遺伝的アルゴリズムを用いた非階層的クラスタリング

加藤常員[†] 小沢一雅[††]

分類操作は，大きく階層的分類と非階層的分類に分けられ，それぞれに多くの手法（クラスタリング手法）が提案されている．本論文では，多値の遺伝子表現を採用した遺伝的アルゴリズムに基づく非階層的クラスタリング手法を提案する．一般にクラスタリング手法は，クラスタの評価基準および分類の具体的な手順を定めるものである．非階層的クラスタリング手法では，データを k 個のクラスタに分割する際，分割数 k は与えられるか，手法の一部として決定される．非階層的クラスタリング手法の主たる戦略として，まず仮のクラスタ分割を与え，評価基準がよりましな分割に逐次改良していく方法（分割最適化法）がある．この方法では，初期の分割あるいはクラスタの核といった初期条件が最終の分割結果に根本的な影響を及ぼし，ときとして局所最適解（分割）に陥る場合も多い．一方，遺伝的アルゴリズムは，局所最適解を回避することができる多点探索の特長を持っている．本研究は，遺伝的アルゴリズムの多点探索能力をいかし，ロバストな非階層的クラスタリングの実現をめざすものである．提案する手法は，優性遺伝をモデルとした新たな遺伝的アルゴリズムによって構成されている．本手法の特性を確認するため，分割最適化法の代表的な手法である k-means 法に基づく比較実験を行った．実験では，2変量データ（点配置パターン）を用い，評価基準は平方和分解の原理に基づくクラスタ内平方和の総和を採用した．実験結果として，在来の手法では数%しか最適な分割を得ることができない対象に対して，本稿で提案する手法によれば，80%に近い割合で最適な分割が得られることを確認した．

Non-hierarchical Clustering by a Genetic Algorithm

TSUNEKAZU KATO[†] and KAZUMASA OZAWA[††]

This paper treats a clustering procedure by using the genetic algorithm. As is well-known, existing clustering procedures are classified into two types; i.e., hierarchical and non-hierarchical types. In this paper, a non-hierarchical clustering procedure based on the genetic algorithm has been presented. We call it *GA procedure*. Our GA procedure can be characterized by its strong power of the multi-point search. From this, the probability to obtain the optimum solution can be made higher than existing procedures. In our experiment, the GA procedure and a typical existing procedure for the so-called k-means method have been compared in terms of probability to obtain the optimum solution and other related aspects. For three kinds of experimental point patterns, the GA procedure has shown its robustness in searching the optimum solution.

1. はじめに

クラスタリングは，データとして与えられる個体の集合を階層的に分類する階層的クラスタリングと特定の分割数へと一気に分割する非階層的クラスタリングに大別される．非階層的クラスタリング手法においては，まずクラスタの評価基準を設定する．次に，仮のクラスタ分割を与え，クラスタの評価基準を用いてよりましなクラスタ分割に改良する方法（分割最適化法）が主流である．分割最適化法には，k-means 法，ファジィ k-means 法，ISODATA 法，ダイナミック・クラスタリングなどの手法[1]が提案されている．これらは，それぞれコンピュータの発展とともに改良され，大量のデータを高速に分割することが可能となっている．しかしながら，n 個の対象を k 群に分割する分け方（場合の数）は，第2種のスターリング数[2]となり，そのすべてについて評価値を計算し，最適解を求めることは事実上できない．すなわち，最終的に得られた分割結果がクラスタの評価基準に対して最適（最大または最小）であるかどうかは完全に保証されてはいない．

非階層的クラスタリングは，一種の組合せ最適化問

† 大阪電気通信大学短期大学部
Junior College, Osaka Electro-Communication University
†† 大阪電気通信大学情報工学部
Faculty of Information Science and Technology, Osaka Electro-Communication University

題であり，いずれの既存手法も評価基準に対して近似的な意味での最適分割を求めるものであるといえる．分割最適化法では，初期の分割あるいはクラスタの核といった初期条件が最終の分割結果に強い影響を及ぼし，局所最適解（分割）に陥る場合も多い[3]．クラスタ分割結果は，クラスタの評価基準によって評価されるものであるが，実践的なクラスタリング手法としては，こうした評価基準を設定することと同様に，最適な分割状態を効率的に探索する安定したアルゴリズムが提供されなければならない．

組合せ最適化問題に有効な手法として遺伝的アルゴリズム（GA）[4)~6)]が知られている．GAは，生物の遺伝と進化の原理をモデル化した確率的探索，学習あるいは最適化の手法であり，複数の個体（解候補）に対して選択や交叉などの遺伝現象をモデルとした操作により解探索を行うアリゴリズムである．従来の単純な並列的解探索に比べ，複数の解候補を協調的に用いるためより良い解を獲得しやすい特長を持っている．一方，現在のところ対象とする個々の問題をGAで解くための一般的な方法は示されていない．問題の表現（遺伝子型へのコーディング）や操作の具体的な処理工程は，個々の問題に応じて設計する必要がある．

