人文科学とコンピュータ　54－1
（２００２．　５．　３１）

分類における重複性の表現手法
－重複クラスタリング－

小沢　一雅

大阪電気通信大学総合情報学部情報工学科

数理的分類法（クラスター分析法）において、従前考慮されることが少なかった重複型分類について基本的な考え方を述べる。通常の分類とは、個体の集合の切断的分類とよぶべきものであるのに対して、重複型分類は個体の共有をゆるす分類であって、こうした差異に関わる得失についても論ずる。個体の集合について具体的に重複クラスタリングを実行することを想定して、一手法の骨子となる原理を提案する。簡単な例題集合をとりあげて、重複クラスタリングを試行する。

An Overlapping Cluster Scheme

Kazumasa Ozawa
Dept of Engineering Informatics
Osaka Electro-Communication University
Neyagawa, Osaka 572-8530, Japan.

Quantitative taxonomy is discussed in relation to realization over a given set of data items. Almost all of existing cluster schemes have provided non-overlapping partition of a given set of data items. This has sometimes been giving unrealistic views over a real world of data in the humanities. This paper presents an overlapping cluster scheme that would play an important role in realistic understanding of a set of quantified data items. Finally overlapping clustering of an example small set of data items has been carried out using the proposed scheme.

１．まえがき

分類は、混沌とした世界（事物の集まり）に秩序をもたらし、認識と理解を創発する機能をもつ。新しい概念を創出していく過程でも、分類は重要な役割をはたす。まさに思考のルールにおけるもっとも重要な部分をなすといっても過言ではない。

分類に関する研究についていえば、分類のあり方そのものをとりあつかう一般論、すなわち分類学（Taxonomy）よりも、むしろ植物分類学や動物分類学といった具体性をおびた

分類学が先行した歴史的経緯がある[1]。一方、多変量データ解析法の一環として、数理的分類法が各方面に普及している。いわゆるクラスター分析法（クラスタリング）とよばれるもの一般がそれである。こうした手法のいくつかを、数理的分類学（Mathematical taxonomy）として体系化した試みもある[2]。

人文科学においても、分類が重要な役割をはたしてきたことは明らかである。土器を縄文土器と弥生土器に分類したことによって古代の認識と理解が深化し、縄文時代と弥生時代という時代区分へと認識が転化していった事例もある。分類が詳細になるにつれ、さらに深い理解のレベルへと到達しうる可能性が生まれていく。一方、旧来の認識を打ち破る新しい概念の創出にあたっても分類が大きな役割をはたしうる。古い認識の基盤になっている分類をいったん破棄し、新しい分類を導入してみることによってまったくちがった様相がみえてくることがある。こうした場合、新しい分類は必然的に新しい概念の創出をともなうのがふつうである。

分類のほとんどは、世界を切断的に分割する切断型分類である。善か悪か、白か黒か、右か左かといった切断型分類は、世界を単純化して認識するのに有効であって成功例も多いが、一面で誤謬に陥りやすい危険性もある。もし切断的でない分類があるとすれば、それは重複をゆるす分類、つまり重複型分類であろう。とりわけ人文科学においては、本来的に重複型分類が適合する事例が多いように感じられる。問題は、重複型分類を実践するにあたって、その原理や手法があまり周知されていない現状である。すでに重複型分類について言及している文献もあるにはあるが、必ずしも実践的な内容をもつものではなかった[2,3]。本稿では、重複型分類の数理的手法（重複クラスタリング）の実践に向けて分類に関する原理を再整理し、要点をまとめる。

2．分類

2．1　分類の意味

分類の対象となるものは事物の集合である。事物の集合といっても漠然としているので、例をあげれば、人の集合、車の集合、土器の集合などがある。人の集合を日本人やアメリカ人などのように国籍によって分割するのも分類であり、車の集合をセダン、ミニバン、トラックなどの種別に分割するのも分類である。集合をなしている事物の個々を個体とよんでおく。人の集合では、個人たちがみな個体であり、車の集合では１台づつの車両たちがそれぞれ個体である。

