分析された結果に形式概念分析を適用し、質的比較を行う例である。この方法は、統計的計量的方法ではクラスター分析に対応するものであるが、これらの方法に比し容易であることを示す。

4.1 量的データの形式概念分析

量的データを質的データに変換し形式概念分析を適用する例として、表5の東京23区の職種区分別人数から23区の分類をとりあげる。表5のデータを質的データに変換する手順は、通常の統計的方法と変わらない(上田, 2003: pp.22-27)。すなわち、表5の各区および合計の職種構成をそれぞれ各区人口合計、合計人口で割って、構成比を求める。さらにそれぞれの合計職種構成比で各区職種構成比を割り算し、構成比の相対比(=特化係数)を計算する。この結果とこれを特徴パターンで表現した結果を表6に示す。

表5 東京23区住民の職種構成(1965年)						
地域	職種区分(千人)					合計
	A	B	C	D	E	
千代田	8	10	15	14	10	57
中央	7	13	24	22	14	80
港	20	27	24	40	23	134
新宿	33	46	39	65	37	220
文京	19	28	25	45	17	134
台東	14	24	43	66	25	172
江東	14	23	31	99	15	182
墨田	10	27	30	111	17	195
品川	24	45	37	99	27	232
目黒	23	36	23	52	16	150
太田	43	77	56	184	35	395
世田谷	62	85	53	102	32	334
渋谷	22	34	25	42	22	145
中野	27	46	34	66	20	193
杉並	48	68	46	64	22	248
豊島	23	44	37	68	25	197
北	21	51	40	99	24	235
荒川	9	21	26	82	12	150
板橋	21	45	34	118	19	237
練馬	30	44	31	70	18	193
足立	15	38	37	143	18	251
葛飾	15	36	33	124	18	226
江戸川	13	31	30	112	16	202
合計	521	899	773	1887	482	4562
A: 管理的職種, B: 事務従事者, C: 販売従事者						
D: 工場労働者, E: サービス職業従事者						
(「国勢調査」, 総理府統計局)						

表6 東京23区住民職種構成											
地域	職種区分—特化係数					地域	職種区分—特徴パターン				
	A	B	C	D	E		A	B	C	D	E
千代田	1.23	0.89	1.55	0.59	1.66	千代田	.	.	++	--	++
中央	0.77	0.82	1.77	0.66	1.66	中央	-	.	++	--	++
港	1.31	1.02	1.06	0.72	1.62	港	+	.	.	-	++
新宿	1.31	1.06	1.05	0.71	1.59	新宿	+	.	.	-	++
文京	1.24	1.06	1.10	0.81	1.20	文京
台東	0.71	0.71	1.48	0.93	1.38	台東	-	-	+	.	+
江東	0.67	0.64	1.01	1.32	0.78	江東	-	--	.	+	-
墨田	0.45	0.70	0.91	1.38	0.83	墨田	--	-	.	+	.
品川	0.91	0.98	0.94	1.03	1.10	品川
目黒	1.34	1.22	0.90	0.84	1.01	目黒	+
太田	0.95	0.99	0.84	1.13	0.84	太田
世田谷	1.63	1.29	0.94	0.74	0.91	世田谷	++	+	.	-	.
渋谷	1.33	1.19	1.02	0.70	1.44	渋谷	+	.	.	-	+
中野	1.22	1.21	1.04	0.83	0.98	中野
杉並	1.69	1.39	1.09	0.62	0.84	杉並	++	+	.	--	.
豊島	1.02	1.13	1.11	0.83	1.20	豊島
北	0.78	1.10	1.00	1.02	0.97	北	-
荒川	0.53	0.71	1.02	1.32	0.76	荒川	--	-	.	+	-
板橋	0.78	0.96	0.85	1.20	0.76	板橋	-	.	.	.	-
練馬	1.36	1.16	0.95	0.88	0.88	練馬	+
足立	0.52	0.77	0.87	1.38	0.68	足立	--	-	.	+	-
葛飾	0.58	0.81	0.86	1.33	0.75	葛飾	--	.	.	+	-
江戸川	0.56	0.78	0.88	1.34	0.75	江戸川	--	-	.	+	-