本論文は非階層的クラスタリングにおいて，最適な分割状態（解）の探索アルゴリズムとしてGAを採用し，在来の手法よりもロバストな分割結果を得ることをめざした一手法を提案する[7)]．すなわち，GAの1つの特長である多点探索性を積極的に活用し，分割最適化法における初期条件の影響を排除し，局所最適解の回避を狙っている．なお，クラスタの評価基準としては，判別分析[8)]などで用いられる平方和分割の原理に基づくクラスタ内平方和の総和を採用した．本論文で提案する手法の有効性を確認するため，在来の代表的手法であるk-means法に基づく比較実験を行った．実験対象として，平面上の点分布データ（点配置パターン）を用いて行った実験結果によれば，在来の手法による最適分割の獲得率が数%であった対象に対しても，本手法では80%近いの獲得率を示した．

2. k-means法

k-means法は，数多くある分割最適化法の中でも，最も標準的な手法と考えられる．k-means法の基本的な枠組みは，クラスタの評価基準とk群の初期分割を与え，平均ベクトル（重心）と平方和とを用いて評価基準に照らしながら分割の改良を行うものである．すなわち，k-means法では，クラスタの評価基準，初期分割および分割の改良手順という3つの要素を設定する必要がある．さらに，それぞれの要素についても多くの提案[1)]があり，全体の組合せによってk-means法は，無数に変形する．

2.1 クラスタ評価基準

n個の個体をk群の空でない互いに排反なクラスタに分割する場合，判別分析などで利用される平方和分割の原理より，次の関係が成り立つ．

$$S_T = S_W(k) + S_B(k) \tag{1}$$

ここで，S_Tは与えられるn個の個体の全平方和，$S_W(k)$はk群のクラスタに分割したときのクラスタ内平方和の総和，$S_B(k)$は同じくk群のクラスタ間平方和の総和を表す．すなわち，n個の個体をs次元の変量ベクトル$x_1, x_2, x_3, \cdots, x_n$，全個体の平均ベクトルを$m$，$k$個の排反なクラスタを$C_1, C_2, C_3, \cdots, C_k$，各クラスタ$C_i$の平均ベクトルおよびクラスタサイズ（所属個体数）をm_iおよびn_iと表し，ベクトルのノルムを$\|\cdot\|$で示すことにすれば，式(1)の各項は次のように表される．

全平方和：

$$S_T = \sum_{i=1}^{n} \|x_i - m\|^2 \tag{2}$$

クラスタ内平方和の総和：

$$S_W(k) = \sum_{i=1}^{k} \sum_{x_j \in C_i} \|x_j - m_i\|^2 \tag{3}$$

クラスタ間平方和の総和：

$$S_B(k) = \sum_{i=1}^{k} n_i \|m_i - m\|^2 \tag{4}$$

本論文では，クラスタの評価基準として，クラスタ内平方和の総和$S_W(k)$を採用する．このとき，与えられたkに対して$S_W(k)$を最小とする分割が最適な分割と判定される．

2.2 初期分割方法

分割最適化法では，初期分割あるいは分割の核（種子点）が最終分割結果を大きく左右する．前述のように，これについて多くの提案がある．

分割の対象となるn個の個体（個体識別のための番号が付与されている）に対して，一般に同値類などの系統的あるいは一意的な初期分割法がよく用いられる．本稿では，次章で提案するGA法との比較を考慮して，一様乱数を用いてn個の個体を空でないk群へ分割する方法を採用する．

2.3 分割改良のアルゴリズム

クラスタ内平方和の総和$S_W(k)$が最小となる分割

状態を探索するにあたって，すべての分割状態を調べるわけにはいかない．

ある1個体だけをあるクラスタから他のクラスタへと移動させることだけに注目すると，$S_W(k)$ の変化は，その個体の乗り換えがあった2クラスタ間だけの変化であることが分かる．すなわち，式(3)からクラスタ C_i から C_j へ個体 x_h が移動したときのクラスタ内平方和の総和の変化量 Δ_k は，

$$\Delta_k = \frac{n_j}{n_j+1}\|x_h - m_j\|^2 - \frac{n_i}{n_i-1}\|x_h - m_i\|^2 \quad (5)$$

で，$\Delta_k < 0$ のとき個体 x_h の移動によって $S_W(k)$ は，減少し，分割は改良されたと判断できる．

Δ_k を改良の判定基準としたクラスタ分割の具体的なアリゴリズムを以下に示す．なお，分割の対象となる個体は，識別のため一連の番号が付与されているものとする．

[k-means 法のアルゴリズム]

Step 1 初期分割を2.2節で述べた方法で与える．

Step 2 各クラスタのクラスタ内平方和，平均ベクトルおよびクラスタサイズを求める．

Step 3 移動させる個体を個体番号の若い順に選択する．ここで個体が所属するクラスタサイズが1（シングルトン）の場合，Step 6に飛ぶ．

Step 4 選択された個体について所属するクラスタから他のクラスタに移動させた場合の Δ_k を計算し，Δ_k が負数かつ最小となるクラスタへ移動させる．いずれのクラスタに対しても Δ_k が負数にならない場合，Step 6へ飛ぶ．

