研究速報

自己組織化マップに基づくデータの連続性を考慮したクラスタリング手法

A Clustering Method Considering Continuity of Data Based on Self-Organizing Map

正会員　今村弘樹[†]，　藤村　誠[†]，正会員　黒田英夫[††]

Hiroki Imamura[†], Makoto Fujimura[†] and Hideo Kuroda[††]

Abstract　For accurate clustering when clusters are close to each other, we developed a method considering the continuity of data based on a self-organizing map.

キーワード：クラスタリング，自己組織化マップ，データの連続性

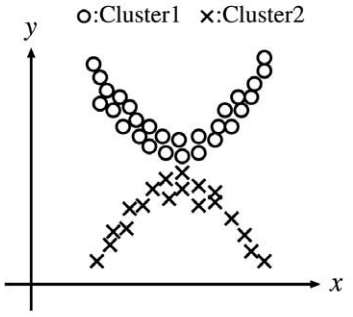

図 1　帯状に連続的な分布をするデータの例
An example of data which continuously distributes like a belt.

1. まえがき

クラスタリング手法の主な手法として，k-means 法[1)〜6)]，fuzzy-c 平均法[7)〜11)]がある．これらの手法は，クラスタリングするデータの分布が正規分布状であることが前提となっている．これに対して，クラスタリングするデータの分布が任意の形状においても高精度にクラスタリングを行うために，任意形状クラスタリング法が提案された[12)13)]．

2006 年 3 月 31 日受付，2006 年 5 月 17 日再受付，2006 年 5 月 25 日採録
†長崎大学 工学部 情報システム工学科
（〒 852–8521 長崎市文教町 1–14, TEL 095-819-2574）
†長崎大学 大学院　生産科学研究科
（〒 852–8521 長崎市文教町 1–14, TEL 095-819-2574）
†Dept. of Computer and Information Sciences, Nagasaki University (1–14, Bunkyou-mach, Nagasaki City, 852–8521)
†Graduate School of Science and Technorogy, Nagasaki University (1–14, Bunkyou-mach, Nagasaki City, 852–8521)

しかし，異なるクラスタが接近している場合，任意形状クラスタリング手法を用いても，本来異なるクラスタが同じクラスタにクラスタリングされてしまうことがある．

ここで，動画像シーケンスを各シーン毎にクラスタリングする場合や照明条件が変化する画像をクラスタリングする場合，各クラスタ毎のデータが，特徴空間において帯状に連続的な分布をすると考えられる (図 1)．そこで，データの連続性を考慮することにより，これらのデータをクラスタ毎に分類できることが期待できる．

ここで，データの大まかな形状を抽出できる自己組織化マップ (SOM)[3)] がある．SOM はコードベクトルをデータにフィッティングさせることにより，データの大まかな形状を抽出する．このため，特に非線形な分布データにおいて，主成分分析や判別分析などに比べ，データの大まかな形状を抽出するのに有効である[4)]．

我々は，この SOM を用いて，コードベクトルをデータにフィッティングさせ，このコードベクトルの連続性に基づき，クラスタリングする手法を提案する．この手法により，異なるクラスタが接近している場合においても，データの分布が帯状に連続的な分布であれば，精度良くクラスタリングできることを実験により示す．

2. 提案手法

ここでは，まず，提案手法の基礎となる SOM のアルゴリズムを示す．次に，SOM に基づく提案手法のアルゴリズムを示す．なお，提案手法では，1 次の SOM を用いることとする．

2.1 SOM のアルゴリズム

1.コードベクトル $\boldsymbol{m}_i(i = 1, 2, \cdots, N_v)$ を定義域内においてランダムに生成する．

映像情報メディア学会誌 Vol. 60, No. 8, pp. 1312〜1316 (2006)

2.以下を $t = 1,2,\cdots,T; n = 1,2,\cdots N$ について繰り返す．

3.式 (1) を満たす i^*

$$i^* = \arg \min_{1 \leqq i \leqq N_v} ||\boldsymbol{m}_i - \boldsymbol{x}_n|| \tag{1}$$