個体の集合は、何もしなければただの集まりにすぎないが、何らかの基準（あるいは判断）をもちこんでこれをいくつかの部分集合に分ければ分類をしたことになる。それぞれの部分集合をクラスター（または、クラスあるいはタクソン）とよぶ。分類によってつくられたクラスターが新しい概念に対応することはよくある。人の集合をマルクスは資本家階級と労働者階級という２つのクラスター（概念）に分割し、これを起点にマルクス的世界観を構築したことはよく知られている。

主観的であれ客観的であれ、とにかくある基準のもとで個体の集合が分類された結果生じるクラスターのそれぞれは、共通性をもった個体のグループとみなされる。ここで共通

性とよんでいるもののもっとも単純な形態は、ある属性についての値が同一になるという場合である。マルクスの例でいえば、階級という属性が資本家と労働者という２つの値をもつと規定されていて、同じ属性値をもつ個体たちが１つのクラスターを形成することになる。この例では、分類の基準が階級という属性をもちこむ形で与えられているが、このタイプの分類はじつに多い。動物分類学でいえば、セキツイという属性（値として有・無の２つ）が導入され、その属性値が「有」である個体たちがセキツイ動物というクラスターにまとめられている。

表1　例題集合

個体	属性A	属性B
1	4	10
2	2	7
3	4	5
4	6	8
5	11	7
6	16	8
7	18	10
8	19	7

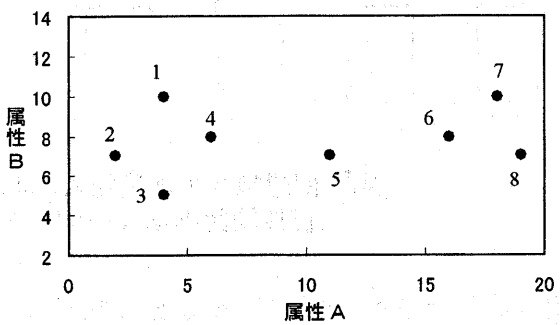

図１　例題集合の散布図

こうした属性値の同一性にもとづく単純な分類が分類のすべてではなく、一般にはもっと多様である。とくに、分析的研究のツールとして普及している数理的分類法（クラスタリング）においては、複数の定量的属性が導入され、属性値（複数の数値の組）から導かれる個体間の距離関係が分類の基準となってきた。すなわち、たがいに「近い」距離関係にある個体たちを１つのクラスターにまとめていくという手法である。こうした定量的手法にあっても、どのような属性を導入するのか、あるいは距離関係をどう規定するのかなど根本的なところには任意性（恣意性）がある。実践の場面では、こうした任意性を最大限に活用してさまざまな試行錯誤をくりかえし、意味のある分類への到達をめざすことになる。表１に８個の個体からなる例題集合を与えている。２つの属性ＡとＢを導入し、それぞれに適当に属性値（数値）を与えている。図１は、例題集合に属する個体たちが２つの属性値を座標とする平面でどのように分布するかを示す散布図である。

２．２　階層性

個体の集合がある分類によって少数のクラスターに分解される一方で、別の分類によれば多数のクラスターに分けられることがある。もし、後者のクラスターのそれぞれがすべて前者のクラスターのどれかに含まれる構造をもっているとき、２つの分類は階層性をもつという。見方を変えると、後者の分類が前者の再分類になっているという構造であって、こうした分類の構造全体を階層分類という。数理的分類法でよく見かけるデンドログラムは階層分類を視覚的に表現する図化法である。図２は、表１の例題集合を単連結法で階層

