確率密度関数の凹性を考慮した遺伝アルゴリズムに基づく パラメータ設定を伴う ISODATA クラスタリング

ISODATA clustering method with parameter estimation based on Genetic Algorithm : GA taking concaveness of probability density function into account

新井 康平*・ト 憲強**

Kohei ARAI and Xianqiang BU

Abstract : An improved ISODATA clustering method with merge and split parameters as well as initial cluster center determination with GA : Genetic Algorithm is proposed. Although ISODATA method is well-known clustering method, there is a problem that the iteration and clustering result is strongly depending on the initial parameters, especially the threshold for merge and split. Furthermore, it shows a relatively poor clustering performance in the case that the probability density function of data in concern can not be expressed with convex function. In order to overcome this situation, GA is introduced for the determination of initial cluster center as well as the threshold of merge and split between constructing clusters. Through experiments with simulated data, the well-known UCI repository data for clustering performance evaluations and ASTER/VNIR : Visible and Near Infrared Radiometer of imagery data, the proposed method is confirmed to be superior to the conventional ISODATA method.

Keywords : Clustering, ISODATA, GA, Concaveness of the probability density function, Nonlinear optimization

1. はじめに

クラスタリング手法は大別すると階層のおよび非階層のクラスタリング法に分けられる¹⁾²⁾。非階層型クラスタリング手法の代表的なものとして k-means 法と ISODATA 法がある。k-means 法は、事前にクラスタ数および初期クラスタ中心を与える必要があり、それらの設定によって計算時間や得られるクラスタ形状などが変わる。一方、ISODATA 法は設定したクラスタ数の一定範囲で自律的に有効なクラスタ数を決めることができる。ISODATA 法は対象データ個体を k-means 法に基づきクラスタに分離し、その後クラスタ内、クラスタ間の統計的指標に基づき、あらかじめ設定した閾値によって当該クラスタを分割・融合し、クラスタ個体を再配置するものであり、この一連の処理を再配置終了判定基準が満たされるまで繰り返す方

法である。したがって、ISODATA 法はクラスタリング精度が比較的高いが、これらのパラメータを設定するために少なからず処理時間を要し、この決定にはしばしば困難を伴う場合がある。最終クラスタ数、クラスタリング精度、計算時間は、これらのパラメータに依存している。特に衛星画像のような多次元データを対象とするクラスタリングの場合、パラメータの数も次元数の増加にしたがって増えるため、その調整は極めて困難を伴う。また、k-means 法と ISODATA 法は対象データの確率密度関数が凸関数であることを暗黙に仮定している。すなわち、凹関数であるようなクラスタの多次元空間における分布の場合は高いクラスタリング精度が期待できない。

本論文は、対象データが衛星画像のように多次元であることを想定し、当該クラスタ個体の確率密度関数が凹関数である場合にも対処し、推定した最適パラメータを用いてクラスタリング精度を向上することを目的とした ISODATA 法に基づくクラスタリング法を提案するものである。パラメータの最適化のため、遺伝アルゴリズム (Genetic Algorithms : GA)³⁾⁴⁾を用いた。GA は大域的最適解を保証するものではないが、比較的短時間にて準最適解を推定できる確率論的最適

* 佐賀大学理工学部

** 佐賀大学大学院

Saga University

〒840-8502 佐賀市本庄 1 番地

「写真測量とリモートセンシング」VOL. 47, NO. 1, 2008

化手法であり、解空間構造が不明であり、決定的な優れた解法が発見されてない、また、全探索が不可能と考えられるほど広大な解空間を持つ問題に有効である。

本論文では、クラスタ個体の分布が凹関数であるようなシミュレーションデータを用いて、まず、本提案手法の有効性を示す。また、クラスタリング精度の比較評価のために多用されている UCI リポジトリのデータセット⁵⁾および衛星画像データを提案手法に適用しクラスタリング精度を評価したところ、従来の ISODATA 法、また、シェープインデペンデントクラスタリング⁶⁾⁷⁾に対して優位性が認められたため、ここに報告する。