A: 管理的職種, B: 事務従事者
C: 販売従事者, D: 工場労働者
E: サービス職業従事者

"--" < 1/1.5, 1/1.5 ≤ "-" < 1/1.25
1/1.25 ≤ "." < 1.25
1.25 ≤ "+" < 1.5, 1.5 ≤ "++"

この後の形式概念分析手順は上述のとおりであるので詳細は省略する。結果を Hasse 図に描いたものが図7である。

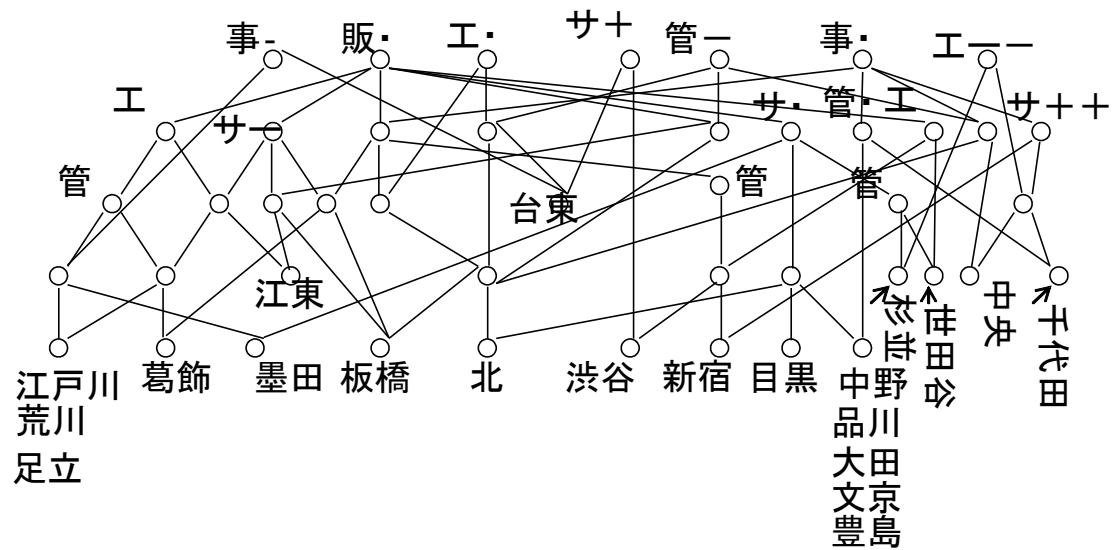


図7 職種構成による東京23区のHasse図

図7からつぎのことがわかる。すなわち20個の属性のうち最上位に出現する属性は7個、属性記述に用いられた属性は16個であり、最終レベルの分類は15個である。しかし、23区を15に分類するのは、分類として細かすぎるといえる。この点については5節で論ずる。

4.2 既存数量分析への適用－文化の国際比較

既に数量分析された結果に形式概念分析を適用する例としての価値観の国際比較データを取りあげる。近年、この種のデータが容易に利用可能であるが(たとえば、『価値観データブック』, 2003), 質的比較自体が目的ではなく、形式概念分析法を統計的方法と比較し、その有効性を示すことにあるので、既に統計的方法で分析されたデータに適用する。

G. Hofstede は、文化の比較を目的とし、世界各国に支社を有する多国籍企業 IBM 社の社員を対象に1980年に40カ国、1991年に50カ国と3地域の質問紙調査によるデータを分析・公表している。ここでは、後者のデータを用いる(G. Hofstede, 1991=1995)。

文化のもっとも中枢にあるのは価値観である。価値観は、肯定的な側面と否定的な側面の両極をあわせ持つ感情であるから、選択肢を示し人々の好みをたずねることが可能である。もちろん、その回答を文字どおりに受け取ってはならない。実際、人々は調査票に記入した回答通りに行動するとは限らないからである。しかし、ある質問に対し、肯定的、あるいは選り好みする人が、B集団よりもA集団に多ければ、この二つの集団では、文化的な違いがあると考えられる。