Step 5 移動があった2つのクラスタのクラスタ内平方和，平均ベクトルおよびクラスタサイズを更新する．

Step 6 すべて個体が一巡するまでStep 3からStep 5までを繰り返す．いずれの個体についても Δ_k が負数にならなくなったら終了し，そうでなければStep 3に戻る．

3. GA 法

GA法と名付けたクラスタリング手法を提案する．GA法の方針は，k-means法と同様にクラスタの評価基準と初期分割を与え，その分割を改良する方法である．

GA法では，まず分割状態を染色体によって表現し，複数の染色体を生成する．次に環境（集団）への適応度を指針として，遺伝操作によってより良い染色体を産み出していく．つまり，複数の初期分割を相互に利用することで，分割の改善を行う．処理の流れを図1に示す．以下にGA法で用いる遺伝操作などについて述べる．なお，分割の対象となる個体数は n で，各個体には個体識別のための一連の個体番号0から $n-1$ が付与されている．また，分割数は k で，各クラスタにも一連のクラスタ番号0から $k-1$ が付されている．

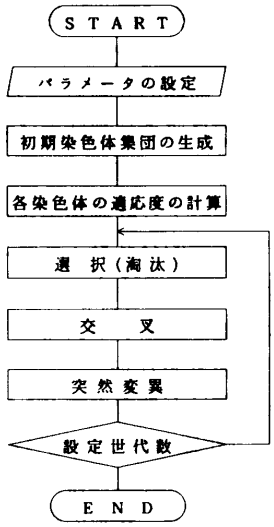


図1 遺伝的アルゴリズムの流れ図
Fig. 1 General flow chart of the proposed genetic algorithm.

(1) 染色体-遺伝子の表現

分割状態を染色体に対応づける．染色体の設計にあたっては，すべての分割状態が染色体に表現でき，かつすべての染色体の表現はただ1つの分割状態を表すように定める．染色体は遺伝子座と個体番号を対応させ，長さ n のストリングで表現する．各遺伝子座の遺伝子は，その遺伝子座が示す個体の所属するクラスタ番号をあてる（図2参照）．

こうした多値の遺伝子表現では，1つの分割状態に対して $k!$ 通りの表現が現われる．一般に対象と染色体の関係は，未熟な染色体集団への収束の危険性などから，一対一対応の表現が望まれる[9]．また，遺伝子座により対立遺伝子の意味に優劣が生じる．クラスタ分割の問題では，遺伝子座や対立遺伝子に対応づける個体番号やクラスタ番号は便宜的な識別子であるため，遺伝子座による優劣を極力排除する必要がある．そこで，あえてGA法では一対多対応の表現（図3参照）を許し，遺伝子座による優劣を染色体集団全体として緩和することを狙った．さらに後述する遺伝操作もこの優劣を排除する方向で設計した．

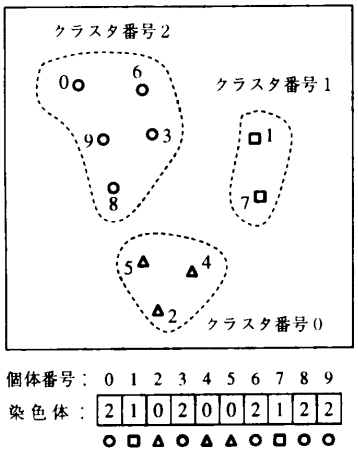


図2 染色体–遺伝子表現
Fig. 2 Genotype.

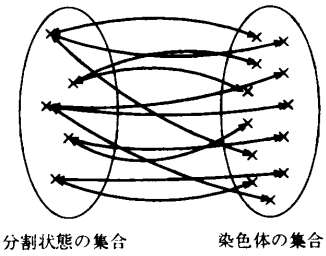


図3 分割状態と染色体の表現の関係
Fig. 3 Relation between the phenotype and genotype.

(2) 適応度

適応度は，次に述べる選択操作で期待値戦略を用いるため，最適な状態を最大とする尺度を設定する．

適応度として，式(4)に示すクラスタ間平方和の総和 $S_B(k)$ を用いる．これは，式(1)の関係からクラスタ評価値 $S_W(k)$ の最小化を最大化にスケーリングしたものである．すなわち，式(1)より

$$S_B(k) = S_T - S_W(k) \tag{6}$$

となる．ここで，全平方和 S_T は分割個数と分割状態にかかわらない固有の定数であるから，クラスタ評価値 $S_W(k)$ が最小化されるとき適応度は最大化されることになる．

(3) 選択（淘汰）操作

もとの集団から適応度に準じて，一様乱数を用いた期待値戦略とエリート保存戦略を併用してもとの集団と同数の染色体を選択する（同時に淘汰を行うことになる）．期待値戦略において，1度選択された染色体の期待値を減らす減少幅はパラメータとして実行時に与える．エリート染色体の保存操作は，期待値戦略による染色体採択終了後にエリート染色体が1つも含まれていない場合，採択された染色体のいずれか（一様乱数により決定）とエリート染色体を差し替える．

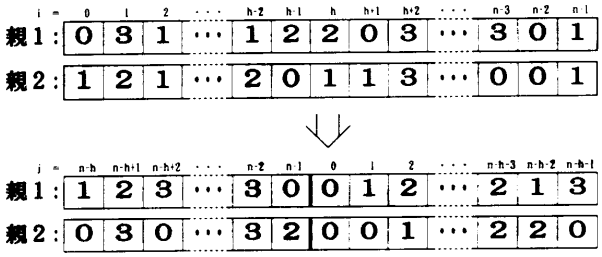


図4 対立遺伝子の置換え
Fig. 4 Exchange between allelomorphs.

