総合距離尺度によるクラスタ分割 (CLAIM 法)

岡田勇二† 加藤常員†† 小沢一雅†††

クラスタリング手法においては，まずクラスタそのものの評価をどのように定義し記述するか，またその評価規準を用いてどのようにクラスタリングを実現するかが重要なポイントとなる．本論文で定義する総合距離尺度とは，データ中の個体間の距離についてクラスタ内外で異なる評価を与え，これらをある重み付けによって総和した最小化関数である．そして，次に本尺度の評価およびクラスタリングの実現手法として，分割最適化型クラスタリング手法の最適化規準に本尺度を採用した新しいクラスタリングアルゴリズム，CLAIM 法を提案する．さらに，人間による視覚的クラスタリングの解が与えられる点パターン（2次元空間のデータ分布）に対して本手法を適用し，その特性と対応関係について述べる．

CLAIM: A New Clustering Algorithm Based on the Integrated Distance Measure

Yuji Okada,† Tsunekazu Kato†† and Kazumasa Ozawa†††

It is important for clustering how to define a cluster and how to evaluate clustering results. In this paper, the integrated distance measure is introduced for optimal partition of a data set, which is defined by the weighted sum of inner and outer cluster distances. A new clustering algorithm, named CLAIM, based on the measure is also presented. CLAIM is experimentally discussed in relation to visual clustering of two-dimensional point patterns.

1. まえがき

パターン認識の分野に限らず，分類操作はわれわれの思考や情報処理の過程において極めて重要な役割を果たしている．分類操作の具体例には，層別する，標識をつける，色分けする，記号をつけるなどさまざまなものがある．こういった一連の操作は，与えられたデータを理解しやすくする，あるいは混沌とした状況を整理して分かりやすいものに変えるという共通の目標がその背景にあると考えられる．いわゆるクラスタリングは，こうした目的を果たすために確定した手続きでデータを分類する操作である．

現在までに，さまざまなクラスタリング手法がコンピュータの進歩とともに改良を加えられながら発展を遂げてきた．過去には実行が困難と考えられてきた手法でさえ，コンピュータパワーの増大によって大量データ（数万から数十万個）への適用が可能になっている．しかし，現実のクラスタリング手法の多くは依然としてデータの縮約的な値（例えば重心などの統計量）に依存した形式をとっている．この場合，厳密に実在値を用いないため，データの持つ本質的な側面を奪うことになりかねない．具体的に言えば，たとえば分散値が等しい場合でも，クラスタの空間的形状が大きく異なることもあり，これが最適なクラスタ分割にとって障害となる事例も，現実には多い．計算速度の点からみると，今日のコンピュータ環境は前述のように次第に制約を緩和しつつあるので，実在値そのものを取り扱う手法に対して再評価の機会が与えられる時にきていると思われる．本稿ではこの立場から，出来る限り実在値に基づいたクラスタリングの評価規準を追求する．

また，こうした手法は，たとえばホップフィールド型ニューラルネットワークや遺伝的アルゴリズムなどの利用を考えるにあたっては，縮約値を用いる手法よりも整合性をもつものと考えられる．

一般論として，クラスタリングは同一のデータについても解（分割結果）は1つではなく，評価規準や利用目的も多様である．実際，クラスタ分析法あるいは数値分類法という呼称で，さまざまなクラスタリング

† グローリー工業株式会社
GLORY Ltd.
†† 大阪電気通信大学短期大学部
Junior College, Osaka Electro-Communication University
††† 大阪電気通信大学工学部
Faculty of Engineering, Osaka Electro-Communication University

が行われているが[1]，植物学や分子生物学など，本来の統計学や統計学データ解析の研究とは異なる別の分野から派生しているものも少なくない[2),3)]．

本稿で提案する CLAIM 法 (Clustering Algorithm Based on the Integrated Distance Measure) では，クラスタリングと人間の視知覚との関連を見いだそうとするところに，1つの目標を定めている．したがって，本手法自体の一般的な応用については今後の課題とし，ここではなるべく人間の主観的な分割結果（視覚的クラスタリング）との比較が容易な形式のデータに限定して考察を行う．

2. 視覚的クラスタリング

一般に，データ集合はベクトル（特徴ベクトル）の集合として定義できる．本稿では，とくに2次元ベクトルとして与えられるデータ集合を，**点パターン**と呼ぶことにする[4]．**図1**は点パターンの範例である．