Hofstede は、文化的相違をもたらす要因を予備調査および文献調査により四つの価値の測定次元、すなわち、権力の格差、不確実性の回避、個人主義対集団主義、男性的価値対女性的価値を設定し、調査を行った。

質問の大部分は5段階の順序で回答するものであり、これに5, 4, 3, 2, 1と点数化し、整理したものを表7に示す。表7中の各価値測定次元の「高」、「中」、「低」は著者が各次元のスコアの度数分布から3段階に区分したものである。4つの測定次元を各々3段階に区分した12個が形式概念分析における属性となる。対象は既に述べたように50ヶ国と3つの地域(アラブ諸国、東アフリカ諸国、西アフリカ諸国)である。

Hofstede は、4つの測定次元に対しつぎの6つの組合せについて散布図を描き、図上にクラスタリング結果を書き入れている。

- | | |
|----------------------|-----------|
| (1) 権力格差指標×個人主義指標 | 6分類 |
| (2) 権力格差指標×男性らしさ指標 | 8分類+5カ国 |
| (3) 男性らしさ指標×個人主義指標 | 9分類+11カ国 |
| (4) 男性らしさ指標×不確実性回避指標 | 10分類 |
| (5) 不確実性回避指標×個人主義指標 | 10分類+13カ国 |
| (6) 権力格差指標×不確実性回避指標 | 11分類 |

ここでは、4つの測定次元の組合せに対する分類ではなく、すべての属性、すなわち $4 \times 3 = 12$ 個の属性を用いた対象の分類を試みる。これが容易に可能なのは形式概念分析の

特徴である。結果はつぎのとおりである。

- (1) 109 個の外延, すなわち Hasse 図は 109 個のノードからなる。
- (2) 深さは 6 である。
- (3) 50 カ国と 3 地域が表 6 に示すように 28 個に分類された。

Hasse 図を描くと複雑なので省略するが, その特徴を捉えるのはそれほど困難ではない。まず深さ 1 のノード, すなわち属性 1 個のノードが 12 個である。これは表 7 の 12 個の属性が他の属性に含まれることはない, つまり 12 個の独立した属性であることを示している。その意味で Hofstede が選択した 4 つの価値測定次元は適切であり, またその程度を 3 分したことにも問題がないことを示している。

つぎに 28 個の分類のうち, 27 個は深さ 4 のノードに出現している。言いかえると, 27 個の分類は 4 つの価値測定次元の程度の組合せできまり, 各分類が異なる文化的特徴を有することを意味する。残り 1 カ国 COS(コスタリカ)は深さ 3, すなわち 3 つの属性できまる。コスタリカは中南米諸国の中で最も政治的経済的に安定した国だといわれているが, その反映であろうか。

