多目的シミュレーテッドアニーリングを用いたクラスタリング

千田 智治

1　はじめに

クラスタリングは，他の様々なデータマイニング技術を利用する祭に，その処理効率をあげる目的で各マイニング処理の前処理として広く用いられる手法であり，その目的は大規模データをいくつかのグループに分割することである．クラスタリングは従来より，k-means 法，凝集法を筆頭に，遺伝的アルゴリズム（GA）を用いた手法や各手法を組み合わせたアンサンブル手法などが用いられているが，本研究では特に 2004 年に J.Handl と J.Knowles により提案された多目的クラスタリングアルゴリズム (Multiobjective clustering with automatick-determination:MOCK) に着目する．MOCK はクラスタリングを多目的最適化問題として捉え，多目的遺伝的アルゴリズムを用いて最適なクラスタリング結果を求める手法である．MOCK はクラスタ数が未知な状況での利用が想定されており，かつ幅広いデータセットに対して高いクラスタリング性能を示すことが報告されている．また，MOCK は最適なクラスタ数を自動的に決定する機能も有している．

MOCK は非常に有望なアルゴリズムではあるが，更なるクラスタリング性能の向上の為に，本研究では多目的シミュレーテッドアニーリング（SA）を用いて MOCK のクラスタリングアルゴリズムを適用する．本研究で用いる多目的 SA には，多目的 GA で一般的によく用いられる NSGAII と比較して同等かそれ以上の精度を示している A Simulated Annealing-Based Multiobjective Optimization（AMOSA）を利用する．

本報告では，多目的 SA について第二章で述べ，進化的計算手法の性能比較を第三章で行う．第四章でクラスタリングを多目的最適化問題として捉えた多目的クラスタリングについての説明を，第五章でその性能比較を行う．最後に第六章で結論を述べる．

2　多目的シミュレーテッドアニーリング

最も代表的な受理確率関数を用いた多目的 SA（MOSA）と，アーカイブの概念を持つ多目的 SA（AMOSA）について説明する．

2.1　MOSA

単目的 SA と MOSA で大きく異なる点は，次状態の受理判定に複数の評価基準が用いられる点である．通常，単目的 SA では Metropolis 基準を用いて受理判定を行うが，MOSA では複数の評価基準を持つためそのまま使用することはできない．これまでに様々な多目的最適化問題に適用可能な受理確率関数が提案されているが，本研究では最も代表的な受理確率関数である Rule SL を MOSA に適用した [1]．この Rule SL を (1) 式で示す．この時，$\Delta f_j = f_j(x) - f_j(x')$ であり，x(現在の解)，x'(次状態) である．また T は温度，$\omega_j(\omega_1 + \ + \omega_p = 1)$ は目的関数 f_j に対する重みを示す．

$$P_r = min\left\{1, \exp\left(\frac{\sum_{j=1}^{p}\omega_j \Delta f_j}{T}\right)\right\} \qquad (1)$$

この受理確率関数は重み係数法により複数の目的関数を単目的化し，単目的 SA で用いられる Metropolis 基準を用いて受理判定を行う．

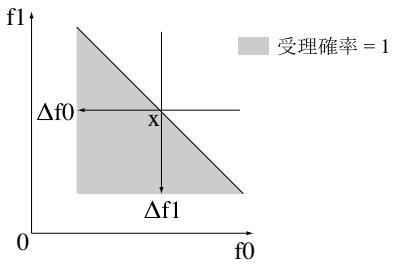


Fig.1　Rule SL(出典：自作)

2 目的最小化問題における各受理確率関数の $f_0 - f_1$ 平面での等高線の模式図を Fig. 1 に示す．x が $f_0 - f_1$ 平面上における現在の解の位置であり，解候補と二つの目的関数値の差を Δf_j で表し，点 x を原点として Δf_0 軸，Δf_1 を描いた．

2.2　AMOSA

2008 年に S.Bandyopadhyay らによって考案された A Simulated Annealing-Based Multiobjective Optimization（AMOSA）[2] はトレードオフの関係の目的関数から解を導出する為，アーカイブの概念を持った多目的最適化アルゴリズムを基盤とした SA である．AMOSA では，非劣解を保存するアーカイブ個体群が一定量を超えた場合，混雑距離を計算して非劣解を削減するクラスタリング処理を行う．また，現在の解と次状態の解の状態及びアーカイブから支配率を求め，受理確率を動的に変化させて解探索を行う．