このうち，例えば図1(a)の点パターンにはいくつかの「かたまり」(クラスタ) が認識される．一方，図1(b)にみられるような一様分布の点パターンでは，このような「かたまり」は認められない．われわれの視覚によるクラスタリングでは，図1(a)のようなクラスタ的 (clustered) な点パターンについては，いくつかに，どのように，分割すべきかはほぼ一意的に決まってくる．すなわち，この場合，分割数は一般に3，かつその数のみであり，分割のされ方も明瞭である．このように，いくつかの「かたまり」としてのクラスタが，視覚的に判然と認められるような場合に注目すれば，比較的対応も容易である．ここでは，こういった視覚的クラスタリングを基軸に考察を行っていく．

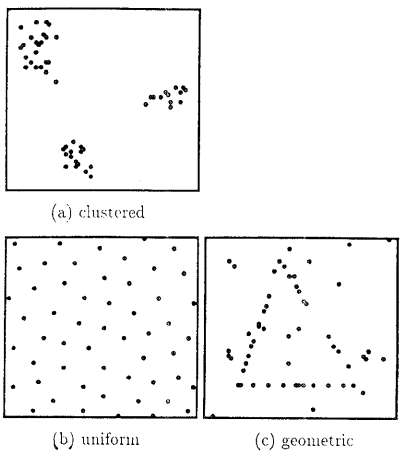


図1 点パターンの実例
Fig. 1 Examples of point patterns.

一方，データの構造がクラスタ的でない場合には，クラスタリングを行うか否かが，それ以前の問題として存在する．図1(b)のような点パターンを自動的クラスタリングの対象とした場合，何らかの分割結果は得られるであろうが，われわれ人間でさえ分割数を特定できないようなデータ構造では解の適否を評価することが難しい．また，図1(c)に示されるような，幾何学的な構造を含む点パターンに対しても同様のことが言える．この場合，点パターン中に潜んだ直線性を検出することに関心があるとすれば，クラスタリングとは異なる手法，例えば Hough 変換[5]のような方法を用いるほうが適している．

以上のことから，本稿では人間が視覚的にみて分割数や分割様式が一意に求まるような（視覚的クラスタリングが可能な）点パターンを用い，そうした解を分割結果として自動生成するクラスタリング手法の構築を目標とする．

3. CLAIM 法

3.1 分割最適化型のクラスタリング手法

既に述べたように，クラスタリング手法は統計的データ解析をはじめとして多種多様な分野で扱われており，その実現方法にもさまざまなアプローチがある．しかし，アルゴリズム的な観点から見れば，それらは大別して階層的な分類手法と非階層的な分類手法の2つに分けることができる．分割最適化型クラスタリング手法は，後者の代表的な手法である．

この手法では，クラスタのまとまりの良さ，あるいはクラスタ間の流れ具合を，何らかの規準で測るようにして，これを最小化（あるいは最大化）する方針で分割を行う．ここに見られる一連のアルゴリズムの流れを簡単に記述すると以下のようになる．

（1）初期配置または初期分割の生成
（2）初期配置に対する最適化基準の算出
（3）再配置および最適化規準の更新・反復
（4）収束の判定と要約

k-means 法[6]を筆頭として，多くの分割最適化型に属する手法がある．本稿で提案するクラスタリング手法も，総合距離尺度という最適化規準を定め，これを最小にする分割結果を求めることから，明らかに分割最適化型の手法に属する．

3.2 CLAIM 法における最適化規準

本稿では個体間距離の差異性に基づいた，ある種の「かたまり」を1つのクラスタとして取り扱う．そのためにまず，クラスタ内の個体同士の距離尺度，およびクラスタ外の距離尺度を定め，さらにこの2つの距離尺度で構成した最適化規準，すなわち総合距離尺度を定義する．

3.2.1 クラスタ内距離とクラスタ外距離

まず，個体 i-j 間のユークリッド距離を d_{ij} とし，次の2つの距離尺度を導入する．ただし，この d_{ij} の値はデータのサイズ等に影響されないように何らかの方法で正規化しておくことが望ましい．ここでは，個体間の最大距離 $d_{\max}$ を基準として正規化を行っている．つまり，$d_{\max}=1$，または $0 \leqq d_{ij} \leqq 1$ である．

● クラスタ内距離和

$$D_{\mathrm{I}}=\sum_{k=1}^{g} D_{\mathrm{I}}^{k}=\sum_{k=1}^{g}\sum_{i\in C_k}\sum_{j\in C_k} d_{ij} \qquad (1)$$

ただし，C_k は k 番目のクラスタ（個体集合）を表す．

これは，同じクラスタに属する個体同士の距離を合計したものである．クラスタ外に存在する個体については考慮しない．したがって，分割するクラスタの数が多くなるほど，1つ1つのクラスタに属している個体の絶対数は少なくなり，距離の合計値である D_{I} の値は小さく評価されることになる．