2. 提案手法

クラスタリングは、対象データ個体の間の類似性や非類似性に基づき、互いに似たもの同士を集めてグループ（クラスタ）を生成して対象データを分類する方法であり、統計解析ではクラスタ分析(cluster analysis)とも呼ばれる⁸⁾。対象データ個体がどの程度似ているかを測る基準は類似度（similarity）と非類似度（dissimilarity）がある⁸⁾。類似度は、たとえば、相関係数のように値が大きいほど対象が似ていることを示す尺度である。一方、非類似度は、分離度とも呼ばれ、その値が大きいほど対象が似ていないことを示す尺度である。一般には非類似度（距離）が良く使われている。クラスタリングにおいて定義する類似度は、一致係数（matching coefficient）や類似比（similarity ratio）などの指標があり、一方、距離の定義は多数ある⁸⁾。

本論文では対象データ個体の確率密度関数が凹関数であるような場合に対処し得る局所平均距離および類似度を用いることとし、これらに基づく適合度関数を用いた GA によって ISODATA 法に必要である初期クラスタ中心、並びに、クラスタ分割・融合のための閾値をも決定するクラスタリング法を提案する。

2.1 局所平均距離と局所類似度

2.1.1 Moving Window 法

局所範囲（Window）を決めると、その範囲内で局所的な個体間距離および類似度が求められる。局所範囲を各次元に沿って少しずつ移動することによって、局所的な距離および類似度を全域にわたってその和また

図1 Moving Window 法

は平均値を求める方法を Moving Window 法という（図1）。本研究では多次元ユークリッド空間においての超球体を Window として使う。同図において半径 r の局所範囲を移動しながら、この範囲内の個体間距離、または、類似度の和および平均を求める。

2.1.2 局所平均距離

式(1)に示すように、局所範囲内の個体間距離の平均値を局所平均距離 $\bar{L}$ という。

$$\bar{L} = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1, j \neq i}^n \delta_{i,j} \|l_i - l_j\| \quad (1)$$

ここで、

$$\delta_{i,j} = \begin{cases} 1 & \text{if } C_i = C_j \\ \frac{1}{2} & \text{if } C_i \neq C_j \end{cases} \quad (2)$$

である。また、 n は局所範囲内の個体数、 l は個体の位置ベクトル、 C_k は k の個体の帰属クラスタを意味する。すなわち、個体間の距離（ノルム）同一クラスタの場合は 1、異なるクラスの場合は半分の重みを考慮した局所範囲のすべての個体間距離の総和である。

2.1.3 局所類似度の和

ある局所の範囲内の直径と同じクラスタに属する個体間の距離の差を類似度とし、その和 DS を下の式の様に求める。

$$DS = \sum_{i=1}^n \sum_{j=1, j \neq i}^n \delta_{i,j} (2r - \|l_i - l_j\|) \quad (3)$$

ここで、

$$\delta_{i,j} = \begin{cases} 1 & \text{if } C_i = C_j \\ 0 & \text{if } C_i \neq C_j \end{cases} \quad (4)$$

である。

2.2 局所範囲の大きさと重みの決定

局所平均距離はクラスタ間分散を検出することが役割である。したがって、局所範囲、すなわち、Moving Window 法での Window サイズはクラスタ間分散をカバーできることが望ましい。Window サイズ、すなわち、超球体の半径を個体間最小距離から少しずつ大きくして、局所平均距離の全域の和を求めると、クラスタ間分散をカバーする時に、局所平均距離がピークとなる。このピークを求めて、その局所範囲における超球体の半径を基準 r_p とする。超球体の半径を変えて、図 2 (a), (b), (c) に示すシミュレーションデータを使って実験した。その結果を表 1 にまとめる。

