クラスタ分割への遺伝的アルゴリズムの適応

1E－1

加藤 常員・小沢 一雅
大阪電気通信大学

１．はじめに

非階層的クラスタリングは、組合せ最適化問題に類似した側面が強く、既存手法のいずれ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。主流の非階層的クラスタリング手法は、初期の分割などの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[1]。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、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。

本論文は非階層的クラスタリングにおいて、最適な分割状態の探索アルゴリズムとして遺伝的アルゴリズム(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手法[2]を提案する。すなわち、ＧＡのひとつの特長である多点探索性を積極的に活用し、初期条件等の影響を排除し、局所最適解の回避を狙ったものでる。

２．ｋ－ｍｅａｎｓ法

ｋ－ｍｅａｎｓ法は、非階層的クラスタリング手法のなかで、もっとも標準的な手法である。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。改良の方針は、１個体のクラスタ間の移動に着目し、評価基準が改善される場合に移動する。すべての個体に対して、移動による評価基準の改善がなくなったときの状態を最適分割とするものである。

３．ＧＡ法

ＧＡ法と名付けた非階層的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。つぎに環境(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。処理の流れを図１に示す。

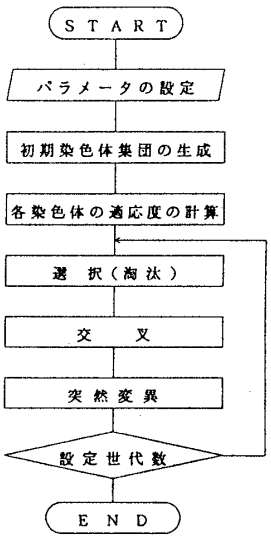


図１　ＧＡ法の流れ図

(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)。

(2) 適応度

適応度は、評価値をスケーリングしたことになるクラスタ間平方和の総和を用いる。

(3) 選択(淘汰)操作

もとの集団から適応度に準じて、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。

(4) 交叉操作

交叉では、優性遺伝をモデルとした多点交叉を行う。

まず、親とすべき染色体対を選択された順序で決定する。決定された染色体対に対して、染色体表現の統一を行う。統一は、遺伝子座による個体の優劣性を補正するように対立遺伝子の置き換える。対立遺伝子の置き換えは、染色体対(両親)ごとに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３参照)。

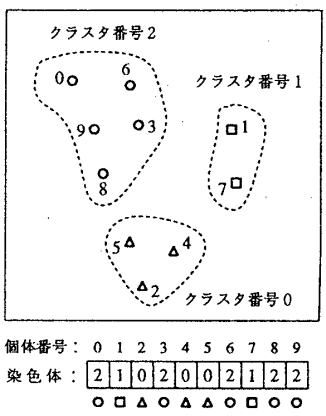


図２　染色体－遺伝子表現

An Application of the Genetic Algorithm to Clustering
Tsunekazu Kato, Kazumasa Ozawa
Osaka Electro-Communication University
Neyagawa-shi, Osaka 572, Japan

$$j \equiv i - h \pmod{n} \qquad (1)$$
$$i = 0, 1, 2, 3, \cdots, n-1$$

j＝0の遺伝子座の位置から対立遺伝子を小さな値の対立遺伝子に逐次的に置き換える。

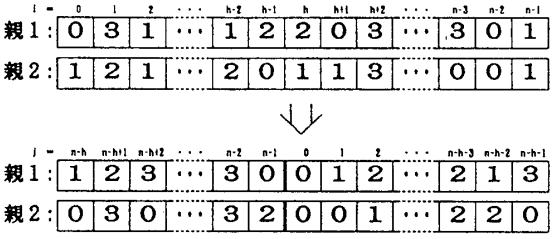


図3 対立遺伝子の置き換え

統一した染色体対について、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も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。

(5) 突然変異操作

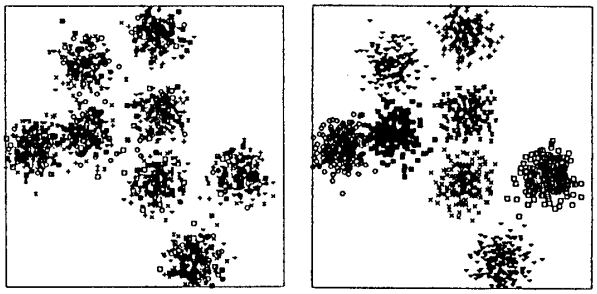
突然変異の操作は、突然変異確率にしたがい、突然変異位置を決定する。一様整数乱数を用いて対立遺伝子を決め、交叉操作と同様に優性のみ置き換える。

以下、ここで示したＧＡ法に対し、先に述べた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を行った。各パラメータの設定値および実験結果等を表１にまとめた。なお、実験パターンは、[0,1)の正方領域に Neyman-Scott の方法[3]を改変した手順で生成した。図４に８分割場合の初期分割および分割結果の一例を示す。

表１の在来法の繰り返し回数、ＧＡ法の最適分割獲得世代数および平均所要時間は、分割数が増すとともに増加している。最適分割一致件数(比率)は、在来法では分割数が増すにつれ急激に減少するが、ＧＡ法では、在来法ほど減少していない。また、両手法の平均所要時間比と最適分割一致件数比との関係を観ると時間比があまり変わらないのに対し、一致件数が分割数が多くなるほど大きくなっている。つまり、ＧＡ法が在来法に較べはるかにロバストな手法であることを示している。



(ａ)初期分割　　(ｂ)分割結果

図4 クラスタ分割

5．おわりに

本報告は、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を示した。今後の課題としては、凝集的でない分布に対してのＧＡ法の枠組みの開発、適切なパラメータ値の設定方法の確立などが挙げられる。

表１ 在来法とＧＡ法との比較実験

点配置パターン	クラスタ数		[分割]	4	8	12
	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	平均所要時間	[秒]	0.07	0.38	1.96
		平均繰り返し回数	[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	パラメータ設定値	集団サイズ	[個体]	30	30	30
		期待値の減小幅		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
		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	平均所要時間比			19.29	25.11	22.59
	最適分割一致件数比			1.09	2.83	18.38

参考文献

(1) Selim, S. Z. and Ismail, M. A.: K-Means-Type Algorithms:A Generalized Convergence Theorem and Characterization of Local Optimality, IEEE Trans. on PAMI, Vol. PAMI-6, No. 1, pp. 81-87(1984).

(2) 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、PRU95-148、pp. 19-24(1995).

(3) Smith S. P. and Jain A. K.:Testing for Uniformity in Multidimensional Data, IEEE Trans. on PAMI, Vol. PAMI-6, No. 1, pp. 73-81(1984).