● クラスタ外距離和

$$D_{\mathrm{O}}=\sum_{k=1}^{g} D_{\mathrm{O}}^{k}=\sum_{k=1}^{g}\sum_{i\in C_k}\sum_{j\notin C_k} (d_{\max}-d_{ij}) \qquad (2)$$

クラスタ外距離和とは，あるクラスタに属する個体と，そのクラスタ外にある個体との距離を，最大距離 $d_{\max}$ から差し引いたものの合計値である．つまり，この場合はクラスタ内距離和とは逆に，分割数が少なくなるほど，クラスタ外の個体数が減少し（同一クラスタ中の個体数は増加するがここでは評価の対象とならないため），値が小さく評価される．

クラスタ外距離を $(d_{\max}-d_{ij})$ と定義したのは，本来，この距離合計値が最大化されるべきものであることに関与している．ここでの基本方針として，個体間距離の短いもの同士が優先的に1つのクラスタに収まるように扱う．その意味で，ある個体からみて，クラスタ外に存在する個体というのは，なるべく距離の値の大きなもののほうが望ましい．ところが，D_{I} が最小化を目指しているのに対して D_{O} が最大化されるのでは，後に述べる総合距離尺度のうえで整合性が悪い．こういった理由から，クラスタ外距離には基準値である $d_{\max}$ からの差をとるように補正を行っている．

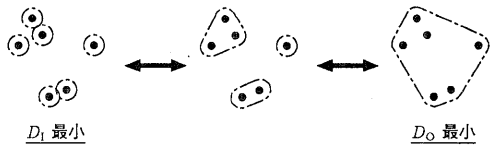


図 2 クラスタ内距離和とクラスタ外距離和の関係
Fig. 2 Relation between inner and outer distances.

3.2.2 総合距離尺度

分割最適化型の手法では，定めた最適化規準を最小（あるいは最大）にするような分割パターンがクラスタリング結果として求められる．ここでは最小化の方針をとるので，例えば距離尺度 D_{I} のみをその規準としてクラスタリングを行った場合，結果として選ばれるのは，クラスタ数の最も多くなるような分割パターン，つまり個本数分のシングルトン（孤立クラスタ）が存在するパターンである．逆に D_{O} のような規準を用いれば，データ空間内の個体すべてが1つのクラスタに属するような分割パターンが選ばれる（**図2**参照）．

そこで，次式のように内外の距離尺度の総和をある重み付けによって（仮にパラメータ α とする）合計し，これら両極端となる分割結果の，いわば中間的な分割パターンに対して最小値を定めようとしたのが本尺度である．

$$D=\alpha D_{\mathrm{I}}+(1.0-\alpha)D_{\mathrm{O}} \quad (0.0\leqq\alpha\leqq 1.0) \qquad (3)$$

$$=\alpha\sum_{k=1}^{g}\sum_{i\in C_k}\sum_{j\in C_k} d_{ij}+(1.0-\alpha)\sum_{k=1}^{g}\sum_{i\in C_k}\sum_{j\notin C_k}(d_{\max}-d_{ij})$$

$$=\sum_{k=1}^{g}\sum_{i\in C_k}\left\{\alpha\sum_{j\in C_k} d_{ij}+(1.0-\alpha)\sum_{j\notin C_k}(d_{\max}-d_{ij})\right\} \qquad (4)$$

既に2章で論じたように，人間の視覚による点パターンの分割は，与えられたデータに潜在する点集合の「融合」と「分離」に対する微妙なバランス感覚が，その根底にあると思われる．この観点からすれば，クラスタ内距離和 D_{I} は分離の度合を，クラスタ外距離和 D_{O} は融合の度合を示す尺度であり，その両者の度合を重み α でバランスをとったものが総合距離尺度 D であると解釈できる．

3.2.3 例　　題

図3のような距離関係を持つ，個体数が3個の点パターンを用いて，実際に総合距離尺度を計算してみる．

このデータに対して，パラメータ α を 0.0 から 1.0 まで 0.1 ずつ増加させていった場合に総合距離尺度がとり得る値を算出し，その最小値に対応する分割パターンを調べたのが**表1**である．

このうち，網掛けの表記部分が，与えたパラメータ値 α に対する総合距離尺度の最小値である．例えば，$\alpha=0.0$ のときには分割パターン1が，その最小値をとることが分かる．そこで，最小値をとるパターンの変遷をまとめると，次の**図4**のようになる．

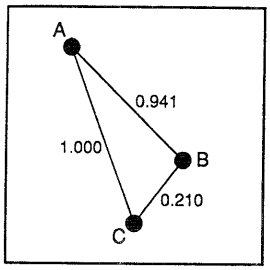
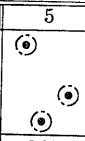


図 3 個体数3の点パターンの例
Fig. 3 An example of point pattern consisting of three items.