表 1 をみると、超球体の半径を基準よりある程度大きくしても、クラスタリング結果に影響はないが、基準値を下回ると大きくクラスタ結果に影響する。これは異なるクラスタ間の個体間距離の最大を包含するように r を設定すれば両クラスタ個体の識別が出来るが、これよりも短く設定すると、すなわち、局所範囲を狭く設定すると両クラスタ個体を分けることが困難になることを意味している。そのため、 r を十分に長く設定すれば良いが、これが処理時間の増大 (r^2 に比例する) に直結するので問題である。したがって、本研

(a) Interrelated data set (b) Nested circle data set

(c) Spiral data set

図 2 2次元シミュレーションデータ

表 1 局所平均距離の Window サイズの影響

Error (%)	$0.5r_p$	$1.0r_p$	$1.5r_p$	$2.0r_p$
Interrelated	98.55	0.00	0.00	0.00
Nested circle	97.30	0.00	0.00	0.00
Spiral	96.77	96.77	0.00	0.00

究では、基準値の 2 倍 $2r$ を超球体の半径とした。

一方、局所類似度の和はクラスタ内の個体間距離と関係している。Window サイズはクラスタ内のすべての個体とその最近隣の距離より長く、クラスタ間分散より小さい方が望ましい。Window サイズが小さいと、クラスタリング結果は局所平均距離によって、クラスタの数が多くなってしまい、極端な場合は、すべて個体自身がそれぞれクラスタになる。したがって、最終クラスタ数から、Window サイズの妥当性が判断できる。孤立個体が存在しないと仮定すると、Window サイズ、すなわち、超球体の直径は他の個体から一番離れた個体と、その最近隣との距離 ($d_{\min}$) より長い場合が適切である。ここで、まず、超球体の半径を $d_{\min}$ に設定しクラスタリングを実施する。最終クラスタ数が想定の数 2 倍を越えるとき、または孤立個体が出る場合は超球体の半径を長くしてクラスタリングを実施する。最終クラスタ数が想定の数 2 倍以内に収まるまでこの過程を繰り返すものとした。

適合度は、まず、距離がある閾値以下の個体を 1 つのクラスタに属するように、すなわち、局所類似度の和 DS を最大値にすることが重要である。局所平均距離 $\bar{L}$ を可能な限り小さくすることに相当する。この観点から重み M を小さくすると良いが、 DS の有効数字は有限であるので、重みの値は $\frac{M}{L}$ の値を DS の有効数字に影響できる大きさにしなければならない。また、 DS を最大値にすることが優先であるので、 $\frac{M}{L} \ll DS$ でなければならない。

2.3 GA

GA とは、生物の進化を参考にした最適化アルゴリズムである⁹⁾。GA では問題の解を個体として表現し、各個体は染色体によって構成される。個体は選択淘汰、交叉、突然変異によって進化し、最適解の探索を行う。初期の GA は変数を 2 進符号化したビットストリングで交叉と突然変異を行っていた。これは変数の連続性を無視している。これに対して、数値そのものを用いて、変数の連続性を考慮した交叉や突然変異を行う GA が提案された。これを実数値 GA (Real-Coded Genetic Algorithms: RCGA) と呼ぶ¹⁰⁾。RCGA は探索に用いる個体を、従来の GA のようにビットストリングを用いるのではなく、実数ベクトルに変換して表現する。したがって、従来の GA の染色体の遺伝子座の構成はビットストリングから実数ベクトルに置き換

える。このとき、実数ベクトルは初期クラスタ中心の多次元ベクトルを用いる。本研究では RCGA に基づいて、初期クラスタ中心の最適化を行う。

2.3.1 適合度関数

クラスタ間分散の総和とクラスタ内分散の総和の比を最大にするような適合度関数を定義する。クラスタリング結果は、対象データ個体の確率密度関数が凸関数とみなせる場合に最適となるようにクラスタを構成してしまう。これに対して、局所類似度の和と局所平均距離を用いて、式(5)のように適合度関数を定義すると凹関数分布に場合にも最適なクラスタの構成が可能となる。

$$F = \sum DS + \frac{M}{\sum L} \quad (5)$$

ここで F は適合度、 M は重みである。局所平均距離はその局所範囲におけるクラスタ内分散を表している。この値が小さいほどクラスタ内分散が小さい。局所範囲における個体がすべて個別のクラスタに属する場合、局所平均距離が最小値となる。この値を適合度にする、範囲内の個体異なるクラスタに属する傾向があり、個体間分散が大きい局所域の両端にある個体が別のクラスタに属するようにクラスタリングすることができる。