温度を T，現在の状態を q，次状態 s とすると次状態は (2) 式の確率で選択される．$E(s,T)$，$E(q,T)$ はそれぞれエネルギー値 s，q である．

$$P_{qs} = \frac{1}{1 - e^{\frac{-(E(q,T)-E(s,T))}{T}}} \qquad (2)$$

2.2.1 アーカイブの初期化

初期化時には単純な山登り法で初期化を行う．次状態の解が現在の解を優越していれば受理することを繰り返し，アーカイブ数 HL 個まで解を保存する．もし，アーカイブ数が HL を越えることがあれば，クラスタリングを適用して HL までアーカイブの削減を行う．クラスタリングについては 2.2.2 項で説明する．

AMOSA には，次状態で生成された解がアーカイブを「より優越している」か「殆ど優越していない」かという優越の量を判断することで受理確率の値を変化させる．この優越の量については 2.2.3 項で述べる．

2.2.2 クラスタリング

AMOSA では，より広域で多様性のある解を形成する為にアーカイブの上限数を越えた解を削減するためにクラスタリング処理を行う．クラスタリングには，よく知られている単結合法 (Single-Link) が用いられており，この解の削減方法は多目的 GA の SPEA でも用いられている手法である．

アーカイブの全てのペアのユークリッド距離を求め，最も短いものから順にクラスタを形成していく．これをアーカイブ数の上限である HL まで繰り返した後，各クラスタ内で他の個体に対して最小平均距離を持つ個体を選択して，残りの個体を全て削除する．

2.2.3 優越の量

AMOSA は，優越の量という概念を持っており，優越の量は次状態の解の受理確率の計算で用いている．2 目的最小化問題における解 A，B の優越の概念を Fig. 2 に示す．

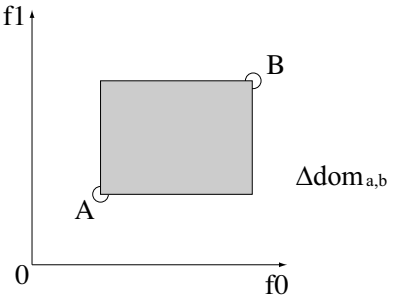


Fig.2　優越の量 (出典：自作)

Fig. 2 に示すように 2 つの解空間においてユークリッド距離を求め，その面積（体積）Δdom を優越の量とする．

2.2.4 AMOSA の主なプロセス

現在の解 ($current-pt$)，次状態の解 ($new-pt$)，アーカイブの 3 種類の解の優越関係によって AMOSA では受理確率を細かく変化させている．Case1〜Case3 の 3 つの状態に分けられる．

Case1：$current-pt$ が $new-pt$ を支配している場合

Case1 ではアーカイブの k 点が $current-pt$ を優越している．k=0 の場合を Fig. 3(a) に $k\geq 1$ の場合

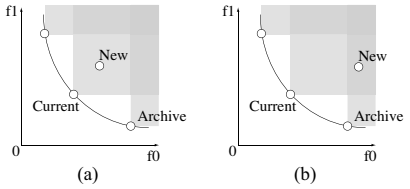


Fig.3　case1(出典：自作)

, $Fig.\ 3(b)$ に示す. Δdom は (3) 式で，受理確率は (4) 式で決定される．その後，次状態の設計変数と評価値を現在の設計変数と評価値に移す．

$$\Delta dom_{avg} = \frac{((\sum_{i=1}^{k} \Delta dom_{i,new-pt}) + \Delta dom_{current-pt,new-pt})}{k+1} \quad (3)$$

$$P = \frac{1}{1+\exp(\Delta dom_{avg} * temp)} \quad (4)$$

Case2：$current-pt$ と $new-pt$ がどの解も支配していない場合

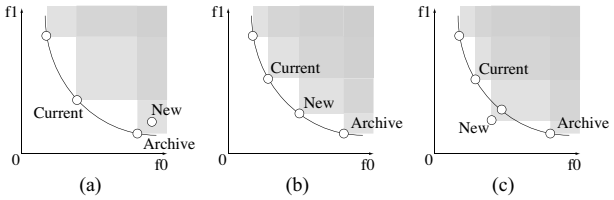


Fig.4　case2(出典：自作)