表 1 総合距離尺度および分割パターン
Table 1 The integrated distance measure and clustering results.

分割数		1	2			3
No.		1	2	3	4	5
分割パターン						
群内距離和		4.302	1.881	0.421	2.000	0.000
群外距離和		0.000	1.579	0.119	1.698	1.698
総合距離尺度	$\alpha=0.0$	0.000	1.579	0.119	1.698	1.698
	0.1	0.430	1.610	0.149	1.728	1.528
	0.2	0.860	1.640	0.179	1.758	1.358
	0.3	1.291	1.670	0.209	1.789	1.189
	0.4	1.721	1.700	0.239	1.819	1.019
	0.5	2.151	1.730	0.270	1.849	0.849
	0.6	2.581	1.761	0.300	1.879	0.679
	0.7	3.011	1.791	0.330	1.909	0.509
	0.8	3.442	1.821	0.360	1.940	0.340
	0.9	3.872	1.851	0.390	1.970	0.170
	1.0	4.302	1.881	0.421	2.000	0.000

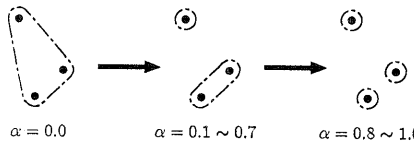


図 4 最小値をとる分割パターンの変遷
Fig. 4 Clustering results changed by parameter α.

このように，狙いどおりパラメータ α を変化させていくことによって，最小値をとる分割数および分割パターンを変化させられることが分かる．しかも，視覚的クラスタリングとしての立場からみると，2分割に2や4のパターンを選ぶことはまずあり得ない．

このことは，すなわち総合距離尺度に適切なパラメータ α を与えることができれば，すべての分割パターン（分割数の異なるものも含む）のうちの特定のパターン，それも視覚的クラスタリングとしても違和感のない分割パターンに対して最小値を定められる可能性を示唆している．

3.3 シミュレーションアルゴリズム

次に，本尺度の有効性を評価するために用意した算法について述べる．基本的な流れはごく一般的な分割最適化型のクラスタリング手法を踏襲しており，比較的簡素なアルゴリズムとなっている．

——シミュレーションアルゴリズム——

Step 1　与えられたデータ（2変量データ）の個体座標に基づいて個体間距離を計算したあと，最大距離を基準に正規化を行う．

Step 2　全個体（仮に n 個とする）を g 個のクラスタへと初期分割する（$1 \leqq g \leqq n$）．（初期分割は分割最適化型手法にとって大変重要な要素であり，一般的な手法ではこれをユーザが与えることになっている．しかし，本手法では今のところ分割数を特定しないため，とりあえず $g=1$，つまり全個体を1つのクラスタとして計算を始めることにしている．）

Step 3　初期分割に対する総合距離尺度 D を求める．

$$D=\sum_{k=1}^{g}\sum_{i\in C_k}\left\{\alpha\sum_{j\in C_k}d_{ij}+(1.0-\alpha)\sum_{j\notin C_k}(d_{\max}-d_{ij})\right\}$$

（ここで任意の個体 p がクラスタ C_k に属しているときの，総合距離尺度に関与している部分が，

$$\alpha \sum_{i\in C_k}(d_{pi}+d_{ip}) +(1.0-\alpha)\sum_{i\notin C_k}\{(d_{\max}-d_{pi}) +(d_{\max}-d_{ip})\} \quad (5)$$

であり，さらに $d_{pi}=d_{ip}$ の関係から，

$$2\left\{\alpha \sum_{i\in C_k} d_{pi}+(1.0-\alpha)\sum_{i\notin C_k}(d_{\max}-d_{pi})\right\} \quad (6)$$

となることに着目する.）

Step 4 移動させる候補の個体 p を選ぶ.
（この選出方法も，単純に個体番号の小さなものから順に選んでいる．なお，番号はデータの読み込み時に割り振ったもので，特別な意味合いは持たせていない.）

Step 5 Step 4 で選びだした個体 p を，クラスタ C_k から他のクラスタ C_l に移動させると仮定して，その分割結果に対して，新たに個体 p が担うことになる総合距離尺度の部分計算をする.

$$2\left\{\alpha \sum_{i\in C_l} d_{pi}+(1.0-\alpha)\sum_{i\notin C_l}(d_{\max}-d_{pi})\right\} \quad (7)$$