局所類似度の和はその局所範囲におけるクラスタ内分散と関係している。この値が大きいほど局所範囲におけるクラスタ内の分散が小さい。局所範囲におけるすべての個体が一つのクラスタに属する時、局所類似度の和は最大値となる。この値を適合度にする、クラスタ数が少なくなる傾向があり、距離の近い個体一つのクラスタに属するようにクラスタリングすることができる。

式(5)によって、局所平均距離と局所類似度の和のバランスを取り、クラスタ内分散とクラスタ間分散を同時に配慮することによりクラスタの形状に依存しないような適合度関数を生成することができる。

2.3.2 交叉方法

RCGA はコード化を行わないため、これに特化した交叉オペレータが必要である。この代表的なものにブレンド交叉 (BLX- α)、単峰性正規分布交叉 (Unimodal Normal Distribution Crossover: UNDX) などがある。BLX- α は、次のように子個体を決める¹¹⁾。

1. 実数値ベクトル空間の 2 つの親個体を a , b とする。
2. 交叉の区間 $[A, B]$ を計算する。

$$A = \min(a, b) - \alpha|a - b| \quad (6)$$

$$B = \max(a, b) + \alpha|a - b| \quad (7)$$

ここで α は係数パラメータである。

3. 子個体を区間 $[A, B]$ から一様乱数で決定する。
- すなわち、従来の GA は染色体の遺伝子座の一部、または、すべてを入れ替える交叉方法と異なり、個体間距離の係数倍を個体ベクトルの小さい方から差し引いたベクトルから個体ベクトルの大きい方に足したベクトルまでの範囲において発生した一様乱数に従って交叉するベクトルを決定する方法である。

2.3.3 突然変異方法

正規分布により突然変異を与えた。すなわち、設定する突然変異の確率に応じて個体の実数ベクトルを正規分布に従って変化させる方法である。

2.3.4 選択淘汰方法

RCGA の選択淘汰方法をトーナメント方式とし、RCGA の収束性能を高めるエリート戦略を採用した。

2.3.5 RCGA の具体的な方法とパラメータ設定

ISODATA 法では実数値のパラメータが多いので、RCGA がより効率的である。本研究では RCGA を使って、ISODATA 法における初期クラスタ中心と分割、融合の閾値の最適化を行う。具体的な方法を以下に示す。

- RCGA の母集団の数を 30 に設定し、世代数を 300 に設定する。
- RCGA を採用しているため、各パラメータの実数値から多次元ベクトルを構成して、そのベクトルに対して交叉と突然変異を行うリアルコーディングを用いる。染色体のベクトルの構成は、

$$C = (S, M, A_1, A_2, \dots, A_m) \quad (8)$$

としている。ここで S は分割の閾値、 M は融合の閾値、 A_m は初期クラスタ中心の第 m 次元の座標である。このように染色体を式(8)で定義し、RCGA を行うことにより最適な初期クラスタ中心と ISODATA の分割および融合の閾値が求められる。

- トーナメント選択のサイズを3に設定する。
- 交叉方法は BLX- α 法を用いる。局所解に陥るのを防ぐこと、収束の速度も考慮して、 α の値は0.5に設定し、交叉確率を70%に設定する。
- 突然変異方法は正規分布による突然変異法を用いて、 σ を0.5に設定する。ここで突然変異確率を5%以下に設定すると局所解に陥りやすい、また突然変異確率が大きくなるとGAの効率が悪くなるので、5%に設定する。

クラスタリング結果を比較するために、一様乱数によりランダムに決定した初期クラスタ中心100組のクラスタリングを実施し、その最尤結果と平均値を計算する。また、幾つかの階層的クラスタリングも行う。