Fig. 4(a) は，$new-pt$ が $k(k\geq 1)$ 個のアーカイブに支配されている場合である. Δdom は (5) 式で，受理確率は (4) 式で決定される．その後，次状態の設計変数と評価値を現在の設計変数と評価値に移す．

$$\Delta dom_{avg} = \frac{\sum_{i=1}^{k} \Delta dom_{i,new-pt}}{k} \quad (5)$$

Fig. 4(b) は，$new-pt$ がどのアーカイブも支配していない場合であり，アーカイブに $new-pt$ の設計変数と評価値を追加する．この際，解がアーカイブ数 HL を越えた場合，クラスタリングを行いアーカイブの削減を行う．

Fig. 4(c) は，$new-pt$ が $k(k\geq 1)$ 個のアーカイブを支配している場合であり，アーカイブにnew-ptの設計変数と評価値を追加する．

Case3：$new-pt$ が $current-pt$ を支配している場合

Fig. 5(a) は，$new-pt$ が $k(k\geq 1)$ 個のアーカイブに支配されている場合であり，受理確率は (6) 式で決定される．その後，そのアーカイブの設計変数と評価値を現在の設計変数と評価値に移す．

$$P = \frac{1}{1+\exp(-\Delta dom_{min})} \quad (6)$$

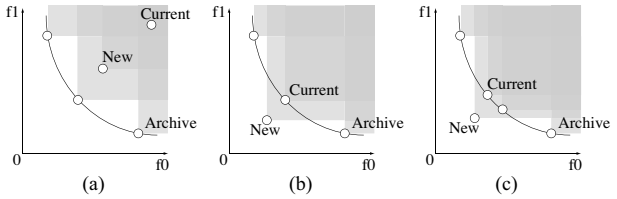


Fig.5　Case3(出典：自作)

Fig. 5(b) は，$new-pt$ がアーカイブを支配していない場合であり，$new-pt$ の設計変数と評価値を $current-pt$ の設計変数と評価値に移す．その際，アーカイブに $new-pt$ を追加する．また，$current-pt$ がアーカイブにある場合は取り除く．

Fig. 5(b) は，$new-pt$ がアーカイブの k 点を支配している場合であり，$new-pt$ の設計変数と評価値を $current-pt$ の設計変数と評価値に移す．その際，優越されている k（$k\geq 1$）個の解をアーカイブから取り除く．

以上のように，AMOSA では解の優越関係を細かく場合分けをして，受理確率や $current-pt$ の決定を繰り返して探索を行っている．

3　数値実験 1

3.1　実験内容

本実験では，多目的 GA と多目的 SA の精度比較を行い，AMOSA の有効性を確認する．多目的 GA は一般的に用いられる NSGAII を，多目的 SA は MOSA と AMOSA を用いて，3 種類の異なる進化的計算手法の数値実験を行う．

3.1.1　対象問題

本実験に用いる対象問題は ZDT4 と KUR である．ZDT4 は単峰性と多峰性の異なる形状の目的関数が 2 個，設計変数が 10 個の対象問題である．KUR は目的関数が 2 個，設計変数が 100 個であり，一方の目的関数において連続する 2 変数間の相互作用を持ち，もう一方の目的関数において多峰性を有する対象問題である．

3.1.2　パラメータ

本実験で用いたパラメータを Table1 に示す．尚，評価計算回数は全手法で 1 試行につき，240000 回に統一している．これは，MOSA の探索が十分に行われる回数を基準に AMOSA，NSGAII も同様に統一することでその他パラメータを決定した．

MOSA では，96 クーリングステップ数× 250 クーリングサイクル数× 10 探索点＝ 240000 回．AMOSA では，2500 クーリングステップ数× 96 温度降下回数＝ 240000．NSGAII では，2400 世代× 100 個体＝ 240000 回となっている．これを全手法 30 試行繰り返し実行した．

3.2　実験結果

MOSA，AMOSA，NSGAII の実験結果を以下に示す．ZDT4 の実験結果を Fig. 6，KUR の実験結果を Fig. 7 とし，30 試行で得られた結果を全て表示している．