Step 6 Step 5 の計算を移動先の候補となるすべてのクラスタ $(l=1,2,\cdots,g+1, l\neq k)$ に対して行ったあと，

$$\alpha \sum_{i\in C_k} d_{pi}+(1.0-\alpha)\sum_{i\notin C_k}(d_{\max}-d_{pi}) > \alpha \sum_{i\in C_l} d_{pi}+(1.0-\alpha)\sum_{i\notin C_l}(d_{\max}-d_{pi}) \quad (8)$$

の関係が成り立つクラスタ C_l を捜し，この C_l を個体 p の移動先のクラスタとする．ただし，候補のクラスタが複数個あるならば，その中で最も右辺が小さくなるようなクラスタを選ぶ．さらにこの関係が見つからない場合は，個体が現在属しているクラスタにとどめておく．（ここで，移動先の候補となるクラスタのうち，C_{g+1} は「空のクラスタ」を意味する．本アルゴリズムにおいては，分割数の増減を許しており，もし個体が $g+1$ 番目のクラスタに移動するならば，分割数は1つ増加することになる．同様に分割数の減少も考慮するために，候補個体がシン

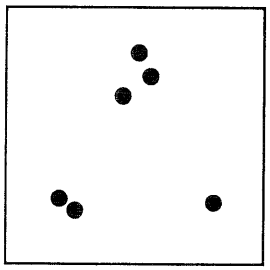


図 5 実験用データ（パターン1）
Fig. 5 Experimental point pattern (pattern 1).

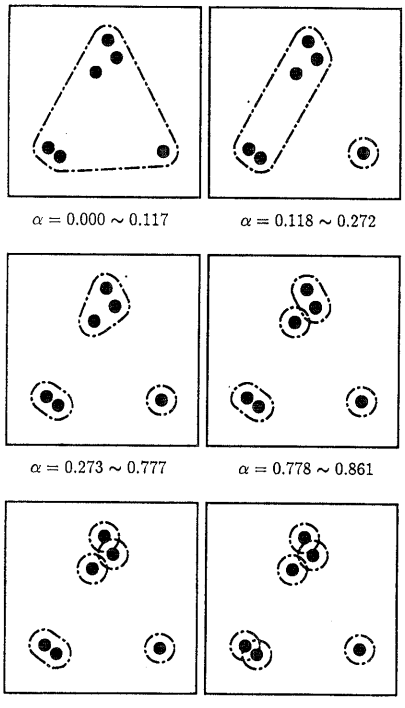


図 6 パターン1についての分割パターン
Fig. 6 Clustering results for pattern 1.

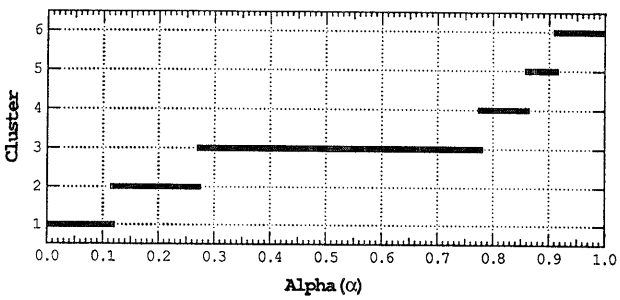


図 7 パラメータ値と分割数
Fig. 7 Number of clusters (pattern 1).

グルトンを形成している場合でもこの演算を行う.)

Step 7 個体pを移動先のクラスタC_lに移し，総合距離尺度Dを更新する.

$$D'=D-2\left\{\alpha\sum_{i\in C_k}d_{pi}+(1.0-\alpha)\sum_{i\notin C_k}(d_{\max}-d_{pi})\right\}+2\left\{\alpha\sum_{i\in C_l}d_{pi}+(1.0-\alpha)\sum_{i\notin C_l}(d_{\max}-d_{pi})\right\} \quad (9)$$

分割数の変化があれば，gの値も更新しておく.

Step 8 Dの値が何回かの繰り返しで変化が見られなくなったら計算を終了させる（それ以外の時は手順 Step 4 へ戻る).

4. 実験および考察

評価実験のために，まずは**図5**の点パターンを用いる．個体数は6で，認識できるクラスタの数は3である.

この点パターンについて，パラメータαを0.000から1.000まで0.001刻みで与え，計1001回のクラスタリング実験を行ってみた．**図6**はこの実験で得られたすべての分割結果，および**図7**はパラメータαと分割数との対応をまとめたグラフである.

分割パターンは全部で6通りのものが得られ，図にはそれぞれの分割をするときのαの範囲を付与している．また，図7のグラフの横軸はαの値，縦軸はそのときの分割数を示している.