2.4 ISODATA 法の改良

前述のように、ISODATA 法は k-means 法により対象データ個体を超平面 (Voronoi 分割)¹²⁾によってクラスタ境界を構成して分ける方法である。すなわち、クラスタの確率密度関数が凸関数と暗黙に仮定しており、対象データ個体の確率密度関数が凹関数の場合は、正確なクラスタ分割は望めない。確率密度関数が凹関数の場合、ISODATA 法では分割と融合過程において対処できる場合もあるが、一度分割と融合によって正しく分類されたクラスタが再配置によって破壊されてしまう可能性がある。

提案手法は図3に示すように、ISODATA 法の終了判定を分割・融合処理の後に行う。すなわち、最後の融合を行った後に個体の再配置を行わないものとした。

また、図4に示すよう対象データ個体の確率密度関数が凹関数の場合、適切な分割の閾値によって分割されたと仮定する。この後にクラスタを融合する場合、融合過程においてクラスタの間の距離が変わるため、融合の順番にクラスタ結果は依存することになる。そのため、適切な融合の閾値があっても、望ましい融合結果にならない可能性がある。

図3 ISODATA 法の処理手順の改良

図4 ISODATA 法の融合過程

図5 提案手法の融合過程

したがって、提案手法はクラスタ結果を融合の閾値だけに依存するように限定するため、図5のように融合の閾値を満足するクラスタを同時に融合することにした。

2.5 計算量の削減

提案手法では初期クラスタ中心を実数値GA (RCGA)によってあらかじめ推定するので、クラスタリング結果はISODATA 法の繰り返しによるクラスタ中心の修正に頼ることなく、ほぼ最適なクラスタ結果が得られる。すなわち、ISODATA 法の繰り返し回数がある程度減らしてもクラスタリング結果はほとんど影響されない。実験ではISODATA 法の繰り返し回数は4に設定した。

Moving Window 法では次元数が増えると、計算量が指数的に増加する。提案手法では各次元に沿って少しずつ移動するのではなく、各個体を中心に順次移動することにより計算量を削減することにした。

3. 実験

3.1 2次元正規分布データ

まず、単純な2次元正規分布に従うデータ個体を用いて実験を行う。使用する2次元正規分布データは正規乱数によってランダムに生成した2組の正規分布データ個体である。それぞれの個体数は100、分散は0.5、平均値の差すなわち2つのクラスタ中心の距離は2である。クラスタリング結果を図6に示す。

この場合は同じクラスタリング結果になっている。すなわち、確率密度関数が凸関数であるような比較的簡単な正規分布データに対しては、ISODATA 法と提案手法に明らかな差はない。ここでISODATA 法は、初期クラスタ中心のみをRCGAにて求めた場合であ

(a) 正解 誤差：0 %

(b) ISODATA 法 誤差：2 %

(c) 提案手法 誤差：2 %

図6 正規分布データを用いた実験

り、従来の初期クラスタ中心を乱数によって与える ISODATA 法の場合、クラスタ結果が乱数に依存し、かつ、大きな誤差を生じている。

3.2 2次元シミュレーションデータ

図2に示すように、データ個体の確率密度関数が凹関数になるデータを生成し、ISODATA 法（初期クラスタ中心のみを RCGA にて決定）と提案手法によってクラスタリングを施し、それらの結果を比較した。

クラスタリング結果を図7，8，9に示す。

(a) ISODATA 法

(b) 提案手法

図7 Interrelated データセットに対するクラスタ結果

(a) ISODATA 法

(b) 提案手法

図8 Nested circle データセットに対するクラスタ結果

(a) ISODATA 法

(b) 提案手法

図9 Spiral データセットに対するクラスタ結果

すべての図形に対して、ISODATA 法におけるクラスタ結果は Voronoi 分割結果となるので正しい結果を得ることができないが、提案手法におけるクラスタ結果は正しく分割できていることが確認できる。