(4) 交叉操作

交叉では，親が持つ形質を継承した子を造り出すことが重要である．ここで継承すべき形質とは，親の分割状態における共通部分であると考えられる．交叉の操作として，以下のような手順で優性遺伝[10]をモデルとした多点交叉を行う．

①染色体対の決定：**(3)** で選択した染色体の中から交叉確率に従い，交叉対象の染色体をランダムに決定する．交叉させる染色体の対は，**(3)** の選択で採択された順に組み合わせる．

②染色体表現の統一：**(1)** で決めた染色体–遺伝子表現は，1つの分割に対して複数の表現を許す．このため，交叉させる染色体対について表現を統一する必要がある．1つの分割状態を表す染色体の表現が同一になるように対立遺伝子（クラスタ番号）の置換えを考える．

遺伝子座は個体の識別番号と対応しているが，識別番号は便宜的なものであるため遺伝子並び（順位）そのものには意味がない．しかし，単純に2つの染色体を遺伝子の置換えによって統一的に表現していく場合，遺伝子座が若いほど同じクラスタに属する個体となる傾向が現れる．この遺伝子座による優劣の補正を考慮した対立遺伝子の置換え処理が必要である．

対立遺伝子の置換えは，①で決定した染色体対（両親）ごとに0から $n-1$ までの一様整数乱数を用いて基準遺伝子座 h を決定し，各遺伝子座 i を次式により j と読み替える（図4参照）．

$$j \equiv i - h \pmod{n} \tag{7}$$
$$i = 0, 1, 2, 3, \cdots, n-1$$

i を j と読み替えた各染色体ごとに $j=0$ の遺伝子座の対立遺伝子と同じ対立遺伝子はすべて0に置き換え（図4では，親1の対立遺伝子2を0，親2の対立遺伝子1を0に置き換える），次に置き換えられていない最小の j の対立遺伝子と同じ対立遺伝

子を1に置き換え（図4の親1では$j=1$の対立遺伝子0を1，親2では$j=2$の対立遺伝子3を1に置き換える），同様の処理をk回（対立遺伝子の数だけ）繰り返し，すべての遺伝子を置き換える．

③優性遺伝多点交叉：②の処理を行った染色体対について，式(7)のjの順序で遺伝子単位で対立遺伝子を比較する．異なる対立遺伝子を持つ遺伝子座が現われるたびに，対立遺伝子を入れ替えた場合の染色体の適応度が各々改善されるか否かを判定する．改善される場合に限り遺伝子を置き換える．一方の適応度が改善され，もう一方の適応度が改善されない場合，改善される方のみ置換えを行う．なお，適応度の改善があったかどうかの判定は，式(5)により行う．すなわち，$\Delta_k < 0$のとき改善されたと判定できる．置換えを行った染色体の適応度は，次式によりそのつど更新する．

$$S_B(k) := S_B(k) - \Delta_k \qquad (8)$$

以上の処理は，一般的な交叉操作と異なり遺伝子の完全な入替えにはなってはいない．優性遺伝子を積極的に残す変則的な多点交叉処理となっている．これを優性遺伝多点交叉と名付ける．

(5) 突然変異操作

突然変異とは形質の局所的変化であって，ある個体が属するクラスタから他のクラスタに所属を換えることである．突然変異の操作は，以下の手順で行う．

①突然変異位置の決定：突然変異確率に従い，突然変異を起こす染色体-遺伝子座をランダムに決定する．

②突然変異対立遺伝子の決定：①で決定した遺伝子座に対して他のクラスタに属するように0から$k-1$の一様整数乱数を用いて対立遺伝子を決め，交叉操作と同様に優性のみ置き換える．置換えを行った遺伝子座を含む染色体の適応度は，式(8)によりそのつど更新する．

(6) 世代

(1)に従い初期染色体集団を生成し，第0世代として適応度を計算する．その際，染色体ごとに式(5)および式(8)の計算に必要な各クラスタ（対立遺伝子）所属の個体数と平均ベクトルも計算し保持する．**(3)**から**(5)**までの一巡をもって1世代とする．世代間で最良の適応度を持つ染色体（エリート染色体）は，継承されるものとする．具体的には，世代開始時にエリート染色体を保持し，世代終了時に保持した染色体よりも高い適応度を持つ染色体が集団内に現われていなければ，集団内の最悪の適応度を持つ染色体と保持しているエリート染色体とを差し替える．アリゴリズムの停止は，世代数で与える．