国	権力格差指標			個人主義指標			男らしさ指標			不確実性回避指標						
	スコア	1 高	2 中	3 低	スコア	4 高	5 中	6 低	スコア	7 高	8 中	9 低	スコア	10 高	11 中	12 低
ARA	80	×			38		×		53		×		68		×	
ARG	49		×		46		×		56		×		86	×		
AUL	36			×	90	×			61	×			51		×	
AUT	11			×	55		×		79	×			70		×	
BEL	65		×		75	×			54		×		94	×		
BRA	69		×		38		×		49		×		76		×	
CAN	39			×	80	×			52		×		48		×	
CHL	63		×		23		×		28			×	86	×		
COL	67		×		13			×	64	×			80		×	
COS	35			×	15			×	21			×	86	×		
DEN	18			×	74	×			16			×	23			×
EGY	64		×		27		×		41		×		52		×	
EQA	78	×			8			×	63	×			67		×	
FIN	33			×	63	×			26			×	59		×	
FRA	68		×		71	×			43		×		86	×		
GBR	35			×	89	×			66	×			35			×
GER	35			×	67	×			66	×			65		×	
GRE	60		×		35		×		57		×		112	×		
GUA	95	×			6			×	37		×		101	×		
HOK	68		×		25		×		57		×		29			×
IND	78	×			14			×	46		×		48		×	
IND	77	×			48		×		56		×		40			×
IRA	58		×		41		×		43		×		59		×	
IRE	28			×	70	×			68	×			35			×
ISR	13			×	54		×		47		×		81		×	
ITA	50		×		76	×			70	×			75		×	
JAM	45		×		39		×		68	×			13			×
JPN	54		×		46		×		95	×			92	×		
KOR	60		×		18			×	39		×		85	×		
MAL	104	×			26		×		50		×		36			×
MEX	81	×			30		×		69	×			82		×	
NET	38			×	80	×			14			×	53		×	
NOR	31			×	69	×			8			×	50		×	
NZL	22			×	79	×			58		×		49		×	
PAK	55		×		14			×	50		×		70		×	
PAN	95	×			11		×		44		×		86	×		
PER	64		×		16		×		42		×		87	×		
PHI	94	×			32		×		64	×			44		×	
POR	63		×		27		×		31			×	104	×		
SAF	49		×		65	×			63	×			49		×	
SAL	66		×		19			×	40		×		94	×		
SIN	74	×			20			×	48		×		8			×
SPA	57		×		51		×		42		×		86	×		
SWE	31			×	71	×			5			×	29			×
SWI	34			×	68	×			70	×			58		×	
TAI	58		×		17			×	45		×		69		×	
THA	64		×		20			×	34			×	64		×	
TUR	66		×		37		×		45		×		85	×		
URU	61		×		36		×		38		×		100	×		
USA	40			×	91	×			62	×			46		×	
VEN	81	×			12			×	73	×			76		×	
WAF	77	×			20		×		46		×		54		×	
YUG	76	×			27		×		21			×	88	×		
平均	56.83019	13	24	16	43.0566	17	21	15	48.73585	16	27	10	65.45283	17	27	9
STD	21.81978				25.4176				18.32521				24.31569			

表8 50ヶ国3地域の分類

No.	国または地域名	No.	国または地域名
1	BRA, EAF, IRA	15	GUA, PAN
2	CAN, NZL	16	YUG
3	ITA, SAF	17	MEX, PHI
4	BEL, FRA	18	IDO, WAF
5	ARG, GRE, SPA, TUR, URU	19	EQA, VEN
6	JPN	20	THA
7	ISR	21	CHL, POR
8	COS	22	FIN, NET, NOR
9	AUT	23	HOK
10	AUL, GER, SWI, USA	24	JAM
11	PAK, TAI	25	GBR, IRE
12	KOR, PER, SAL	26	SIN
13	COL	27	IND, MAL
14	ARA	28	DEN, SWE

5 他の方法との比較・関連

形式概念分析の基本的考え方、方法および適用例を述べてきたが、ここで同様な発想、あるいは目的、効果をもたらす方法と比較しその長短を考察しておきたい。採り上げるのはクラスター分析とブール代数的アプローチである。