3.3 実データのデータセット

UCI リポジトリデータセットは、アメリカのカリフォルニア大学アーバイン校 (University of California, Irvine) の知識発見の研究者らが公開しているデータアーカイブである。これはウェブページ*から誰でもアクセスできる。本研究ではその中の Iris, Wine, New thyroid, また、R の Ruspini の各データセット、さらに、Chernoff の fossil データセットを用いて実験を行う。

1. Iris データセット Iris データセットは3種類の

*<http://mllearn.ics.uci.edu/MLRepository.html>

アイリス花のデータである。データセットの総個体数は150であり、種類ごとに50個体を含むものである。このデータセットのデータは4次元である。それらは萼片の幅と長さ、花卉の幅と長さであり、単位は cm である。Iris データセットは最も知られているクラスタリング用のデータセットの1つである。

2. Wine データセット Wine データセットは3種類のイタリアワインの化学分析データである。データセットの総個体数は178であり、各種別の個体数はそれぞれ59, 71, 48である。このデータセットのデータは13次元である。各次元のデータの範囲には大きな差がある。全数値が1以下の次元と1,000以上の数値まで含む次元がある。このため、Wine データセットのクラスタリングは難しいとされている。本研究では正規化 (Min : 0, Max : 1) した Wine データセットを使用した。
3. Ruspini データセット Ruspini データセットは R と S-plus に含まれたものである。カテゴリ数が4つの4次元データである。総個体数は75であり、それぞれは23, 20, 17, 15の個体を含むものである。
4. New thyroid データセット New thyroid データセットは UCI にある伝染病に関する5次元データである。総個体数は215である。また、カテゴリの数は3であり、カテゴリ1は150の個体、カテゴリ2は35, カテゴリ3は30を含むものである。
5. fossil データセット fossil データセットは Chernoff による3種類の石灰石に関する6次元のデータである。総個体数は87であり、各カテゴリは40, 34, 13の個体を含むものである。

これらの5つのデータセットはクラスタリング手法の比較によく使われている基準的なデータセットである。これらのデータセットに対して実験を行った。その結果を表2にまとめる。

表2 各データセットの実験結果

Error (%)	Proposed	k-means (ICCD)	ISODATA (Ran-Avg.)	Single linkage
iris	10.67	10.67	21.33	32.00
wine	4.49	5.62	34.65	61.24
ruspini	0.00	0.00	12.43	0.00
fossil	4.60	4.60	34.42	13.79
new thyroid	13.02	13.95	31.63	29.77

表2から、初期クラスタ中心をランダムに設定した場合の ISODATA 法に比べると、提案手法は誤差が20.23%も下がることが分かる。また、k-means 法 (ICCD) および Single linkage 法に比べても、提案手法の方の誤差がはるかに小さいことが示されている。

3.4 衛星画像

2004年12月7日に ASTER/VNIR : Visible and Near Infrared Radiometer によって撮影された佐賀地域の衛星画像から抽出した鹿児島市付近の一部 (図10) を用いてクラスタリングを実施した。図中に示す標本抽出域を設定し、30×30の同一クラスタ画像データ個体とした。これら個体のクラスタは海、植物、淡水、市街地の4つのクラスタ種別に相当している。画像データ個体の次元は3であり、各次元の最大値は255である。図11に当該画像を示す。

提案手法と従来の ISODATA 法を当該衛星画像に適用し、クラスタ結果を比較する。従来の ISODATA 法では Ward 法¹²⁾に基づく結果を初期クラスタ中心に使うものとした。それらの結果を表3および図12に示す。

図10 佐賀県鹿児島市付近の衛星画像 (Band : 1, 2, 3 N)

図11 クラスタを代表する標本抽出画像

表3 衛星画像を用いた実験結果

	Proposed	ISODATA
Error	0.00	3.89
elapped time (s)	25,567.445	1.134

(a) 提案手法

(b) ISODATA 法

(a) 原画像

(b) ISODATA

(c) 提案手法

図13 衛星画像を用いたクラスタ結果

(c) 左図は提案手法、右図は通常の ISODATA によるクラスタ結果のクラスタに含まれる個体データの ASTER/VNIR バンド 1, 2, 3 N の 3 次元特徴空間における分布のある断面における 2 次元分布