をデータ $\boldsymbol{x}_n$ に対する勝利ユニットとする．

4.勝利ユニットとその周辺のユニットを

$$\boldsymbol{m}_i := \begin{cases} \boldsymbol{m}_i + \alpha(t)\{\boldsymbol{x}_n - \boldsymbol{m}_i\} \\ \quad \text{if } i \in i^* + N_c(t) \\ \\ \boldsymbol{m}_i \\ \quad \text{if } i \notin i^* + N_c(t) \end{cases} \tag{2}$$

により更新する．ただし，

$$\alpha(t) = \frac{0.7}{1 + [t/7]} \tag{3}$$

$$N_c(t) = \left\{ i : |i| \leqq \left[14 \exp\left(-\frac{t^2}{50} \right) \right] \right\} \tag{4}$$

とし，$[q]$ は，q を超えない最大の整数を表す．また，“:=”は右辺を左辺に代入することを表し，“=”は右辺と左辺が等しいことを表す．それぞれを明確に表すためにこれらの記号を用いる．

2.2 提案手法のアルゴリズム

1.SOM を実行し，

$$E(\boldsymbol{m}_i(t)) = \arg \max_{1 \leqq i \leqq N_v} ||\boldsymbol{m}_i(t) - \boldsymbol{m}_i(t-1)|| \quad (t \geqq 2) \tag{5}$$

が

$$E(\boldsymbol{m}_i(t)) < Th_1 \tag{6}$$

となった時点で $\boldsymbol{m}_i(t)$ の更新を停止する．ただし，式 (5) における $\boldsymbol{m}_i(t)$ を t 回目の繰り返し計算における $\boldsymbol{m}_i$ とし，Th_1 は，式 (6) における閾値を表す．式 (5) は，各 $\boldsymbol{m}_i$ の更新時における変化の最大値を表し，その変化が式 (6) における閾値以下となったらすべての $\boldsymbol{m}_i$ の変化がなくなったとみなし，SOM の処理を終了する．

2.$\boldsymbol{m}_i$ のラベル値を $l(\boldsymbol{m}_i)$ とし，$l(\boldsymbol{m}_i) := 0 (i = 1,2,\cdots,N_v)$ とする．また，$\boldsymbol{m}_i$ の除外フラグを $e(\boldsymbol{m}_i)$ とし，$e(\boldsymbol{m}_i) := 0 (i = 1,2,\cdots,N_v)$ とする．まず，初期値として $i := 1$，$l(\boldsymbol{m}_1) := 1$ と設定する．

3.$i+1 \leqq N_v$ かつ，$l(\boldsymbol{m}_{i+1})$ が 0 ならば，$l(\boldsymbol{m}_{i+1}) := l(\boldsymbol{m}_i)$ とする．

4.$i+2 \leqq N_v$ かつ，

$$\frac{(\boldsymbol{m}_{i+2} - \boldsymbol{m}_{i+1})^T (\boldsymbol{m}_{i+1} - \boldsymbol{m}_i)}{||\boldsymbol{m}_{i+2} - \boldsymbol{m}_{i+1}|| ||\boldsymbol{m}_{i+1} - \boldsymbol{m}_i||} < Th_2 \tag{7}$$

かつ，

$$||\boldsymbol{m}_{i+2} - \boldsymbol{m}_{i+1}|| - ||\boldsymbol{m}_{i+1} - \boldsymbol{m}_i|| > 0 \tag{8}$$

ならば，$l(\boldsymbol{m}_{i+2}) := l(\boldsymbol{m}_{i+1}) + 1$ とする．ただし，式 (7) の左辺は，ベクトル $(\boldsymbol{m}_{i+2} - \boldsymbol{m}_{i+1})$ と $(\boldsymbol{m}_{i+1} - \boldsymbol{m}_i)$ のなす角における cos の値を表し，Th_2 は，式 (7) における閾値を表す．$\boldsymbol{m}_{i+2}$ が $\boldsymbol{m}_{i+1}$ と $\boldsymbol{m}_i$ と異なるクラスタであれば，$\boldsymbol{m}_{i+2}$ と $\boldsymbol{m}_{i+1}$，$\boldsymbol{m}_{i+1}$ と $\boldsymbol{m}_i$ の間が不連続となり，式 (7) で，左辺が閾値以下となる．また，その際に式 (8) に示すように，$\boldsymbol{m}_{i+2}$ と $\boldsymbol{m}_{i+1}$ 間の距離が $\boldsymbol{m}_{i+1}$ と $\boldsymbol{m}_i$ 間の距離よりも大きくなる．