以下，ここで示したGA法に対し，前章で述べたk-means法を在来法と呼ぶ．

4. クラスタ分割実験

在来法とGA法との分割のロバスト性を比較する実験，およびGA法の染色体集団とエリート染色体の適応度の世代による遷移を確認する実験を行った．各パラメータの設定値および実験結果などを**表1**にまとめた．

4.1 クラスタ分割の実験対象パターンの作成

クラスタ分割の実験対象としたデータは，視覚的検証が容易な平面上の点配置パターン（2変量データ）を採用した．[0,1)の正方領域にNeyman-Scottの方法[11),12)]を改変した以下の手順で生成した．

[Neyman-Scottの方法]

Step 1 パラメータとしてクラスタ数k，各クラスタに属する平均個体数μ，および各クラスタの稠密度（標準偏差）σを与える．

Step 2 k個のクラスタの核を2次元ベクトルm_i $(i=0,1,2,\cdots,k-1)$として，2次元一様乱数により決定する．なお，クラスタの核の間の距離が4σ以下のものがあるときは，再度k個の核を決定し直す．

Step 3 平均μのポアソン乱数を用いて各クラスタに所属する点数n_i $(i=0,1,2,\cdots,k-1)$を決める．

Step 4 各クラスタごとに平均m_i，分散σ^2の2次元正規乱数を用いてn_i個の点を生成する．

実験に用いる点配置パターンは，クラスタ分割数4，8，および12の3種類について乱数系列を換えながら各25パターン，合計75パターン作成した．その一例を**図5**に示す．なお，クラスタの平均個体数などのパラメータの設定値は表1に示す．

4.2 クラスタ分割実験の概要

(1) 比較実験

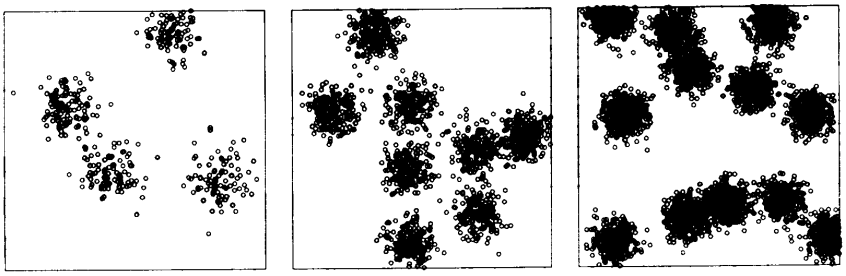
4，8，12分割の各25パターンに対して乱数系列を換えて各20回，合計1500回の試行を在来法，GA法の双方について行った．比較のため，式(5)のΔ_kの計算（正負判定）回数，個体のクラスタ移動回数，所要時間および分割状態を観測した．なお，GA法のパラメータの設定値は表1に示している．

(2) 遷移実験

GA法について図5(c)の点配置パターンに対して，**表2**に示すような交叉確率と突然変異確率のみが異なる2種類のパラメータセットを準備し，乱数系列を換えながら各100回の試行を行い，エリート染色体およ

表1 在来法とGA法との比較実験
Table 1 Comparative experiment of an existing procedure with the GA procedure.

点配置パターン	パラメータ設定値	クラスタ数［分割］	4	8	12
		クラスタ平均個体数［個体］	100	200	400
		クラスタ稠密度	0.06	0.05	0.04
		パターン数［件］	25	25	25
	パターン平均個体総数［個体］		398.08	1599.00	4799.80
在来法	パターンごとの試行件数［件］		20	20	20
	クラスタ数ごとの試行件数［件］		500	500	500
	実験結果	平均 Δ_k 計算回数［回］	4534.82	89862.93	702629.55
		平均個体移動回数［回］	356.94	2118.25	7842.85
		平均所要時間［秒］	0.07	0.38	1.96
		平均繰返し回数［回］	3.79	8.04	13.31
		最適分割一致件数［件］	458	174	21
		最適分割一致比率［%］	91.60	34.80	4.20
GA法	パラメタ設定値	集団サイズ［個体］	30	30	30
		期待値の減少幅	0.75	0.75	0.75
		交叉確率	0.60	0.60	0.60
		突然変異確率	0.03	0.03	0.03
		打ち切り世代数［世代］	20	35	55
	パターンごとの試行件数［件］		20	20	20
	クラスタ数ごとの試行件数［件］		500	500	500
	実験結果	平均 Δ_k 計算回数［回］	49999.90	487807.64	2288121.80
		平均個体移動回数［回］	11709.68	78031.59	290764.18
		平均所要時間［秒］	1.35	9.54	44.27
		平均獲得世代数［世代］	9.12	20.99	34.79
		平均獲得 Δ_k 計算回数［回］	40379.80	409038.97	1979536.74
		平均獲得個体移動回数［回］	10961.82	76513.52	288224.37
		最適分割一致件数［件］	500	492	386
		最適分割一致比率［%］	100.00	98.40	77.20
GA法/在来法	平均 Δ_k 計算総回数比		11.60	5.92	3.65
	平均個体移動総回数比		33.29	37.28	37.39
	平均所要時間比		19.29	25.11	22.59
	平均獲得 Δ_k 計算回数比		9.37	4.97	3.15
	平均獲得個体移動回数比		31.17	36.55	37.06
	最適分割一致件数比		1.09	2.83	18.38