分類した場合のデンドログラム表示である。

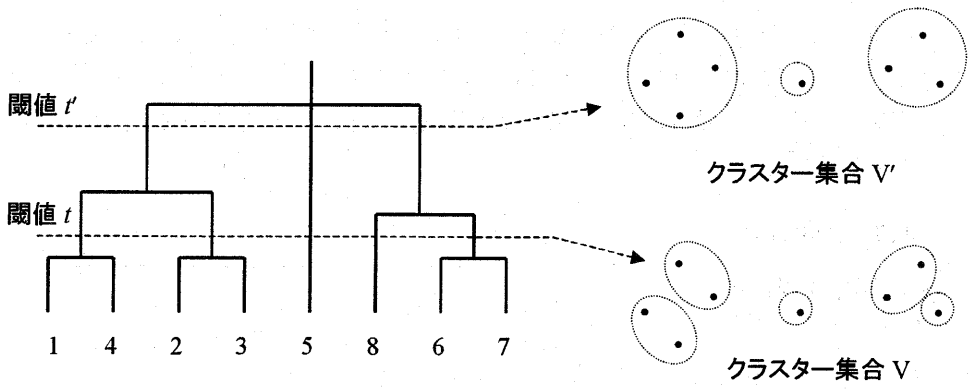

図2 例題集合の単連結法による数理的分類
（個体間距離はユークリッド距離を使用）

階層分類は、大分類、中分類、小分類、細分類などのように多段構造になるのがふつうである。植物分類学や動物分類学など基本的に階層分類とみなされる具体例はきわめて多いが、一般的にいえば分類は必ずしも階層性をともなわなければならないものではない。小分類と大分類との間にいかなる階層性が存在しない分類があったとしても理論的におかしくはない。

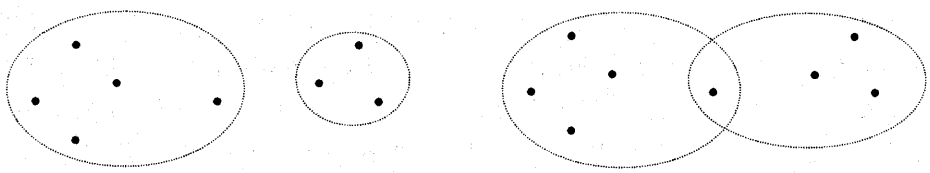

図3 切断型分類と重複型分類

2．3 重複型分類

これまで「分類」「分割」「分解」などの用語を用いてきた。これらの語感からは、複数のちがったクラスターが共通の個体を含まない状態を意味する切断的分類を連想させるかもしれない。本稿では、これらの用語をこうした切断的分類に限定しない意味で用いていることを断っておく。

現実に行われている分類のほとんどは切断的分類であるが、中にはそうでないものもある。たとえば、国籍による人の分類は、二重国籍の人が存在するから明らかに切断的ではない。切断的分類でない分類とは、ちがったクラスターが共通の個体を含むことをゆるす分類、つまり重複型分類である（図３参照）。一方、広く用いられている数理的分類法のす

べてが切断的分類を前提としたものといっても過言ではない。実際、重複型分類が実践的に用いられた事例を筆者は知らない。

個体の集合に関する新しい認識や理解の創発をめざして分類を実践する場合、重複性などの複雑化要因をそれが現実に存在するとしてもすべて無視し、切断的分類を強行することによって単純化された明晰な世界認識に到達できる可能性がある。重複型分類は、明らかにこうした目的には適合しない。重複型分類は、むしろ個体たちの相互関係の実態をできるかぎり忠実にとらえたいときにのみ有効といえよう。前述のように、人文科学においては、他分野に比べて重複型分類が適合する場合が多いように思われる。

3．原理

重複クラスタリングを実践するにあたって、必要となる数理的分類法の原理について切断的分類を含む分類一般とも関連づけながら簡単にまとめる。

3．1 個体間の距離概念

数理的分類法によって個体の集合を分類する場合の出発点は、個体たちのたがいの「近さ」や「遠さ」をはかる尺度として“距離”を定量的に規定することである。いま N 個の個体の集合 $X=\{1,2,\cdots,N\}$ があるとする。個体間距離の全体は、すべての２個体 $i,j\in X$ の距離 $d(i,j)$ を逐次定量的に与えることによって規定される。距離 $d(i,j)$ をどのように与えていくかについては、つぎの２通りの方法に大別される。

