計測自動制御学会東北支部 第 328 回研究集会 (2020.3.16)
資料番号 328-2

離散 Data Swarm Clustering による位相保存性の検討

A Study on Topology Preservation in Discrete Data Swarm Clustering

○宮城島 悠太[†]　岩井 俊哉[†]

○Yuta Miyagishima[†],　Toshiya Iwai[†]

† 日本大学

† Nihon University

キーワード：群知能(swarm intelligent), クラスタ分析(cluster analysis), データスワームクラスタリング(data swarm clustering), ボイド(boid), 自己組織化マップ(SOM)

連絡先：〒963-8642 福島県郡山市田村町徳定中河原 1 日本大学大学院工学研究科 情報工学専攻 岩井研究室 岩井俊哉，Tel: 024-956-8819, Fax: 024-956-8863, E-mail: iwai@cs.ce.nihon-u.ac.jp

1. はじめに

群知能とは，鳥や魚などの群れの自己組織的な振る舞いを模倣した人工知能技術の総称である．鳥や魚の群れをなす行動を模倣した粒子群最適化法[1]などの最適化アルゴリズムや鳥の群の振る舞いのシミュレーションモデルとして有名な Boid を応用した Data Swarm Clustering(以下，DSC と略記)[2]などのクラスタ分析法が例として挙げられる．群知能に基づくアルゴリズムは，発見的手法に基づく場合が多く，必ず解が求まる保証がない．しかし，アルゴリズムの簡便さ，多様な問題への適応性及び解探索能力の高さから着目されている．また，自律分散型マルチロボットシステムにおける協調行動に群知能を応用したスワームロボティクス分野で，近年多くの研究が行われている[3]．

高次元特徴ベクトルで表現されるデータオブジェクトを，低次元空間上に写像することで自己組織的にクラスタを形成するクラスタ分析アルゴリズムがある．人工ニューラルネットワークモデルである自己組織化マップ(Self-Organizing Maps, 以下，SOM と略記)[4]，Neural-Gas[5]及び t-distributed Stochastic Neighbor Embetting (t-SNE)[6]が,その代表的なアルゴリズムである.SOM では，高次元データオブジェクトをユニットが配置された低次元空間（以下,マップと呼ぶ）に写像し，各データオブジェクトに対応する Best Matching Unit（以下，BMU と略記）がマップ上に自己組織的に配置され，クラスタが形成される．形成されたマップ空間上の BMU 同士の近隣関係と，対応するデータオブジェクトの特徴ベクトル空間上の近隣関係が一致することが望ましい．この性質を「位相保存性」と呼び，SOM は位相保存性のよいクラスタ分析法の一つである．

また, Veenhuis ら[1]により提案された DSC モデルでは,データオブジェクトを付与された Boid である Datoid が類似したデータオブジェクトを付与された Datoid と群れを形成することで,クラスタが形成され，最終的に同一クラスタに属す Datoid が一点に集まる.この DSC モデルに基づき, Datoid を離散空間に配置させる離散 DSC モデルが考案され，SOM と同程度の位相保存性が示された[7]．

本研究では，離散 DSC モデルの重複回避規則である交換方式を連鎖方式に変更したモデルを提案し，連鎖方式の離散 DSC モデルと交換方式の離散 DSC モデル及び SOM との位相保存性を比較した．

2. DSC モデルの概要[2]

D個の属性を持つデータオブジェクトをD次元特徴ベクトル$\boldsymbol{o}$として表現し，特徴ベクトル空間上のデータオブジェクト$\boldsymbol{o}_i$と$\boldsymbol{o}_j$の距離r_{ij}に基づき$\boldsymbol{o}_i, \boldsymbol{o}_j$間の類似度を式(1)(2)で定義する類似度関数$S(\boldsymbol{o}_i, \boldsymbol{o}_j)$で表す．

$$S(\boldsymbol{o_i}, \boldsymbol{o_j}) = 1 - \frac{r_{ij}}{r_{max}} \qquad (1)$$