(a) 4分割 (b) 8分割 (c) 12分割

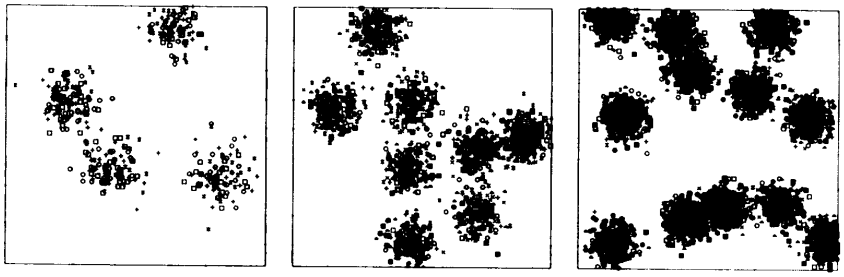
図5 実験用点配置パターン
Fig. 5 Experimental point pattern data.

び染色体集団の適応度の遷移を観測した．

4.3 クラスタ分割実験の結果

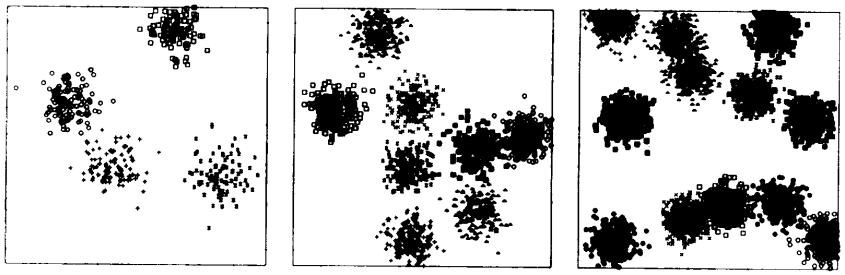
(1) 比較実験

表1に実験結果をまとめて示した．在来法の項目は，パターンごとの500回の試行に対して，平均の Δ_k の計算回数（平均 Δ_k 計算回数），平均の個体のクラスタ移動回数（平均個体移動回数），Sun Super SPARC+（50 MHz）で実行に要した平均時間（平均所要時間），収束した繰返し回数（平均繰返し回数），分割結果が最適分割と一致した件数（最適分割一致件数）およびその比率（最適分割一致比率）である．GA法の項目は，平均 Δ_k 計算回数，平均個体移動回数，平均所要



(a) 4 分割 (b) 8 分割 (c) 12 分割

図 6 初期分割
Fig. 6 Initial partition for the patterns presented in Fig. 5.



(a) 4 分割 (b) 8 分割 (c) 12 分割

図 7 最適分割と一致した分割結果
Fig. 7 Successful clustering result for the patterns presented in Fig. 5.

表 2 遷移実験におけるパラメータ測定値
Table 2 Parameter sets of the transition experiment for the GA procedure.

パラメータ項目	パラメータセット	
	set 1	set 2
集団サイズ [個体]	30	30
期待値の減少幅	0.75	0.75
交叉確率	0.60	1.00
突然変異確率	0.03	1.00
打ち切り世代数 [世代]	30	30
試行件数 [件]	100	100

時間，打ち切り世代でのエリート染色体と同じ染色体を最初に獲得した世代（獲得世代）の平均（平均獲得世代数），獲得世代までの Δ_k の計算回数の平均（平均獲得 Δ_k 計算回数），獲得世代までの個体のクラスタ移動回数の平均（平均獲得個体移動回数），最適分割一致回数および最適分割一致比率である．また，GA 法/在来法の項目は，在来法の各値に対する GA 法の各値の比である．

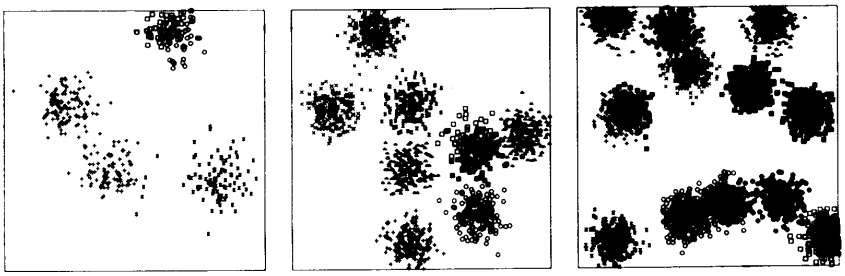
図 5 の各パターンの初期分割の一例を図 6 に，最適分割と一致した分割結果を図 7 に，最適分割と一致しない分割結果の一例を図 8 に示す．各図では同一クラスタに属する個体を同じ記号で示している．なお，最適分割と一致しない分割結果は，k-means 法，GA 法ともに図のような分割状態になる（GA 法では，4 分割の最適分割と一致しない分割は得ていない）．