Table1　MOSA・AMOSA・NSGAII のパラメータ

MOSA	
クーリングステップ数	96
クーリングサイクル	250
探索点	10
最高温度	1.0
冷却率	0.90
非劣解を格納する配列の要素数	100
近傍数	0.5
AMOSA	
クーリングステップ数	250
温度降下回数	96
最高温度	200.0
最低温度	0.0000001
アーカイブ数	100
近傍数	0.5
NSGAII	
個体数	100
世代数	2400
アーカイブ数	100
トーナメント数	2
遺伝子長	設計変数× 20
突然変異率	1.0 ／遺伝子長

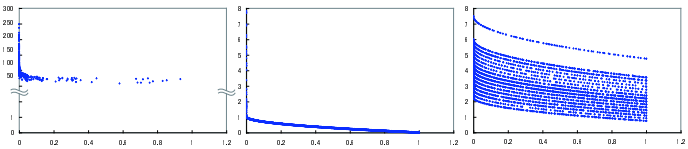


Fig.6　ZDT4 の実験結果 (出典：自作)

Fig. 6 より，探索性能の良い順に AMOSA，NSGAII，MOSA であることがわかる．AMOSA で得られた解が真のパレート最適解と一致し，30 試行全てにおいて安定した精度の高さを示している．NSGAII では，両目的関数値が 1 以下である真のパレート最適解には到達することはできなかったが，MOSA と比較して良い結果を示した．MOSA は試行回数に関わらず探索があまり進まず，他の手法に比べて精度が悪いことがわかる．

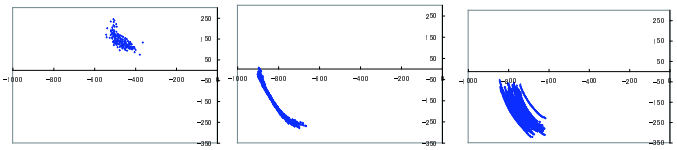


Fig.7　KUR の実験結果 (出典：自作)

Fig. 7 より，AMOSA, NSGAII 共にほぼ同等の結果を示した．ZDT4 の時と同様，AMOSA は全ての試行において安定した精度を示している．MOSA も ZDT4 の時と同様に AMOSA，NSGAII よりもかなり精度が悪い．

以上より，AMOSA は他の多目的 SA や多目的 GA と比較して精度の高い進化的計算手法であることがわかった．

4 多目的クラスタリング

4.1 クラスタリング

クラスタリングは，大量のデータをいくつかのグループに分割する場合や，その後のマイニング処理の効率化を図る場合などに用いられる汎用的なデータマイニング手法である．一般にクラスタリングにおける良いクラスタソリューション（分割）とは，クラスタ内の均質性が高く，かつクラスタ間の分離性が高い分割が行われたものとする．均質性とは，近接したデータが同じクラスタに含まれると高くなる評価指標であり，分離性とは，離れたデータが異なるクラスタに含まれると高くなる評価指標である．

4.2 多目的クラスタリング

本研究で用いる多目的クラスタリングアルゴリズム Multiobjective clustering with automatick-determination（MOCK）は，2004 年に J.Handl と J.Knowles が提唱した進化的アルゴリズムベースのクラスタリングアルゴリズムの一種である [3) 4)]．多目的 GA の PESAII を用いて均質性と分離性双方の評価指標に基づいた目的関数を用いて 2 つの評価指標を同時に最適化することにより，それぞれの評価指標を用いた場合の利点を生かし欠点を補いながら，質の高いクラスタソリューションを発見できるよう設計されたクラスタリングアルゴリズムである．

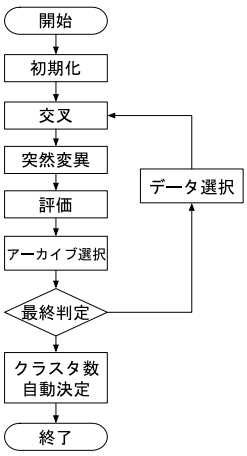


Fig.8 MOCK の Flow

MOCK のアルゴリズムの Flow を Fig. 8 に載せる．MOCK では，初期化時にグラフベースの個体表現を基に，より良質な初期個体を得るために最小全域木を用いた初期化処理が行われる．この初期化では，凝集法で得られる解に似た良質な初期個体群を生成することが可能である．また，MOCK の交叉は，どちらか一方の親の遺伝子を受け継ぐ一様交叉を用いており，突然変異には探索空間サイズを大幅に削減する近傍突然変異が用いられる．Flow に示すように遺伝的操作により幅広いクラスタ数を持つ解からなるパレート解集合を得るステップと，得られたパレート解集合から最適なクラスタ数及びクラスタ境界を持つ最終解を決定するステップに分けられる．