$$r_{max} = \max_{i,j} r_{ij} \qquad (2)$$

データオブジェクト$\boldsymbol{o_i}$を付与された離散Datoidの運動するマップ空間上での位置ベクトルを$\boldsymbol{x}_i$で表し，Datoid$\boldsymbol{x}_i$と$\boldsymbol{x}_j$のユークリッド距離$d(\boldsymbol{x}_i, \boldsymbol{x}_j)$を実距離と呼ぶ．実距離の他に，類似度関数を考慮したマップ空間上の Datoid 対の距離として，類似度距離SD_{ij}と非類似度距離DD_{ij}を Datoid の運動規則に用いる．類似度距離と非類似度距離を，それぞれ式(3)(4)に定義する．

$$SD(i,j) = S(\boldsymbol{o}_i, \boldsymbol{o}_j) d(\boldsymbol{x}_i, \boldsymbol{x}_j) + \left(1 - S(\boldsymbol{o}_i, \boldsymbol{o}_j)\right) d_{max} \qquad (3)$$

$$DD(i,j) = \left(1 - S(\boldsymbol{o}_i, \boldsymbol{o}_j)\right) d(\boldsymbol{x}_i, \boldsymbol{x}_j) + S(\boldsymbol{o}_i, \boldsymbol{o}_j) d_{\max} \qquad (4)$$

ここで，d_{max}とはマップ空間の最大距離である．本研究ではマップ空間を正方形とし，d_{max}を正方形の対角線の長さとした．式(3)(4)より，類似度／非類似度距離は Datoid に付与したデータオブジェクト間の類似度が高い／低いと実距離に近づき，類似度が低い／高いと実距離より大きくなる．

Boid の運動規則を構成する 3 つの効果である接近回避，速度合わせ及び結合を，類似度距離と非類似度距離に基づき変更した運動規則に従い Datoid の位置は更新される．以下に，その 3 つの効果について説明する．

①非類似 Datoid への接近回避

着目している i 番目の Datoid に対して非類似度距離で最近傍（非類似最近傍）にいる Datoid が実距離で閾値T_d以内にいるとき，i 番目の Datoid は非類似最近傍の Datoid に対して接近を回避する．その速度の変化分$\Delta V_{separation}$を式(5)に示す．

$$\Delta V_{separation} = \left(1 - S(o_i, o_{n_{i,similar}})\right) W_s \times (x_i - x_{n_{i,dissimilar}}) \quad (5)$$

ここで，添え字$n_{i,dissimilar}$は i 番目の Datoid の非類似最近傍にいる Datoid の識別番号である．また，W_sは調整係数である．

②類似 Datoid との速度合わせ

着目している i 番目の Datoid は，類似度距離で最近傍(類似最近傍)にいる Datoid と速度を合わせる．速度合わせのための速度変化分$\Delta V_{alignment}$を式(6)で表す．

$$\Delta V_{alignment} = W_a (v_{n_{i,similar}} - v_i) \qquad (6)$$

ここで，v_iは i 番目の Datoid の速度ベクトルであり，添え字$n_{i,similar}$は i 番目の Datoid の類似最近傍にいる Datoid の識別番号である．またW_aは調整係数である．

③類似 Datoid の群れへの結合

着目している i 番目の Datoid は，類似度距離が近い K 番目までの Datoid(類似 K 近傍)の中心へ向かう．その速度変化分$\Delta V_{cohesion}$は式(7)で表す．

$$\Delta V_{cohesion} = S(o_i, o_{n_{i,similar}}) W_c R_{0,1} \times (c_{i,similar} - x_i) \quad (7)$$

$$c_{i,similar} = \frac{1}{k} \sum_{k}^{i=1} x_{i_j} \qquad (8)$$

ここで，式(8)で表される$c^{(t)}_{i,similar}$は i 番目の Datoid に対する類似 K 近傍の Datoid の重心を表す位置ベクトルである．式(8)のx_{i_j}は i 番 Datoid の j 番目の類似近傍 Datoid の位置ベクトルである．また$R_{0,1}$は[0, 1]の範囲の一様乱数であり，W_cは調整係数である．