5. $$E(\boldsymbol{x}_a) = \arg \min_{1 \leqq a \leqq N} ||\boldsymbol{x}_a - \boldsymbol{m}_i|| \tag{9}$$

を満たす $\boldsymbol{x}_a$ に対して，

$$||\boldsymbol{x}_a - \boldsymbol{m}_i|| > Th_3 \tag{10}$$

ならば，$e(\boldsymbol{m}_i) := 1$ とする．ただし，Th_3 は，式 (10) における閾値とする．式 (9) により $\boldsymbol{m}_i$ と距離が最も近い $\boldsymbol{x}_a$ を抽出し，式 (10) で，この $\boldsymbol{x}_a$ と $\boldsymbol{m}_i$ との距離が閾値以上ならば，データ $\boldsymbol{x}_a$ から孤立した $\boldsymbol{m}_i$ として，7. のラベリングの際に $\boldsymbol{m}_i$ を除外する．

6.もし，$i = N_v$ ならば，7. へ，そうでなければ，$i := i+1$ として，3. へ．

7.各 $\boldsymbol{x}_u (u = 1,2,\cdots,N)$ に，

$$E(\boldsymbol{m}_j) = \arg \min_{1 \leqq j \leqq N_v} ||\boldsymbol{x}_u - \boldsymbol{m}_j||, \text{ ただし} e(\boldsymbol{m}_j) = 0 \tag{11}$$

を満たす $\boldsymbol{m}_j$ における $l(\boldsymbol{m}_j)$ をそれぞれ割当てる．式 (11) で $\boldsymbol{x}_u$ と距離が最も近い $\boldsymbol{m}_j$ を抽出し，$\boldsymbol{m}_j$ のラベル値を $\boldsymbol{x}_u$ に割当てる．

3. 実　験

提案手法の有効性を評価するために，コンピュータにより生成した人工データに対して，クラスタリングを行った．ここで，提案手法との比較のために，任意形状に対して高精度に，かつ，高速にクラスタリング可能な文献[13]の手法を用いた．実験で使用した人工データは，

$$y = \alpha(x - \beta)^2 + \gamma \tag{12}$$

を満たす 2 次元の点 (x, y) に対して，生起確率が

$$p(e) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{e^2}{2\sigma^2} \right) \tag{13}$$

をみたす e_x，e_y をそれぞれ，x，y に加えたものとする．ただし，ここでは，クラスタ数を 2 とし，クラスタ 1 に含まれるデータの式 (12) における各パラメータはそれぞれ，$\alpha = 1.0$，$\beta = 0.0$，$\gamma = 0.25$ とし，クラスタ 2 に含まれるデータの式 (12) における各パラメータはそれぞれ，$\alpha = -1.0$，$\beta = 0.0$，$\gamma = -0.25$ とした．また，クラスタ 1，クラスタ 2 共に，式 (13) における $\sigma = 0.5$ とし，クラスタ 1，クラスタ 2 におけるデータ数は共に 50 とした．実際に用いた人工データを図 2 に示す．また，文献[13]の手法のパラメータ σ は，18.0 とし，提案手法の SOM において，Nc=3，Th_1=5.0×10^{-3} とし，用いたコードベクトル数は 50，コードベクトル配置の定義域は $-5 \leqq x \leqq 5$，$-5 \leqq y \leqq 5$

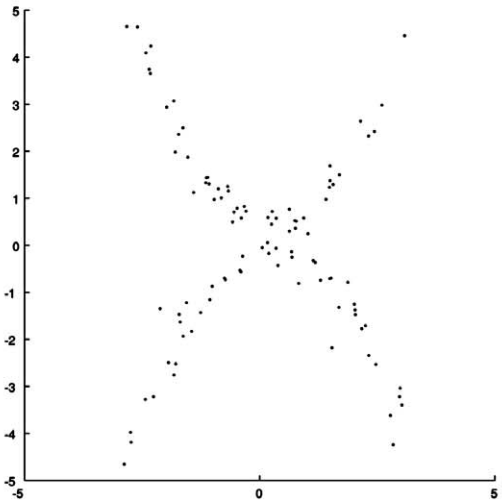

図 2　生成した人工データ#1
Generated synthesis data#1.