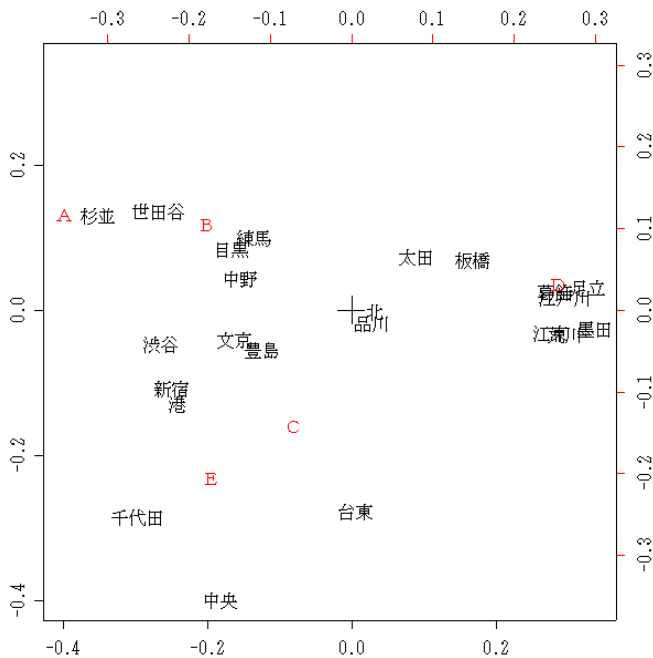
5.1 クラスター分析との比較

形式概念分析の基本的考え方は、与えられた属性の中から独立な属性を見出し、これらの属性によって記述し、結果として対象を分類されることである。記述に必要な属性数および分類数はデータから自動的に決まり、分析者が介入する必要はない。せいぜい数量化されたデータを質的データに変換する段階で、その程度および区分について検討する必要があるだけである。しかし、属性の論理積で分類するため、分類が細かくなりがちである。目的によっては大まかな方が対象の全体像をつかみやすい場合も少なくない。

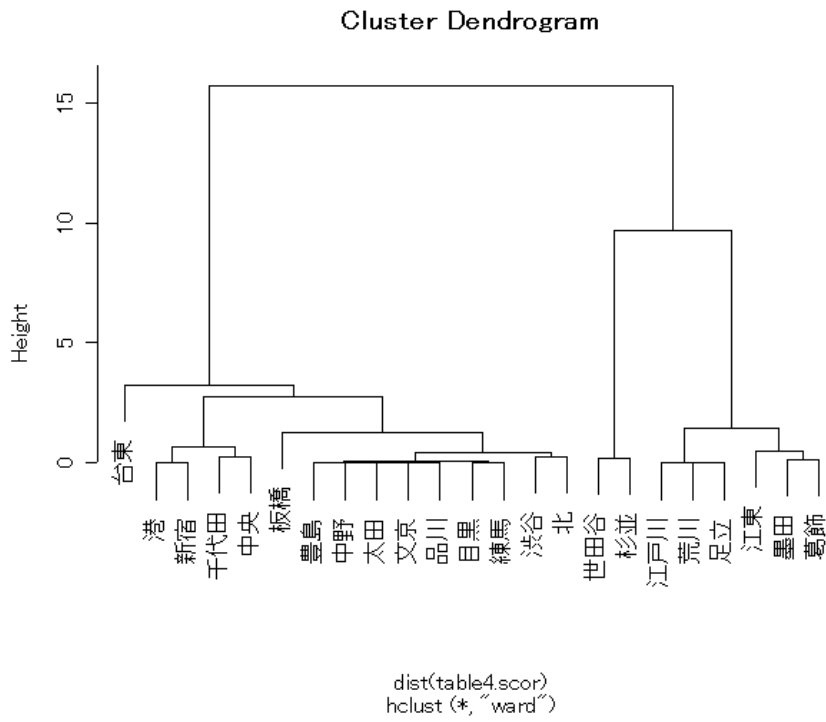
対象をクラスターに分類する代表的手法にクラスター分析がある。しかし、クラスター分析を適用するには、類似度を計る基準や計算手順を減らすための工夫などによってさまざまな手法があり、これらを充分知った上で使う必要がある。また、結果のよみ方にも留意しなければならない。例えば、階層的手法と非階層的手法によって異なるクラスターになることが少なくないが、その場合、さらに何らかの情報を加味することが必要である(上田, 2003, pp.123-129)。

岡太(2002)は、クラスター分析と多次元尺度構成法(MDS)の応用を概観した論文でさらに明確に指摘する。すなわち、「クラスター分析と MDS を、それぞれ独立に、完全データに適用することが望ましい。離散的構造を用いるクラスター分析と、多次元空間という連続的構造を用いる MDS とを、同一データの分析に併用することで、データに潜んでいる情報を相異なる側面からとりだすことができる」。