以上の 3 つの効果からなる運動規則より，i 番目の Datoid の速度を式(9)により更新する．また，i 番目の Datoid の位置の更新式を式(10)に示す．

$$v_i^{(t+1)} = w_I^{(t)} v_i^{(t)} + \Delta V_{separation}^{(t)} + \Delta V_{alignment}^{(t)} + \Delta V_{cohesion}^{(t)} \quad (9)$$

$$x_i^{(t+1)} = x_i^{(t)} + v_i^{(t+1)} \qquad (10)$$

ここで，$w_I^{(t)}$は慣性係数であり，本研究ではその値を時間ステップ毎に線形減少させた．

DSC モデルでは Datoid が連続空間内を運動し，類似したデータオブジェクトを付与された Datoid 同士が群れを形成し，クラスタ分析が実現する．しかし，類似した Datoid 同士は無限に接近することが可能なため，形成されたクラスタは一点に収束することが多い．それに対して，SOM のように Datoid 同士の配置の位相保存性が実現されるモデルとして離散 DSC モデルがある．

3. 離散 DSC モデルの概要

マップ空間を離散化して，離散 Datoid を離散空間上に配置するモデルである．本研究では，マップ空間を正方格子で離散化する．離散 Datoid の位置の更新は次の 2 ステップに分かれる．

1) 2 章で説明した Datoid の従う運動規則に基づき連続空間上に離散 Datoid の移動位置を定める．
2)その移動位置に最も近い格子点に離散 Datoid の位置を更新する．

ただし，更新先の格子点に既に他の離散 Datoid がいた場合，離散 Datoid 同士が位置を重複することがないように重複回避処理を施す．本研究で使用する 2 つの重複回避処理について説明する．

既存の離散 DSC モデルでの重複回避処理では，更新した離散 Datoid と更新位置に既にいた離散 Datoid の位置を単に交換する．以下では，この重複回避処理を交換方式と呼ぶ．全離散 Datoid の更新を終えて繰り返し回数を一つ増やす．

本研究で提案する重複回避処理では，更新した離散 Datoid の更新位置にいた離散 Datoid に対して更新を行う．これを，重複がなくなるまで連鎖的に繰り返す．以下では，この重複回避処理を連鎖方式と呼ぶ．この連鎖的に行われる離散 Datoid の更新も含めて全離散 Datoid の位置の更新を終えて繰り返し回数を一つ進める．連鎖方式の離散 DSC モデルのアルゴリズムの擬似コードを図 1 に示す．

```
入力：n Datoid数＝データオブジェクト数
      {o_i} n個のD次元データオブジェクト
      Niter 繰り返し回数
      Ws, Wa, Wc 調整係数
出力：マップ上のDatoidの分布
初期化：n個のDatoidの位置・速度ベクトルをランダムに設定
        繰り返し回数をカウントする変数stepを0に設定
        更新したDatoid数をカウントする変数countを0に設定
        Datoidの識別番号(0～n-1)を配列turnにランダムに格納
while step<Niter do
  count ← 0
  while count < n do
    Datoid_turn[count]の位置を更新
    countを1増やす
    while 重複有 do
      重複したDatoidの位置を更新
      重複したDatoidの識別番号をturn[count]の内容と交換
      countを1増やす
    end
  end
  stepを1増やす
  turnの内容をランダムにシャッフルする
end
```

Fig.1 連鎖方式の離散 DSC モデルのアルゴリズム

図 1 中の配列 turn は 0 から n-1(n は Datoid 数)の数がランダムに格納された大きさ n の配列であり，配列内容の値を識別番号に対応させ，順番に Datoid を更新することで，ランダム非同期式で Datoid を更新している．以下では，離散 Datoid のことを単に Datoid と呼ぶ．