各分割数には第2種のスターリング数，$P(n, g)$通りの分割パターンが存在し，全体では$P(6,1)+P(6,2)+\cdots+P(6,6)=203$通りの分割通りが考えられる.

$$P(n,g)=\frac{1}{g!}\sum_{i=1}^{g}\binom{g}{i}(-1)^{g-i}\cdot i^n \quad (10)$$

しかし，ここでは個体数が比較的少ない数なので，局所的なパターンへの収束は見られない．与えたパラメータαに対する総合距離

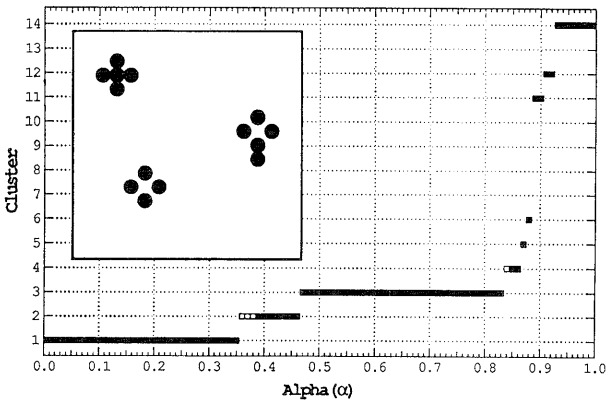


図 8(a) パターン2と分割数
Fig. 8(a) Pattern 2 and number of clusters.

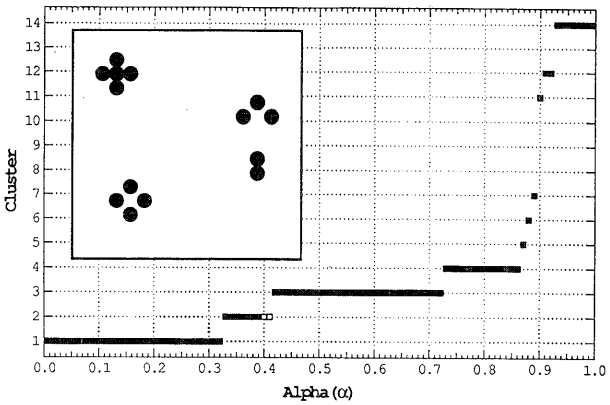


図 8(b) パターン3と分割数
Fig. 8(b) Pattern 3 and number of clusters.

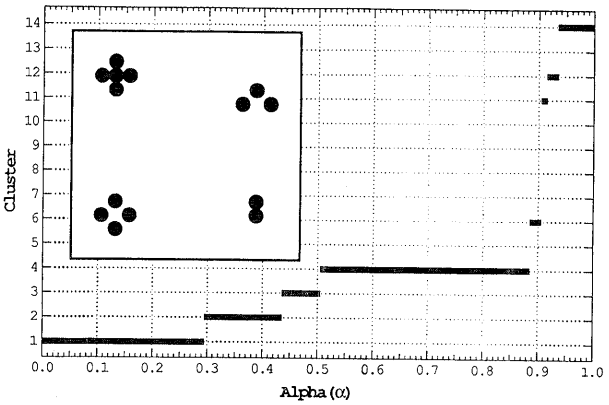


図 8(c) パターン4と分割数
Fig. 8(c) Pattern 4 and number of clusters.

尺度の最小値をもつ，各分割数あたり1つだけの分割パターンが得られている．これらはおそらく，分割数を限定した条件下では最も視覚的に望ましいものであると思われる．

さらに，この6種の分割パターンのうちで，どれが最も望ましい分割数のパターンであり，なおかつその結果が得られるパラメータαはどの値であるかを考える．先にも述べたとおり，本点パターンで認識される分割数は3である．図7のグラフでは，与えたパラメータに関して，分割数は一定の変化をとるのではなく，それぞれの分割数によって異なる幅（αの範囲）を持つことが確かめられる．この場合，広範囲のαに対して分割数3のクラスタリング結果が得られている．

本実験だけでは断定できないが，最適な分割パターンに対するαは，他の分割パターンのときよりも広い範囲を保つ傾向がある．そうなれば，最適なクラスタリング結果が得られるαは，この広い範囲のいずれかの値であればよく，確率的にもそれを指定できる可能性が高くなっていることが言えよう．

次に**図8**（a）から図8（c）までの3つの点パターンを用いてαのとる幅の変化を調べてみる．これはクラスタの位置を少しずつ変えて配置した点パターンに対する実験である．各クラスタは同じ個体数で構成されているため，互いに似かよってはいるものの，点パターン全体として視覚的に認識される分割数は異なる．