以上を考慮して、表5、表7に対し対応分析とクラスター分析を試みた。結果の一部を図8、図9に示す。これらの分析結果は対象の特征的傾向をよく表現しているが、直感的に見る限り、クラスター分析に比し対応分析の方がより形式概念分析の結果に近いといえる。これはある意味で当然である。対応分析は、クラスター分析が基本的に近いものどうしを対にしようとする傾向があるのに対し、似たカテゴリどうしほど近づけて表現すると考えられからである。一方、クラスター分析の結果の一部は、形式概念分析でいえば、一段階浅いレベル、すなわち属性の一部を捨象し、クラスターを構成する傾向が見られる。たとえば図8のクラスター、(杉並, 世田谷), (千代田, 中央)は、図7では最終分類より一段上(つまり、浅い)で対象を分類しているといえる。

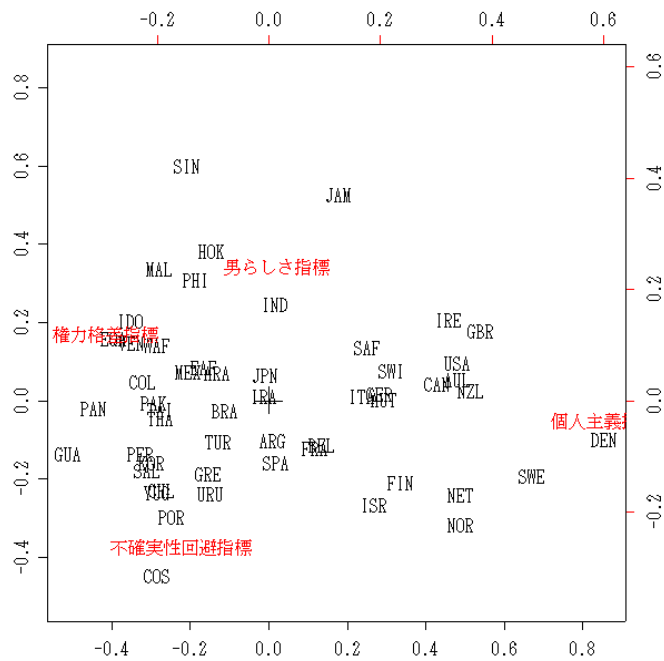


a) 対応分析

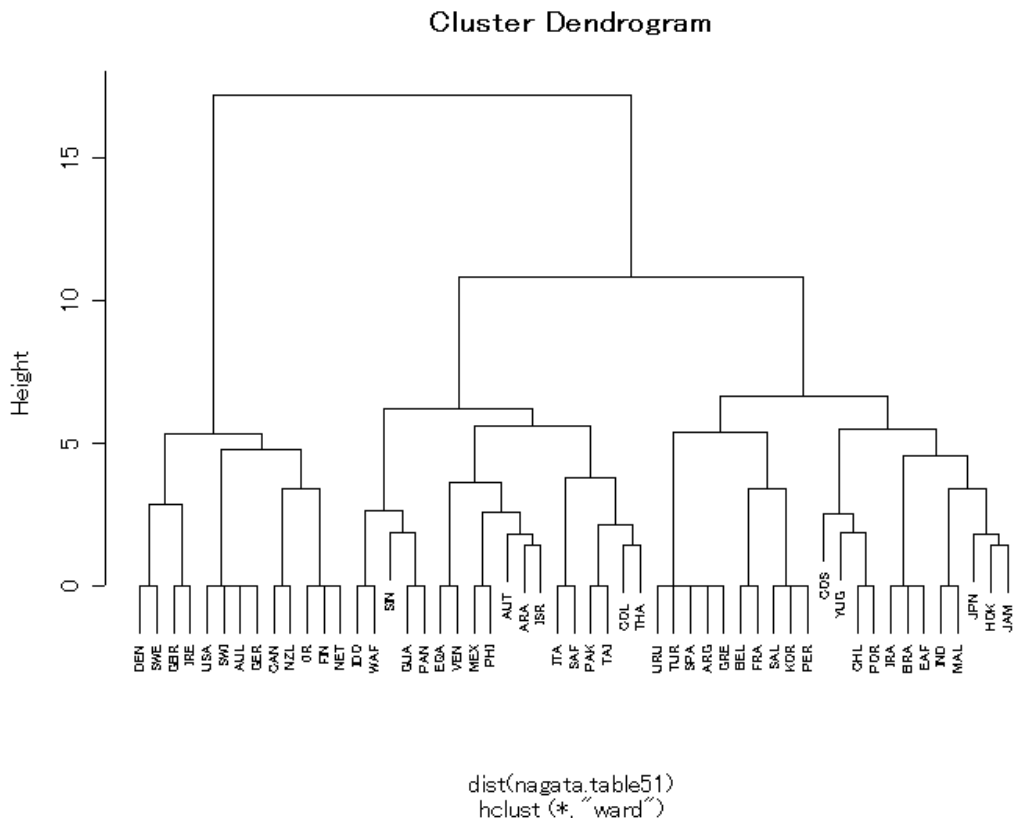


b) クラスタ分析

図8 表5の対応分析とクラスタ分析



a) 対応分析



b) クラスター分析

図9 表7の対応分析とクラスター分析結果

クラスター分析に関する同様の傾向は図9にも現れている。すなわち、表8の形式概念分析で複数の国に分類されたものは、クラスター分析の結果と全く同じであるが、1国に分類されたものはクラスター分析ではクラスターを構成するよう分類される傾向がある。

以上の事実はつぎのように考えることができる。すなわち、多変量解析の結果をクラスターに分類するあるいはクラスター分析の結果を読み、解釈するにはある種の判断が必要となるが、考え方を全く異にする形式概念分析はその判断に有効な示唆を与える。なぜなら、形式概念分析から得られるデータの属性記述は分類の質的な解釈を可能にし、クラスターの意味づけをより強固なものにするからである。