4. 数値実験の内容

離散 DSC モデルと SOM により 3 つのデータセットのクラスタ分析の数値実験を行い，位相保存性を比較した．

3 つのデータセットについて説明する．一つは，自作の Circle データセットである．これは属性数 D=200，データオブジェクト数が 150 であり，データオブジェクトの識別番号が近いほどデータオブジェクト間の類似性が高く，類似性は識別番号に対して周期境界条件を満たす．つまり，類似性の高い Datoid 対を隣接させて Datoid を並べると，識別番号に沿って円環状に Datoid が並ぶ．二つ目は，UC Irvine Machine Learning Repository[8]で提供されている Iris（アヤメ）データセットであり，三種類の Iris の種からそれぞれ 50 個体，合計 150 個のデータオブジェクトから構成されている．各データは 4 つの属性(D=4)を持つ．種の異なる 3 つのクラスへのクラス分類用のベンチマークデータセットであり，各データオブジェクトが属す種を識別するクラスラベルが付与されている．三つ目は，Swiss Roll データセット[9]である．各データオブジェクトは，二次元上の 4 点（7.5,7.5），（7.5,12.5），（12.5,7.5），（12.5,12.5）にピークを持つガウス混合モデルからランダムにサンプリングされて作成された二次元上の点(x, y)を次式で変換した 3 次元空間上の点(X, Y, Z)である．

$$X = x\cos x \tag{11}$$

$$Y = y \tag{12}$$

$$Z = x\sin x \tag{13}$$

従って，属性数(D)は 3 であり，本研究ではデータオブジェクト数を 200 個とした．また，変換前の 2 次元データ点が 4 つのガウス分布のいずれから生成されたか識別するラベルが各データオブジェクトに付与されている．このラベルをクラスラベルと呼び，クラス数は 4 となる．

Iris データセットと Swiss Roll データセットでは，全てのデータオブジェクトがクラスごとにクラスタ分類されるわけではないと考えるが，殆どのデータオブジェクトがクラスラベル毎にクラスタ分類されると期待される．

位相保存性を定量的に評価するために，類似度平均，Goodman-Kruskal の順序連関係数（以下，G-K 係数と略記）及び Spearman の順位相関係数（以下，S 係数と略記）を測定した．

類似度平均は，着目する Datoid とそれに隣接する 8 近傍の Datoid との類似度の平均をとり，さらに全 Datoid に対してそれを平均した量である．G-K 係数は，2 対のデータオブジェクト対の特徴ベクトル空間上での距離の大小関係と対応する 2 対の Datoid 対の実距離の大小関係の一致度を表す測定量である．また，S 係数は，特徴ベクトル空間上のデータオブジェクトの全ペア間距離の順位と対応する Datoid 対の実距離の順位に対する Pearson の積率相関係数である．ただし，これらの評価量は-1 から 1 の値を取り，位相保存性が良いほど大きい値を取る．Datoid の更新等において乱数列を使用しているため，40 回の数値実験で評価量の統計平均を行う．

G-K 係数と S 係数は，Datoid 対の類似度の大小関係の比較に基づく位相保存性の厳密な評価量であるのに比べ，類似度平均では類似した Datoid

が隣接しているかを判定する位相保存性のおおまかな評価量と考えられる．

データオブジェクトの各次元の特徴を平等に評価するために，データオブジェクトの次元成分ごとに最大値を 1，最小値を 0 に正規化するデータの前処理を行った．

Datoid の運動できるフィールドは 100×100 の 2 次元正方格子として，繰り返し回数を 30000 回とした．また，Datoid の更新式(5)-(7)中の調整係数の値を Ws=0.8， Wa=0.1， Wc=0.2 と Ws=0.5， Wa=0.5， Wc=0.5 の二組で数値実験を行った．前者をパラメータセット 812，後者をパラメータセット 555 と呼ぶ．これらの値は位相保存性に関するパラメータ値の数値的解析に基づいて決定したものである．

比較する SOM のマップは Python のライブラリ Somoclu を用いて作成した．繰り返し回数を離散 DSC モデルの数値実験と同じく 30000 回とした．なお，更新に伴いマップが収束することを確認している．Somoclu でも，乱数列を使用しているため 40 回の数値実験で評価量の統計平均を行う．