得られた分割パターンは図8（a）で11通り（2分割と4分割が各々2通り），図8（b）では10通り（これも2分割が2通り），そして図8（c）が8通りであった．

図8（a）と図8（c）に関しては，視覚的に認められるクラスタ数もそれぞれ3分割と4分割のようにはっきりしており，これは本実験においてパラメータαが最も広い範囲をとる分割パターンと一致している．図8（b）については，3分割か4分割か意見の分かれるところであろう．この場合，興味深いことに3分割となる範囲が広いものの，4分割をとる範囲も図8（a）に比べて広くなってきている．

ただし，常に分割数1となるパラメータαの範囲が広いことに注意したい．これには，例えば次のような理由が考えられる．

- 標本空間全体に分布している個体の数が多くなるほど，空間内の密度が高くなるため，全体を1つのクラスタと見なしやすくなる．また，点パターンがあまりクラスタ的でなくなる場合，すなわち一様分布に近くなる場合にも1分割の範囲は広くなる．これはデータの構造に起因する問題である．
- 分割手法において，初期分割数を1で開始しているため，αの値が小さいほど1分割のままで状態遷移計算が終わってしまう可能性が高く，いわゆる全分割パターン中のグローバルミニマに到達できない．実際，与えられた点パターンに対して，配置個体それぞれがシングルトンである場合から状態遷移を開始すると，この範囲は少し変化する．こちらは初期値選択問題や最小値探索問題に関わる，アルゴリズム上の問題である．

他にもまだ別の要因が存在するかもしれないが，いずれにしても，これに関しては今後さらに検討を行う余地が残されている．

この実験ではまた，分割数すべてが選ばれるのではないことも分かる．個体数が多くなったため，14個体の場合，分割パターンは全部で約2億通り，分割数だけでも14通りが考えられるが，5分割から13分割までの間は，分割できるαの値が非常に限られているか，もしくはまったく分割しないかのどちらかである．この実験ではαを0.001刻みで変化させたため，得られなかった分割数も，もしかするとそれよりも微少なパラメータ値の変化によって分割可能なのかもしれないが，他の3分割や4分割の得られるαの範囲に比べると非常に狭い範囲であることに変わりはない．

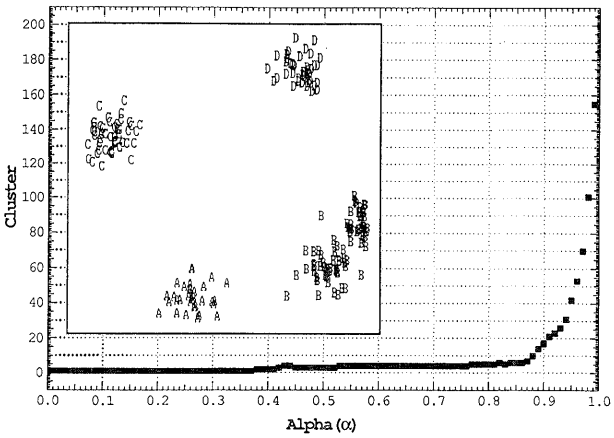


図9 パターン5と分割数
Fig. 9 Pattern 5 and number of clusters.

さらに個体数を増やしてみるとどうなのであろうか．**図9**は確率モデルである Neyman-Scott process[7),8)] にしたがって生成した200個体，4クラスタの点パターンの一例である．

分割結果は，分割数3や5の近辺でローカルミニマに陥っている兆しは見られるものの，1分割をのぞき，4分割となるαの範囲がかなり広いことが確かめられる．個体数が増加すると，さらに選ばれることのない分割数が多くなり，分割可能性の高いものだけが結果として現れてくる．図はその分割数4の場合に見られた分割パターンを示しており，クラスタごとに異なる記号で個体を表記している．

このように実験結果を見る限りでは，比較的広いαの範囲を持つ分割パターンが，人間の「主観的な」クラスタリング結果に対応していると考えられる．

ただ，実験に本手法を適用するにあたって，パラメータαを徐々に変化させて，しらみつぶしに分割パターンの安定域を調べるのでは，実用性に問題が残る．今後，何らかの方法でもって適切なαのパラメータ値の決定方法を獲得しなければならない．残念ながら現段階では，この方法はまだ確立していない．しかし，データ構造の簡単な予備パターン解析の導入や，パラメータと分割数との相関をさらに調べることによって，妥当なクラスタリング結果が得られる可能性は高い．

5. むすび

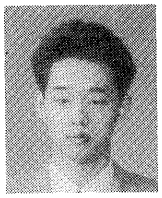
本稿では，視覚の融合と分離の概念に基づく，総合距離尺度というクラスタ評価規準を定め，それに基づいた分割最適化型の新しいクラスタリング手法を提案した．また，点パターンを用いた実験によって，その有効性についても検討を行った．