【直接法】 ２つの個体 $i,j\in X$ の２対比較によって、距離 $d(i,j)$ を数値として直接与える方法である。与え方は、主観的でもかまわない。和歌の集合や人間の顔の集合を例にとると、２つの個体が似ていると判断すれば、0 に近い数値を与え、似ていないと思われる場合は程度に応じて大きな数値を与えればよい。

【間接法】 適当に導入された属性（複数）に関して、各個体の属性値をまず数値として与える。与え方は主観的であってもかまわないし、計測など機械的操作によってでもかまわない。２つの個体 $i,j\in X$ の距離 $d(i,j)$ は、それぞれの属性値全体の差を適当に数値化すればよい。属性の導入、属性値の決め方、および距離の数値化の方法にはすべて恣意性があり、全体としてみればつねに主観が左右するとみてよい。表１の例題集合が間接法の具体例である。

直接法または間接法のいずれを用いるにせよ、集合 X に属する個体間の距離は２個体の対のすべてについて与えられるから、全体では N^2 個の距離が与えられることになる。こうした個体間距離の全体 $\{d(i,j)\}$ は、$d_{ij}=d(i,j)$ とおいてつぎの個体間距離行列 D として表現される。

$$D=\begin{bmatrix} d_{11} & d_{12} & \vdots & d_{1N} \\ d_{21} & d_{22} & \vdots & d_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ d_{N1} & d_{N2} & \vdots & d_{NN} \end{bmatrix} \qquad (1)$$

距離の基本的性質から、任意の $i,j\in X$ について $d(i,i)=0$ および $d(i,j)=d(j,i)$ が

成り立つように設定されていると、上の行列は対角成分がすべて0である対称行列になる。

3．2　クラスター間距離

個体の集合 X を分類した結果として生成されるクラスターたちは、すべて X の部分集合である。いま、K 個のクラスターが生成されたとし、これらの集合を $V=\{v_1, v_2, \cdots, v_K\}$ と書く。クラスターは X に属する個体の集まりであるから、明らかに $v_1 \cup v_2 \cup \ldots \cup v_K = X$ となる。特別な場合として、単独の個体 $x \in X$ のみのクラスター $\{x\}$ を考えなければならないことがある。これをシングルトンとよぶ。

いま、2つのクラスター $v, v' \in V$ の間の距離を $m(v, v')$ と書き、これをクラスター間距離とよぶ。クラスター間距離は、数理的分類法において中心的な役割をはたすものであるが、これをどのように規定するかについても任意性がある。ただし、ここでは少なくとも任意の $v, v' \in V$ についてつぎの条件がみたされることを仮定する。

$m(v, v)=0$　〈反射律〉　　$m(v, v')=m(v', v)$　〈対称律〉

具体的にクラスター間距離 $m(v, v')$ を与える手続きとしては、個体間距離にもとづく方法が標準的である。もっとも単純なクラスター間距離として知られている単連結法を例にとると、$m(v, v')$ はつぎのように与えられる。

$$m(v,v') = \min_{i \in v-v', i' \in v'-v} d(i,i') \qquad (2)$$

3．3　重複クラスタリング

個体の集合 $X=\{1, 2, \cdots, N\}$ をいくつかのクラスターに分類するもっとも原理的な考え方を要約する[3]。ここでの分類はもちろん重複型分類を含めている。まず、初期クラスター集合として、すべてシングルトンからなるクラスター集合 $V_0=\{\{1\}, \{2\}, \cdots, \{N\}\}$ を分類の出発点とする。初期クラスター集合におけるクラスター間距離は、たとえば式（2）によるにせよ、明らかに個体間距離に一致する。