5. 数値実験の結果と考察

Circle データセットを離散 DSC でクラスタ分析したマップ空間上の Datoid の分布を Fig. 2 に示した．図中の点が Datoid を表し，Datoid に付与したオブジェクトの識別番号順にグラデーションをつけている．左図が連鎖方式で，右図が交換方式の結果である．両図ともに，似た色の Datoid が隣接して円環状に配置されていることが分かる．従って，特徴空間上でのオブジェクトの近隣関係が，マップ空間でもおよそ再現されており，定性的には位相保存が実現していることが分かる．Circle データセットを SOM によりクラスタ分析したマップを Fig.3 に示す．図中の白い点は BMU であり，背景に U-Matrix をモノトーンで描いている．また，BMU の横に対応するデータオブジェクトの識別番号を示した．識別番号が近い BMU がマップ上の近い位置に配置されているが，所々識別番号が近い BMU が遠くに配置されている．従って，Fig. 3 でも定性的には位相保存が実現していることが分かる．3 つのデータセットにおける G-K 係数，S 係数及び類似度平均を，それぞれ Table 1，2，3 にまとめた．Circle データセットでは，離散 DSC モデルでの G-K 係数と S 係数は，SOM より大きな値を取っている．従って，定量的には SOM より離散 DSC モデルでのクラスタ分析の方が高い位相保存性を示したことが分かる．重複回避処理の違いでは，G-K 係数，S 係数の値は交換方式より連鎖方式で大きい値を示した．一方，類似度平均の値は交換方式の方が大きくなった．従って，SOM より離散 DSC モデルのほうが高い位相保存性を示し，離散 DSC モデルの方式では連鎖方式のほうがより高い位相保存性を示すことが分かる．

Iris データセットを離散 DSC でクラスタ分析したマップ空間上の Datoid の分布を Fig. 4, 5 に示した．Fig. 4, 5 の左図が連鎖方式，右図が交換方式の結果である．Fig. 4 では図中の点が Datoid を表し，クラスごとに色をつけている．同一の色の Datoid が隣接してクラス毎に集合していることが分かるが，赤と緑のクラスが一つのクラスタ内に配置してしまった．Fig. 5 では，識別番号 1 のデータオブジェクトに対する類似度の値に比例させて Datoid にグラデーションをつけている．Fig.4 で赤と緑で色づけられているクラスタ内の Datoid が，Fig.5 では明確に色で分類できていないことが分かる．つまり，データオブジェクトの類似性からこの 2 つのクラスを明確にクラスタ分類できないと考えられる．しかし，Fig. 4, 5 では，似た色の Datoid がマップ上に隣接して配置されていることから，定性的に位相保存が実現されていることが分かる．Iris データセットを SOM によりクラスタ分析したマップを Fig. 6 に示す．図中の点は BMU であり，背景に U-Matrix をモノトーンで描いている．また，BMU をクラスごとに色づけている．Fig.6 の BMU と Fig.4 の Datoid の属すクラスは同一の色で描いた．似た色の BMU がマップ上に隣接して配置されていることから，定性的に位相保存が実現されていることが分かる．Table 1 より，Iris データセットにおいて連鎖方式のパラメータセット 812 以外の離散 DSC モデルでの G-K 係数の値は SOM での値よりも大きい値をとった．また，交換方式の G-K 係数の値は，連鎖方式の値より大きくなった．S 係数では，離散 DSC モデルの両方式での値が SOM での値より大きくなった．一方，Circle データセットと同様に，S 係数の値は連鎖方式の方が交換方式より大きい値を示した．類似度平均は交換方式のほうが連鎖方式に比べ高い値をとった．このことから，Iris データセットでは離散 DSC モデルのほうが SOM より高い位相保存性を示すが，離散 DSC モデルでの交換方式と連鎖方式のいずれの位相保存性が高いか判断しがたい．