他方、属性記述によって対象を分類する形式概念分析では一般に分類が細くなるが、多変量解析のクラスター解析等のさまざまな方法から得られる全体的見通しは、どのレベルでクラスターとするかについて有効な示唆を与えると考えられる。したがって、両者を適切に併用することによって、より具体的な質的比較分析が可能になる。

5.2 ブール代数的アプローチとの関係

Ragin は、統計的計量的方法と質的方法の短所を補い、長所を生かす第3のアプローチ

として質的比較によるブール代数的方法を提案した (Ragin, 1984)。その特徴はつぎのとおりである。(1)社会現象の多様性と因果関係の複雑性を分析できる。(2)論理的で体系的な比較ができる。(3)分析手続きが客観的である。(4)数多くの事例を処理できる。(5)より節約的なモデルを提供する。

形式概念分析は、つぎの点で Ragin のブール代数的方法を包含し、さらにより多様な分析を可能にする方法である。すなわち、(1)ブール代数的方法のように変数値は二値に限らず、任意の名義尺度、順序尺度を扱うことが可能である。(2)社会現象の多様性をそのまま扱うことができる。(3)分析手続きが客観的である。(4)より節約的なモデルを想定しない(結果としてより節約的なモデルが導出される可能性はある)。

しかし、形式概念分析と Ragin の方法の相違は分析目的にある。すなわち、ブール代数的アプローチは対象全体を説明する、より簡潔な因果関係の最小論理式を与えようとするのに対し、形式概念分析は対象を記述するとともに、対象間の比較構造を表現することである。

形式概念分析では深さ 1 のノードの論理式を論理和で結合すれば、すべての対象を記述する論理式を得る。その意味で形式概念分析はブール代数的方法を包含するといえる。しかし、形式概念分析にとってこの論理式は副産物に過ぎず、本来のねらいは対象の記述であり、その結果得られる対象の比較である。