ここで、適当な閾値 $t(>0)$ について、$m(v, v')<t$ が成立するすべてのクラスター（シングルトン）の組 $v, v' \in V_0$ を検出する。この操作は、$m(v, v')<t$ の意味で、たがいに近い関係にあるシングルトンの組をすべてみつけることである。この結果、V_0 を構成する N 個のシングルトン各々についてたがいに近い関係にあるか否か、つまり隣接関係の全容が明らかになる。そこで、もっとも原則的な考え方は、たがいに隣接しあうシングルトンのグループをもってクラスターとみなすことである。「たがいに隣接しあう」という意味は、そのグループに属するどの2組のシングルトン v, v' をとってみても、$m(v, v')<t$ が成立するということであって、いわゆる強連結な状態をいう。このとき、他とは隣接しないままで孤立するシングルトンも当然ありうる。こうした孤立シングルトンは、それじしんで強連結であり、$m(v, v')<t$ の意味で1つのクラスターとみなされる。かくして、新たに生成されたクラスターの集合を V_1 とすると、V_1 は初期クラスター集合 V_0 に属するクラスター（シングルトン）のクラスター集合といえる。V_1 は一般に重複クラスタリングになっている。

上に述べた V_0 から V_1 をつくりだす操作を V_0 $[t]\rightarrow$ V_1 と書くことにする。いま、さらに

新たな閾値 $t' > t$ を設定して、V_1 上に隣接関係を考えると、新たなクラスター集合 V_2 を生成することができる。すなわち、$V_1[t'] \rightarrow V_2$ である。こうしたクラスター生成操作をつぎつぎと反復する過程は、$t < t' < t'' < \cdots$ なる閾値の系列を適当に与えることによって

$$V_0[t] \rightarrow V_1[t'] \rightarrow V_2[t''] \rightarrow \cdots \qquad (3)$$

のように実現される。この過程は、すべてのクラスターがただ1つのクラスターに統合される段階（最終クラスター集合 V_∞）で終了する。なお、$V_\infty = \{\{1, 2, \cdots, N\}\}$ である。

3．4 階層性と重複性

クラスター間距離にもとづいた上記(3)の過程によって生成されるクラスター集合全体は階層性をもつ。もし、クラスター間距離ではなく、個体間距離によって直接隣接関係を定義する手法を用いる場合にはこうした階層性は保証されない。

一方、上記(3)の過程で生成される1つのクラスター集合 $V_k\ (k = 1, 2, \cdots)$ に属するクラスターたちは一般にたがいに共有の個体をもつ重複クラスターになっている。いま、2つのクラスター $v, v' \in V_k$ を考える。v に属する個体の数を $|v|$ と表すことにすると、2つのクラスター v, v' の重複度 $g(v, v')$ はつぎのように定式化できる[2, 3]。

$$g(v, v') = \frac{|v \cap v'|}{|v \cup v'|} \qquad (4)$$

3．5 例題集合

表1に示される例題集合について、3．3節に述べたクラスター生成操作を具体的に適用した場合、(3) の過程がどのようになるかをしらべてみる。

まず、例題集合の個体間距離をユークリッド距離で測ることにすると、式 (1) に示される個体間距離行列 D はつぎのようになる。

$$D = \begin{bmatrix} 0 & 3.6 & 5.0 & 2.8 & 7.6 & 12.2 & 14.0 & 15.3 \\ 3.6 & 0 & 2.8 & 4.1 & 9.0 & 14.0 & 16.3 & 17.0 \\ 5.0 & 2.8 & 0 & 3.6 & 7.3 & 12.4 & 14.9 & 15.1 \\ 2.8 & 4.1 & 3.6 & 0 & 5.1 & 10.0 & 12.2 & 13.0 \\ 7.6 & 9.0 & 7.3 & 5.1 & 0 & 5.1 & 7.7 & 8.0 \\ 12.2 & 14.0 & 12.4 & 10.0 & 5.1 & 0 & 2.8 & 3.1 \\ 14.0 & 16.3 & 14.9 & 12.2 & 7.7 & 2.8 & 0 & 3.1 \\ 15.3 & 17.0 & 15.1 & 13.0 & 8.0 & 3.1 & 3.1 & 0 \end{bmatrix}$$