4.3 評価関数

MOCK では，分離性を表す評価関数である Overall Deviation と均質性を表す評価関数 Connectivity を用いる．この 2 つの評価関数を同時に最適化することでクラスタリングを行なう．この評価関数より，解集合内のクラスタ数の多様性を維持したまま解の質を高めることができる．

4.3.1 Overall Deviation

Overall Deviation はクラスタの分離性に基づく大域的な評価指標である．Overall Deviation は各クラスタの中心点 μ_k から同じクラスタ内の各データ i までの距離の総和で定義される．この指標を最小化することにより，コンパクトなクラスタが生成される．この評価指標の数式を,(7) 式に載せる．

$$Dev(C) = \sum_{C_k \in C} \sum_{i \in C_k} \delta(i, \mu_k) \tag{7}$$

4.3.2 Connectivity

Connectivity はクラスタの均質性に基づく局所的な評価指標として用いられる．Connectivity はあるデータ i の $1 \sim L$ 番目までの近傍 (近い順を j とする) が，データ i と異なるクラスタに存在する場合，$1/j$ のペナルティを与えるという指標であり，全てのデータについてのペナルティ値の総和が Connectivity 値となる．この指標を最小化することにより，近傍のデータ同士がより高い確率で同じクラスタに存在することになる．この評価指標の数式を (8) 式に載せる．

$$Conn(C) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nn_i(j)},$$
$$where\ x_{r,s} = \begin{cases} \frac{1}{j} & if\ \nexists C_k : r, s \in C_k \\ 0 & otherwize \end{cases} \tag{8}$$

4.4 グラフベースの個体表現

MOCK ではグラフベースの個体表現を行ない，任意のクラスタ数を持つ個体を同時に表現可能である．また，クラスタ数が未知の問題において，最適なクラスタ数を特定する機能を持っている．Fig. 9 に，個体表現を示す．

与えられた各データがグラフ内の各点をそれぞれ表し，各遺伝子が各エッジに対応する．エッジは両端のノードに対応するデータが同じクラスタに属する事を意味する．また，各遺伝子と各ノードを 1 対 1 に対応させ，グラフ

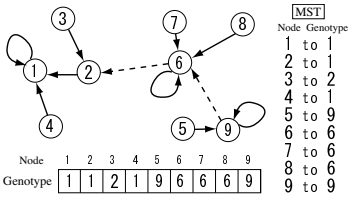


Fig.9　グラフベースの個体表現 (出典：自作)

内の各エッジを遺伝子に変換する為，各ノードから出るエッジが 1 本に限定された有向グラフを用いる．

4.5　最小全域木

MOCK の初期化には，与えられたグラフより最小全域木 (Minimum Spanning Tree:MST) を生成する．MST とは，エッジが最小の数で生成できるグラフ（Fig. 10）のことをいう．

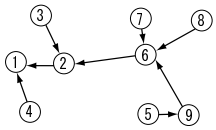


Fig.10　最小全域木 (出典：自作)

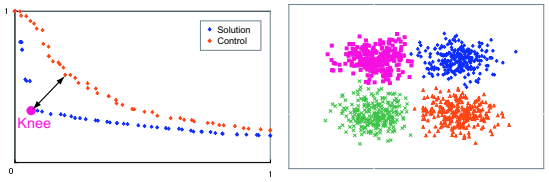
最初に与えられたデータセットに対して完全な MST を作成する．初期個体の i 番目の個体は $(i-1)$ 番以上に長いエッジを取り除いて作成される．このように，グラフベースの個体表現と MST を用いた初期化アルゴリズムより，任意のクラスタ数を持つ質の高い初期個体を生成することが可能となる．