6 おわりに

以上、形式概念を用いて質的データを比較分析する方法を述べるとともに、他の方法との比較を試みた。形式概念分析の特徴を改めて整理するとつぎのとおりである。

- 1) 形式概念分析は、数理的方法であるが、統計的計量的方法ではない。
- 2) 対象の属性を質的データによって記述する。
- 3) 対象の属性を全体として処理し、その過程で属性間の関係を明らかにし、対象記述に必要な属性を特定することができる。
- 4) 対象は属性の論理積によって記述され、その記述にしたがって対象は分類される。
- 5) 形式概念分析によって得られる情報(外延と属性、その深さ、分類結果等)から対象と属性間の全体構造を把握することができる。Hasse 図によって視覚的把握が一層容易になるが、ノード数が多い場合、見通しのよい Haase 図を描くことは容易ではない。

形式概念分析の具体的適用にあたっては以下の点を考慮する必要がある。

- 1) 属性の評価値データが多い場合、統計的計量的処理をしてから形式概念分析を適用するのが適切である。
- 2) 対象データが多い場合、対象の一部に形式概念分析を適用し、属性の何らかの特徴を見出し、その後全体的な分析に移行するのが適切であろう。
- 3) 形式概念分析は、属性の論理積によって対象を記述するため分類が細かくなり、大まかな分類には他の方法との併用が有効であろう。

クラスター分析等との比較から、形式概念分析と統計的計量的方法を併用すると、クラ

スター分けへの示唆、分析結果の解釈あるいは意味づけにとってきわめて有効であることを指摘した。

形式概念分析が、社会科学的データの解析ツールとして広く用いられるためには、形式概念分析を支援する、使いやすいツールを開発するとともに、多くの具体的データに対し適用する必要がある。

謝辞

本稿は「価値判断過程の形式化」の研究から生まれたものである。その過程で日頃、種々議論し、有益なコメントを頂いた田中一氏および高橋徹氏に感謝する。また金明哲氏は原稿を読み、データ解析の方法について有益な示唆と一部計算までお手伝い頂いた。経済学部の谷澤弘毅氏には社会科学のデータに関しご相談にのって頂いた。以上の方々に記して感謝の意を表す。

引用文献

- Davey & Priestley(2002) : *Introduction to Lattices and Order*(2nd Ed.), Cambridge University Press
- 電通総研, 日本リサーチセンター(2004) : 世界 60 カ国価値観データブック, 同友館
- Ganter,B. and R. Wille(1999) : *Formal Concept Analysis*, Springer
- Hofstede,G.(1991=1995) : *Cultures and Organizations:Software of the mind*. McGraw-Hill International (UK) Limited. (岩井紀子・岩井八郎『多文化世界：違いを学び共存への道を探る』有斐閣)
- _____ (1980=1984) : *Culture's Consequences*. SAGE publications. (萬成博・安藤文四郎監訳『経営文化の国際比較：多国籍企業の中の国民性』, 産能出版)
- 鹿又伸夫・野宮大志郎・長谷川計二編著(2001) : 『質的比較分析』, ミネルヴァ書房
- 森典彦・田中英夫・井上勝雄(2004) : 『ラフ集合と感性：データからの知識獲得と感性』海文堂
- 長田博泰(2004) : 「価値判断過程の解析と形式化」, 『社会情報(札幌学院大学社会情報学部紀要)』, Vol.13, No.2: 79-98
- 岡太彬訓(2002) : 「社会学におけるクラスター分析と MDS の応用」, 『理論と方法』 Vol.17, No.2: 167-181
- Ragin,C.C., S.E.Mayer, and K.A.Drass (1984) : *Assessing Discrimination : A Boolean Approach*, *American Sociological Review* Vol.49(April:221-234)
- _____ (1987=1993) : *The Comparative Method:Moving Beyond Qualitative and Quantitative Strategies*. University of California Press.(鹿又伸夫監訳『社会科学における比較研究－質的分析と計量的分析の統合に向けて』 ミネルヴァ書房)
- 上田尚一(2003) : 『質的データの解析－調査情報のよみ方』, 朝倉書店
- Wille, R.(1982) : *Restructuring lattice theory:an approach based on hierarchies of concepts*, *Ordered sets*(ed. I.Rival), Reidel, Dordrecht-Boston, pp.445-470

引用 Web サイト

Freese, R. : <http://www.math.hawaii.edu/~ralpf/LatDraw/>