D を基礎にしてクラスター間距離を式 (2) に与えた（拡張）単連結法で評価することにする。隣接関係を規定する閾値の系列を、$5.0 < 6.0 < 12.0$ と設定した場合、それぞれの閾値に対応して生成するクラスター集合を図4に示す。

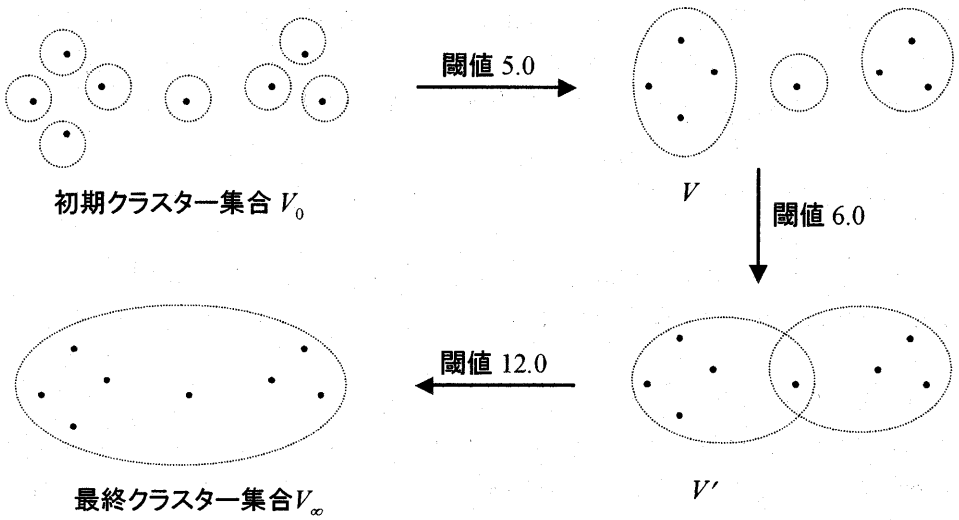

図4　例題集合の重複クラスタリング

図４において、中間段階で生成する２つのクラスター集合 V および V' は、それぞれ具体的な分類を示している。V は切断型分類になっているが、V' は重複型分類である。全体として階層性が保たれていることも確認できる。

４．むすび

本稿では、分類が新しい知見を獲得するための思考のルールとして大きな役割をはたすという視点から、従前の数理的分類法を補完する重複型分類のあり方について概論的な考察を述べた。紙面の都合で詳述できなかった部分も多いが、機会を改めてくわしく論じたいと思う。とくに、人文科学において本来数理的分類法が有効とみられる問題も多いと感じているので、具体的な事例に焦点をあてた考察も必要であろう。一方、本稿で述べた重複クラスタリングは隣接関係における強連結性にもとづいた手法であったが、これを実践するにあたっての問題点についても言及する必要がある。さらに、まったく異なった原理にもとづいて重複クラスタリングを実現する手法についても検討の余地が残っている。重複型分類は、実態に即した世界の認識と理解にとって有効なルールを提供すると考えられるので、人文科学との関連において考察を継続したいと考えている。

【参考文献】

[1]中尾佐助,『分類の発想』, 朝日選書 409, 朝日新聞社, 1990.

[2]Jardin, N. and Sibson, R., *Mathematical Taxonomy*, Wiley, 1977.

[3]Ozawa, K., A Stratificational Overlapping Cluster Scheme, *Pattern Recognition*, Vol.18, Nos. 3/4, 279-286, 1985.