(2) 遷移実験

図 9 と図 10 はパラメータセット 1（交叉確率 0.60，突然変異確率 0.03）およびパラメータセット 2（交叉確率 1.00，突然変異確率 1.00）での 100 回の試行をまとめたグラフである．グラフの横軸は世代，縦軸は最適分割を表す適応度に対する百分率（最適適応度獲得比率）を表し，実線のグラフは各世代でのエリート染色体の平均適応度の最適適応度獲得比率，破線のグラフは各世代でのすべての染色体の平均適応度の平均の最適適応度獲得比率を表す．なお，エリート染色体の平均適応度の最適適応度獲得比率は，図 9 では 10 世代で 91.5%，28 世代で 99.9%，図 10 では 7 世代で 91.9%，19 世代で 100%に達している．

4.4 クラスタ分割実験結果への考察

表 1 の実験結果の各項目が示す傾向は，最適分割一致件数および比率は，分割数（個体数）が増すとともに減少し，他の項目は増加している．この傾向は在来法，GA 法ともに同じである．しかし，最適分割一致



(a) 4 分割 (b) 8 分割 (c) 12 分割

図8 最適分割と一致しない分割結果
Fig. 8 Unsuccessful clustering result for the patterns presented in Fig. 5.

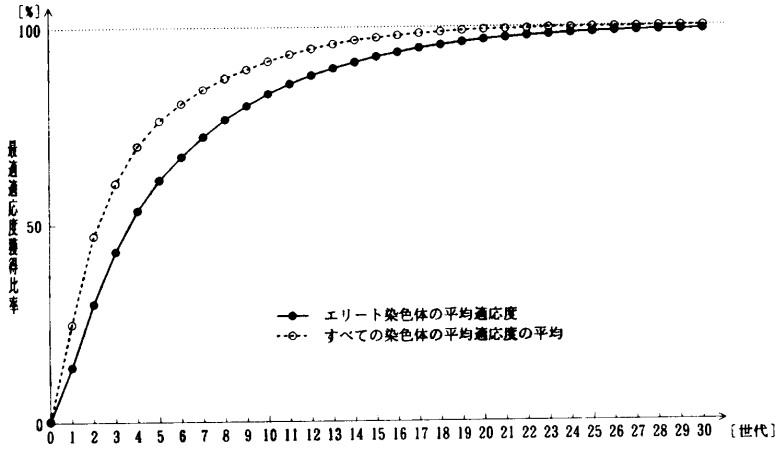


図9 世代-獲得適応度平均（交叉確率：0.60，突然変異確率：0.03）
Fig. 9 Genereation vs. the mean evaluation value for pattern (c) presented in Fig. 5 (probability of crossover: 0.60; probability of mutation: 0.03).

件数（比率）の減少傾向は，在来法とGA法との減少の度合いに大きな差がある．GA法/在来法の項目の最適分割一致件数比が示すようにGA法は在来法に比べ，4分割ではほぼ1倍，8分割では約3倍，12分割では約18倍で最適分割を得ている．一方，平均所要時間比などは，わずかに増加する（Δ_k の計算回数比は約3分の1に減少している）．すなわち，結果を得るに要する時間などのコストは，問題の大きさ（分割数，個体数）が大きくなるにしたがって，在来法，GA法の両手法とも同程度の増加傾向を示すが，最適分割を獲得する状況は，GA法が在来法を大きく上回る．つまり，GA法は在来法に比べてロバストであり，分割最適化法の短所である初期条件（初期分割，乱数系列など）の影響を強力に排除する能力があることを示している．

また，分割数が増すにつれ最適分割一致件数（比率）が減少し，在来法の平均繰返し回数およびGA法の平均獲得世代数が増加するのは，クラスタの評価基準との関係が大きいからと考えられる．規模の大きな問題（クラスタ数や個体数が多い分布）に対し平方和分解の原理に基づく評価値では，1個体の移動による評価値の増減はわずかになり，最適分割を表す値に近い局所値が多くなるためと思われる．

図9においてエリート染色体の平均適応度とすべての染色体の平均適応度の平均がともに単調増加している．一般的には染色体の平均適応度の平均が単調増加する場合は初期収束や局所最適分割に陥る可能性が高いと思われる．しかしながら，本手法の交叉処理および突然変異操作によって適応度が必ず改善される方向に推移するため適応度が低下することはない．一方，交叉操作の際に染色体の先頭位置をランダムに決めるため探索空間を縮めることなく最適分割にたどり着くものと思われる．そのため交叉確率および突然変異確率をともに1.00にしたときも図10が示すように着実に最適分割を獲得し，必要な世代数も減少している．このことは，交叉確率および突然変異確率が大きいほ