Swiss Roll データセットを離散 DSC でクラスタ分析したマップ空間上の Datoid の分布を Fig. 7, 8

に示した．Fig. 7 と 8 のマップの描写方法は，それぞれ Fig. 4 及び 5 と同一である．Fig. 7, 8 ともに,似た色の Datoid がマップ上の隣接した位置に配置されていることから，定性的に位相保存が実現されていることが分かる． Swiss Roll データセットを SOM によりクラスタ分析したマップを Fig. 9 に示す．Fig. 9 のマップの描写方法は，Fig. 6 と同一である． 似た色の BMU がマップ上に隣接して配置されていることから，定性的に位相保存が実現されていることが分かる．Swiss Roll データセットでの G-K 係数と S 係数の値は離散 DSC モデルの交換方式が最も大きい値をとった．類似度平均も交換方式のほうが大きい値を示した．このことから，Swiss Roll データセットでは離散 DSC モデルの交換方式が高い位相保存性を示すことが分かる．

6. まとめ

本研究で用いたデータセットに対するクラスタ分析では，離散 DSC モデルの連鎖方式・交換方式ともに SOM よりも高い位相保存性を示した．しかし，データセットによって位相保存性が高い方式が異なることが分かった．また，調整係数の値によっても位相保存性の善し悪しが異なる．

交換方式は単純なアルゴリズムであるが，類似度平均が大きいことより,類似性の近い Datoid をマップ上の近隣に集める能力が高いと思われる．連鎖方式と交換方式で，位相保存性に顕著な差が見出せなかった．従って，連鎖方式に比べ交換方式の計算時間が少ないことから，交換方式により効率的なクラスタ分析ができていると考えられる．

7. 今後の課題

本研究では離散 DSC モデルと SOM により位相保存性の比較を行ったが，Neural-Gas や t-SNE との比較を行う．離散 DSC モデルでの位相保存性のよい調整係数の推奨値を求め，交換方式と連鎖方式を比較する．異なるデータセットにも適用する．

文　献

1) J. Kennedy and R. C. Eberhart, “Swarm Intelligence”, Morgan Kaufmann Publishers, San Francisco, California
2) アジス・アブラハムら編，「群知能とデータマイニング」，255－278 頁，東京電機大学出版（2012）.
3) 田村康将ら，特集「生物の群行動に学ぶロボットシステム」，計測と制御，87－144 頁，