4.6　クラスタ数自動決定アルゴリズム

MOCK の大きな特徴は，クラスタ数の自動決定を行えることである．MOCK のクラスタ数自動決定は Tibshirani らの Gap 統計を元に開発された．これはクラスタサイズを決定する統計的な手法である．Gap 統計はクラスタリングアルゴリズムをクラスタ数を変数とする関数として捉えた場合，最も適切なクラスタ数において特徴的な「Knee（膝）」が見られるであろうという期待に基づいた手法である．

Gap 統計では，与えられたデータと同じ領域内にランダムにデータを生成する．そのランダムデータ（コントロールデータ）に MOCK のクラスタリング処理を行い，元のデータのクラスタリング結果で得られたパレートと比較し，コントロール曲線から最も遠い点を最適解として選択する．

Fig. 11(a) では，左下の点「Knee」がコントロール曲線から最も遠い点であり，この点を最適解とする．Fig. 11(b) より，クラスタ数 4 のクラスタリング結果を示している．4 つの島を作るデータ群がそれぞれクラスタを形成しており，人間が視覚的に判断してクラスタリングを行う結果とほぼ一致している．このように，MOCK は得られたパレート解の中から最も良いと思われる解を自



(a) Kneeの決定　　(b) 最適なクラスタリング結果

Fig.11　クラスタ数自動決定アルゴリズム (出典：自作)

動的に決定するメカニズムを有する．

4.7　提案手法

本報告では 2.2 節で説明した AMOSA をベースとするクラスタリングアルゴリズム AMOSA-MOCK を提案する．J.Handl らが提案した MOCK では，多目的遺伝的アルゴリズム PESAII をベースとしているが，進化的計算手法の段階での精度比較を行い易くするために NSGAII をベースにして MOCK のプログラムを組み直した．この手法を NSGAII-MOCK とする（Fig. 12）．

Fig.12　MOCK の手法一覧 (出典：自作)

本研究の目的は，NSGAII-MOCK よりも精度の良いクラスタリングアルゴリズムの提案であり，現段階では AMOSA-MOCK の性能を確認し問題点の発見を行う．

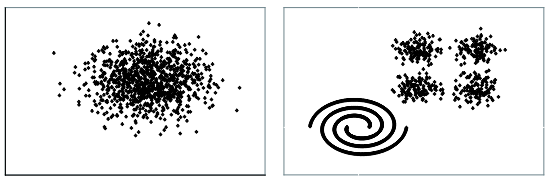
5　数値実験 2

5.1　実験内容

本実験では，以下のテストデータセットを用いて AMOSA-MOCK と NSGAII-MOCK のクラスタリング性能を比較する．

5.1.1　テストデータセット

データ数が 13000 の Square1_13000 とデータ数が 1500 の SpiralSquare の異なるデータ数・形状であるテストデータセット 2 種類を用いる（Fig. 13）．



(a) Square1_13000　　(b) SpiralSquare

Fig.13　テストデータセット (出典：自作)

Square1_13000 はデータが中央に集中した形状をしているため，k-means 法のような非階層型クラスタリング手法には解き易いが，凝集法のような階層型クラスタリング

手法では解く事が困難な問題である．一方 SpiralSquare は，渦巻き上の形状をしたデータ群があるため，階層型クラスタリング手法には解き易いが，非階層型クラスタリング手法では解く事が困難な問題である．

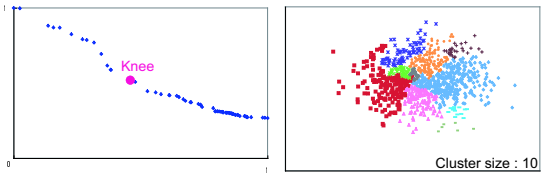
5.1.2　パラメータ

AMOSA-MOCK，NSGAII-MOCK 共に，最大クラスタ数 25，近傍数 20 とし，用いる遺伝子数はテストデータ数と同じとする．また，近傍とはあるデータから近接したデータの事を意味し，近傍数 20 はあるデータからユークリッド距離の近いデータを順に 20 個ということを意味する．