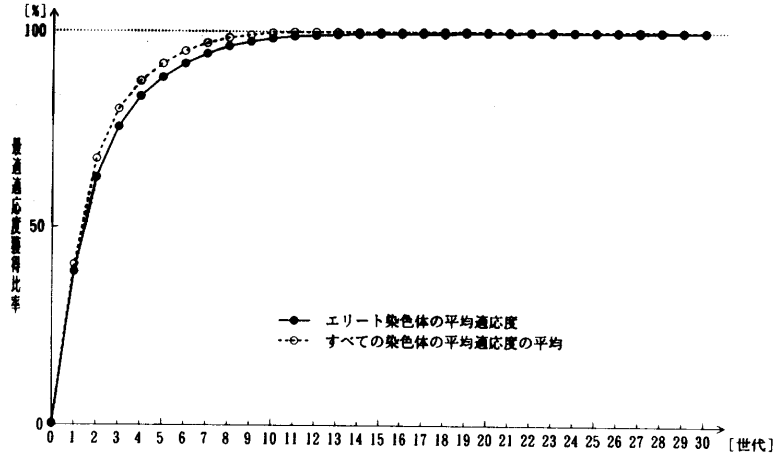


図10 世代-獲得適応度平均（交叉確率：1.00，突然変異確率：1.00）
Fig. 10 Genereation vs. the mean evaluation value for pattern (c) presented in Fig. 5 (probability of crossover: 1.00; probability of mutation: 1.00).

ど最適分割の発見が若い世代でなされると考えられる．一方，計算量の観点から見れば集団のサイズや打ち切り世代数を含め，適切なパラメータ値を設定することが必要である．

以上のことよりGA法は，在来法に比べてロバストであり，少なくとも本稿で用いた凝集的な分布データに対してきわめて有効と考えられる．

5. おわりに

本論文は，GAを用いた非階層的クラスタリング手法を提案し，在来法との比較実験を行った．実験結果よりGA法がロバストな手法であることが判明した．とくに優性遺伝をモデルとした優性多点交叉操作に特長があり，強力な機能を発揮することが実験により確認された．計算量的には，在来法の20～25倍程度である．この値を大きいと見るかあるいは小さいと見るかは意見の別れるところであるが，大域的な最適分割を得ることと引き換えるに許容できる範囲であると筆者は考えている．少なくとも本論文の実験で用いたレベルの凝集的な分布データについては，パラメータの設定にも依存するが，在来の手法よりもはるかに高い確率で最適分割を獲得できると考えている．

一方，凝集的でない分布に対しても，GA法の枠組みは適応度の計算を換えるだけで適用が可能と思われる．なお，この点については今後の検討課題としたい．また，先に述べた計算量の視点で見たときの適切なパラメータ値の設定方法の確立，および非階層分割であっても分割数が不明で動的に分割数が決定されるクラスタスタリング手法，たとえばISODATA法に対応したGAの枠組みの開発なども今後の検討課題となろう．さらに，既存の手法とGA法との中間的あるいは融合的な手法がどのような振舞いをするかなども興味のあるところである．

参考文献

1) Anderberg, M.R.（著），西田英朗（訳）：クラスター分析とその応用，p.442，内田老鶴圃，東京 (1988).
2) 仙波：組合せアリゴリズム，p.171，サイエンス社，東京 (1989).
3) Selim, S.Z. and Ismail, M.A.: K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality, *IEEE Trans. PAMI*, Vol.PAMI-6, No.1, pp.81–87 (1984).
4) Goldberg, D.E.: *Genetic Algorithms in Search, Optimization and Machine Learning*, p.412, Addison-Wesley, Reading (1989).
5) Michalewics, Z.: *Genetic Algorithms + Data Structures = Evolution Programs*, p.250, Springer-Verlag, Berlin (1992).
6) 樋口，北野：遺伝的アルゴリズムとその応用，情報処理，Vol.34, No.7, pp.871–883 (1993).
7) 加藤，小沢：遺伝的アルゴリズムを用いたクラスタリング，電子情報通信学会技術研究報告，PRU 95-148，pp.19–24 (1995).
8) 奥野，久米，芳賀，吉澤：多変量解析（改訂版），p.430，日科技連，東京 (1981).
9) 山村，小野，小林：形質の遺伝を重視した遺伝的アルゴリズムに基づく巡回セールスマン問題の解法，人工知能学会誌，Vol.7, No.6, pp.1049–1059 (1992).
10) 新津，佐藤，福田，柳澤，金谷，高岡：現代生物学，p.101，丸善，東京 (1987).

11) Smith, S.P. and Jain, A.K.: Testing for Uniformity in Multidimensional Data, *IEEE Trans. PAMI*, Vol.PAMI-6, No.1, pp.73–81 (1984).
12) Diggle, P.J.: On Parameter Estimation and Goodness-of-fit Testing for Spatial Point Patterns, *Biometrics*, Vol.35, pp.87–101 (1979).

(平成7年8月8日受付)
(平成8年9月12日採録)

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．平成2年同大学院担当（情報工学）．平成7年同大学情報工学部教授，同学部長．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等の研究に従事．電子情報通信学会，IEEE，英国BMVA，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」（単訳；雄山閣），「パターン情報数学」（森北出版，近刊）．