クラスタリングは組合せ最適化問題であり，計算の複雑性を考慮するとすべての可能な組合せの中から最適解を得るのは非常に難しい．ここで述べられたクラスタリングアルゴリズムはごく基本的なものであり，個体数の非常に多いデータを扱うためにも，理論の整理や高速化を検討する必要性がある．ただし，冒頭でも述べたとおり，筆者らはニューラルネットワーク[9),10)]や遺伝的アルゴリズム[11)]を応用したこの種の問題解決手法も検討中である．この場合，総合距離尺度のような，簡素な計算式の方が整合性が高いと考えている．

参考文献

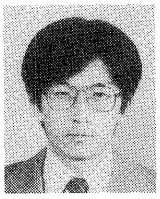
1) Jardine, N. and Sibson, R.: *Mathematical Taxonomy*, Wiley (1977).
2) Ozawa, K.: CLASSIC: A Hierarchical Clustering Algorithm Based on Asymmetric Similarities, *Pattern Recogn.*, Vol. 16, No. 2, pp. 201-211 (1983).
3) Ozawa, K.: A Stratificational Overlapping Cluster Scheme, *Pattern Recogn.*, Vol. 18, Nos. 3/4, pp. 279-286 (1985).
4) Hoffman, R.L. and Jain, A.K.: Sparse Decompositions for Exploratory Pattern Analysis, *IEEE Trans. Pattern Analysis Machine Intelligence*, Vol. PAMI-9, No. 4, pp. 551-560 (1987).
5) Duda, R.O. and Hart, P.E.: Use of the Hough Transformation to Detect Lines and Curves in Pictures, *Comm. ACM*, Vol. 15, No. 1, pp. 11-15 (1972).
6) Jain, A.K. and Dubes, R.C.: *Algorithms for Clustering Data*, Prentice-Hall, Englewood Cliffs (1988).
7) Smith, S.P. and Jain, A.K.: Testing for Uniformity in Multidimensional Data, *IEEE Trans. Pattern Analysis Machine Intelligence*, Vol. PAMI-6, No. 1, pp. 73-81 (1984).
8) Diggle, P.J.: On Parameter Estimation and Goodness-of-fit Testing for Spatial Point Patterns, *Biometrics*, Vol. 35, pp. 87-101 (1979).
9) 岡田，加藤，小沢：ホップフィールドモデルによるクラスター分析についての一考察，第43回情報処理学会全国大会論文集，分冊1，1B-9 (1991).
10) 岡田，加藤，小沢：総合距離尺度によるクラスター分割（CLAIM 法），電子情報通信学会技術研究報告，PRU 92-106, pp. 29-36 (1993).
11) 加藤，小沢：遺伝的アルゴリズムによるクラスター分割，第47回情報処理学会全国大会論文集，分冊2，6N-8 (1993).
12) 楠原，辰巳：誘引関係に基づく離散型クラスタリングシステム，電子情報通信学会技術研究報告，PRU 88-121, pp. 17-24 (1988).
13) 大隅：統計的データ解析とソフトウェア，日本放送出版協会 (1989).

（平成5年7月14日受付）
（平成6年9月 6 日採録）



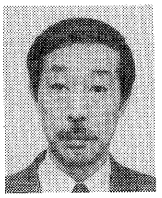
岡田 勇二（正会員）

昭和44年2月生．平成3年大阪電気通信大学工学部経営工学科卒業．平成5年同大学大学院情報工学専攻博士前期課程修了．工学修士．現在，グローリー工業(株)第2金融機器事業部勤務．パターン認識，ニューラルネットワーク等に興味をもつ．



加藤 常員（正会員）

昭和33年生．昭和57年大阪電気通信大学工学部経営工学科卒業．昭和57～59年ミネベア(株)勤務．平成元年岡山理科大学大学院理学研究科博士課程修了．理学博士．昭和63～平成2年日本学術振興会特別研究員．平成2年大阪電気通信大学短期大学部講師，現在に至る．情報処理技術の考古学への応用研究に従事．



小沢 一雅（正会員）

昭和17年生．昭和41年大阪大学基礎工学部電気工学科卒業．昭和47年同大学院博士課程修了．工学博士．同年大阪電気通信大学工学部講師．昭和54年同教授．レーザOCRの研究を経て，パターン認識，コンピュータ考古学の研究に従事．電子情報通信学会，IEEE，英国BMVAおよびCAA各会員．著書「情報理論の基礎」(国民科学社)，「数理考古学入門」(共訳；雄山閣)，「前方後円墳の数理」(雄山閣)，「考古学における層位学入門」(訳；雄山閣，近刊)，「パターン情報数学」(森北出版，近刊)．