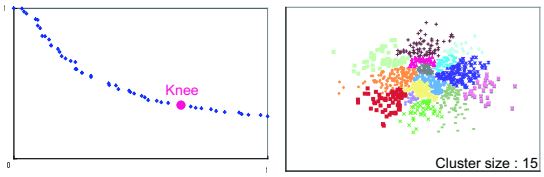
最大クラスタ数	25
近傍数	20
遺伝子数	テストデータ数

5.2　実験結果

AMOSA-MOCK，NSGAII-MOCK の実験結果を以下に示す．

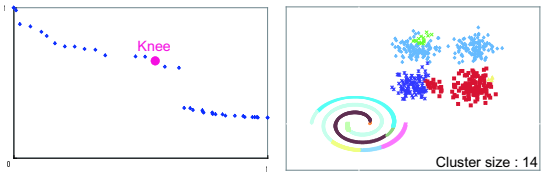


(a) AMOSA-MOCK

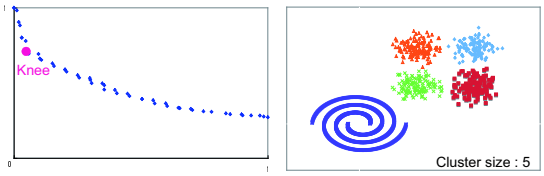


(b) NSGA2-MOCK

Fig.14　Square1_13000 の実験結果 (出典：自作)



(a) AMOSA-MOCK



(b) NSGA2-MOCK

Fig.15　SpiralSquare の実験結果 (出典：自作)

Fig. 14(a)，Fig. 15(a) より，AMOSA-MOCK は NSGAII-MOCK よりも解の精度が悪い事がわかる．このパレート解の精度が良ければ，テストデータを均質性，分離性を考慮してクラスタリングを行うため，適切なクラスタリング結果を示す．Fig. 14(b)，Fig. 15(b) より，NSGAII-MOCK の方がクラスタリング結果の質が高い．Square1_13000 の場合，AMOSA-MOCK では大小様々なクラスタが形成されているが，NSGAII-MOCK では，均等にクラスタリングされている．また，SpiralSquare の場合，データ群の作る島ごとにクラスタリングを行えている．

以上より,AMOSA-MOCK は NSGAII-MOCK と比較して精度が悪く，アルゴリズムの改良を検討する必要であることがわかった．AMOSA の特性上，交叉は行わず突然変異のみで次状態を生成するため，次状態の生成部分で改良の余地がある．

6　まとめと今後の課題

多目的最適化問題をクラスタリングに適用した多目的クラスタリングアルゴリズム MOCK は，均質性と分離性の 2 つの評価指標を同時に最適化するアルゴリズムを有する．MOCK では，多目的 GA の SPEAII をベースとしてクラスタリング処理を行うが，クラスタリングの更なる精度向上の為に本研究では多目的 SA を用いて MOCK を適用した．

様々な多目的 SA が考案されているが，他の進化的計算手法と比較して精度の高い AMOSA を利用した．数値実験では，多目的 GA で一般的に用いられている NSGAII や，他の多目的 SA よりも精度が良いことがわかった．しかしながら，MOCK へ AMOSA のアルゴリズムをそのまま適用しても精度の良いクラスタリング結果を得る事ができなかった．それは，AMOSA は初期状態に生成した初期解を探索段階で全て利用するのではなくて，ランダムに選択した 1 点のみを利用して探索を進めている為だと考えられる．

今後は，AMOSA において次状態の生成方法を新たに考案することで，均一でかつ幅広いパレート解を求めるアルゴリズムの考案が必要である．

参考文献

1) 単目的および多目的シミュレーテッドアニーリングにおける温度パラメータ, 實田 健，修士論文，2004.

2) A Simulated Annealing-Based Multiobjective Optimization Algorithm:AMOSA, Sanghamitra Bandyopadhyay, Ujjwal Maulik, and Kalyanmoy Deb, IEEE TRANSACTION ON EVOLUTIONARY COMPUTATION, VOL. 12, NO. 3, JUNE 2008.

3) Julia Handl and Joshua Knowles. Multiobjective clustering with automatic determination of the number of clusters, Technical Report No. TR-COMPSYSBIO-2004-02, UMIST, Department of Chemistry, August 2004.

4) MOCK の大規模データへの適用，真武 信和，廣安 知之，三木 光範，ISDL Report No. 20060712009，2006.
http://mikilab.doshisha.ac.jp/dia/research/report/2006/0712/009/report20060712009.html