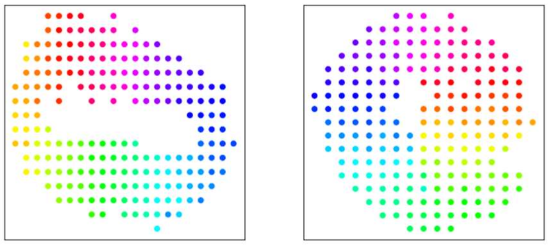

Fig. 2 識別番号順に色付けした Datoid の分布（Circle データセット，パラメータ 555）

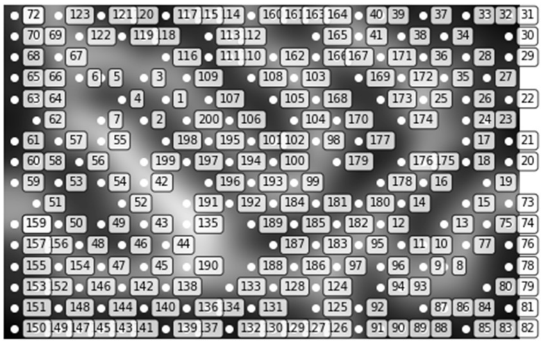

Fig. 3 SOM によるマップ（Circle データセット）

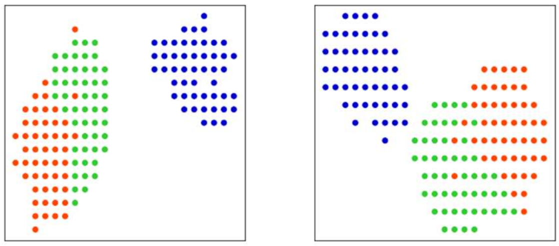

Fig. 4 クラス毎に色付けした Datoid の分布（Iris データセット，パラメータ 555）

Fig. 5 類似度で色付けした Datoid の分布（Iris データセット，パラメータ 555）

（2020）.
4) Kohonen T, “Self-Organizing Maps”. Springer-Verlag, Berlin Heidelberg (1995).
5) T. Martinetz and K. Schulten, “A “Neural-Gas” Network Learns Topology”, Artificial Neural Networks, Elsevier(North-Holland), pp. 397-402 (1991).
6) L. Maaten and G. Hinton, “Visualizing Data using t-SNE”, Journal of Machine Learning Research, Vol. 9, pp. 2579-2605 (2008).
7)K. Yoshida and T. Iwai, “Topology preservation in discrete data swarm clustering”, The proceedings of AROB, Vol. 24, pp. 1082-1086 (2019).
8)UC Irvine Machine Learning Repository https://archive.ics.uci.edu/ml/index.php （2020.3.11 アクセス）.
9)http://people.cs.uchicago.edu/~dinoj/manifold/swissroll.html （2020.3.11 アクセス）.

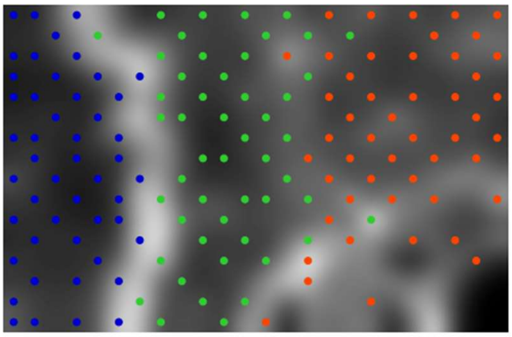

Fig. 6 SOM によるマップ（Iris データセット）

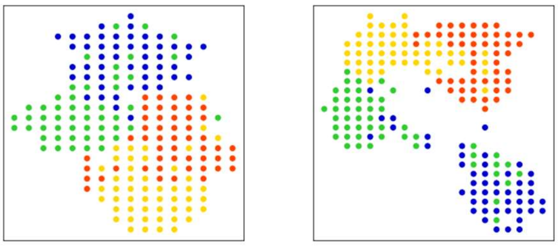

Fig. 7 クラス毎に色付けした Datoid の分布
（Swiss Roll データセット，パラメータ 555）

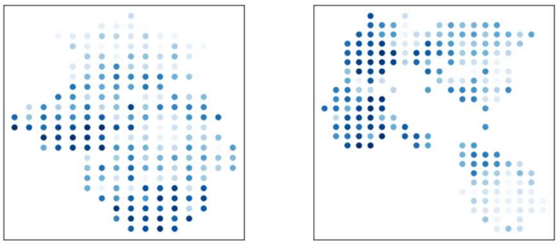

Fig. 8 類似度で色付けした Datoid の分布
（Swiss Roll データセット，パラメータ 555）

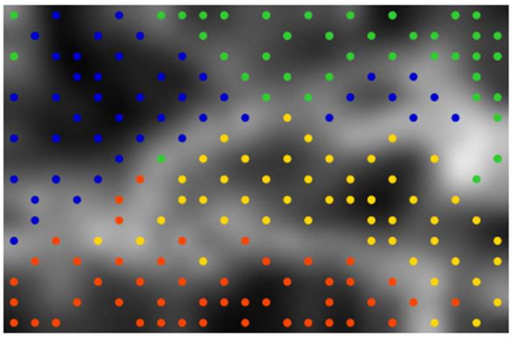

Fig. 9 SOM によるマップ
（Swiss Roll データセット）

Table 1 GK 係数の比較

	連鎖方式		交換方式		SOM
	555	812	555	812	
Circle	**0.325**	0.318	0.314	0.309	0.196
Iris	0.549	0.489	0.573	**0.575**	0.512
Swiss Roll	0.281	0.392	0.345	**0.493**	0.454

Table 2 S 係数の比較

	連鎖方式		交換方式		SOM
	555	812	555	812	
Circle	**0.389**	0.354	0.245	0.248	0.243
Iris	0.815	**0.847**	0.781	0.763	0.69
Swiss Roll	0.419	0.395	0.49	**0.665**	0.626

Table 3 類似度平均の比較

	連鎖方式		交換方式	
	555	812	555	812
Circle	0.571	0.580	**0.875**	0.872
Iris	0.863	0.875	0.898	**0.903**
Swiss Roll	0.726	0.795	0.845	**0.852**