図12 衛星画像サンプリングデータ散布図

この表から分かるように、提案手法は対象データ個体の確率密度関数が凹関数であるような衛星データに対しても有効であることが示された。また、図13(a)の原画像中央付近から、前述の 4 つのカテゴリを含む 32*32 画素の一部（図13(a)）を抽出して、提案手法と従来の ISODATA 法によるクラスタリングを適用してクラスタ結果を比較した。クラスタリングの結果は図13(b)および図13(c)に示す。ここで提案手法と通常の ISODATA 法のクラスタ結果から各クラスタに含まれるデータ個体の ASTER/VNIR バンド 1, 2, 3 N の 3 次元特徴空間における分布を調べてみると、ある断面における 2 次元分布は図12となっており、個体の確率密度関数が凹関数で表現される分布になっていることが分かる。また、ISODATA 法ではこの分布に対応したクラスタが構成できていないのに比較し、提案手法では適切なクラスタが構成できていることが分かる。

図13から、衛星画像データについても従来の ISODATA 法よりも提案手法の方が、特に淡水の分類

については、より良い結果が得られることが分かる。

4. おわりに

シミュレーション、UCI リポジトリデータセットおよび ASTER/VNIR 画像を用いたクラスタリング実験結果から、すべての場合において提案手法は従来手法に比べ優位性が認められた。クラスタリング対象データ個体の確率密度関数が凹関数の場合においても提案手法は従来手法よりも分離度が高く、良好なクラスタリング結果が得られることが分かった。したがって、提案手法は ISODATA 法のパラメータ（初期クラスタ中心および分割・融合の閾値）の設定が困難な場合と、ISODATA 法の確率密度関数が凹関数の場合に正確なクラスタリングが期待できないという問題を克服し得ることを確認した。

提案手法の計算時間は従来手法よりも実数値遺伝アルゴリズムによるパラメータ設定の分だけ余計な時間が係るが、それにも増してクラスタ精度向上が著しいため、クラスタ精度を重視するような場合には有効であると考ええる。

(受付日2007.7.12, 受理日2008.1.9)

参考文献

- 1) R.J. Williams, *Simple statistical gradient-following algorithms for connectionist reinforcement learning*, Machine Learning, 8, 3-4, 229-256, 1992.
- 2) C. Yi-tsui, *Interactive Pattern Recognition*, Marcel Dekker Inc., New York and Basel, 1978.
- 3) John H. Holland, *Adaptation in Natural and Artificial Systems*, Ann Arbor, The University of Michigan Press, 1975.
- 4) John H. Holland, *Adaptation in Natural and Artificial Systems*, Massachusetts Institute of Technology, 1992.
- 5) UCI Repository, <http://www.sgi.com/tech/mlc/db/>
- 6) S Bandyopadhyay, *An automatic shape independent clustering technique*, Machine Intelligence Unit, January 06 Pattern Recognition, Vol. 37, No. 1, 2004.
- 7) Kohei Arai and Ali Ridho Barakba, *Method for shape independent clustering in case of numerical clustering together with condensed clustering*, Proc. of the SCI symposium 2004.
- 8) 新井康平, パターン認識の基礎理論, 学術図書出版, 1999.
- 9) L.J. Eshleman and J.P. Schaffer, *Real-Coded GA and internal schemata*, Foundations of Genetic Algorithms, 2, 187, 1993.
- 10) E.G.M. de Lacerda, *Model Selection of RBF Networks via Genetic Algorithms*, Pernambuco Federal University Informatics Center (Report), 2003.
- 11) 小野 功, 佐藤 浩, 小林重信, 単峰性正規分布交叉 *UNDX* を用いた実数値 GA による関数最適化, 人工知能学会誌, 14, 6, 1999.
- 12) 高木幹雄, 下田陽久編著, 新井康平共著, 画像解析ハンドブック, 東京大学出版会, pp.641-684, 1991.