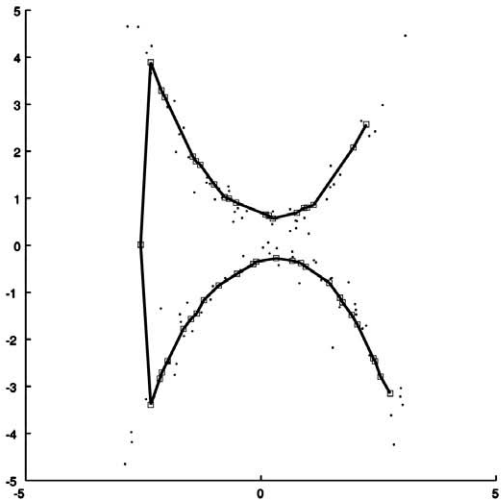

図 4　SOM によるコードベクトルのマッチング結果#1
The matching result#1 of code vectors by SOM.

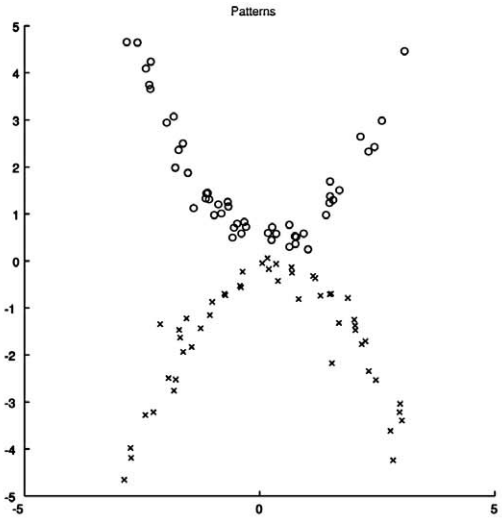

図 3　正解のクラスタリング結果#1
The correct clustering result#1.

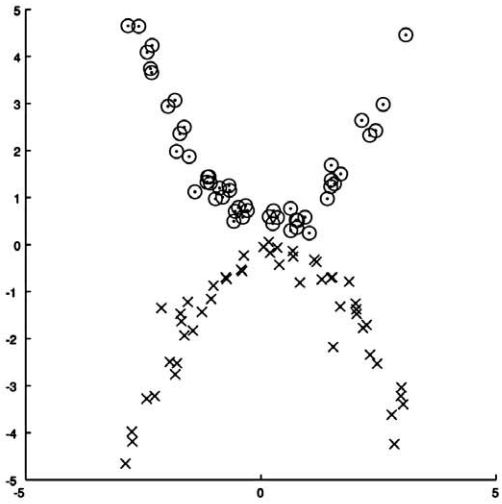

図 5　提案手法によるクラスタリング結果#1
The clustering result#1 by the proposed method.

とした．また，提案手法のクラスタリングにおけるパラメータは，それぞれ，Th_2=cos(π/2)，Th_3=1.0 とした．

図 3 に，正解のクラスタリング結果，図 4 に SOM によるコードベクトルのマッチング結果，図 5 に提案手法によるクラスタリング結果，図 6 に文献 [13] のクラスタリング結果を示す．提案手法によるクラスタリング結果は，正解のクラスタリング結果と同等な結果であった．これは，SOM によるコードベクトル（図中の連結線上の点）が，各クラスタの大まかな形状にマッチングでき，このコードベクトルの連続性に基づき，クラスタリングした後，これらのコードベクトルとの距離により，データをクラスタリングしたため，良好な結果が得られたと考える．これに比べ，文献 [13] の手法によるクラスタリング結果は，全データが一つのクラスタにクラスタリングされた，これは，各クラスタ内のデータの距離に対して，異なるクラスタ間の距離が短いためであると考える．

次に，図 2 のデータと異なる図 7 に示すデータに対して，クラスタリングを行った．　図 7 に，生成した人工データ，図 8 に，正解のクラスタリング結果，図 9 に SOM によるコードベクトルのマッチング結果，図 10 に提案手法によるクラスタリング結果，図 11 に文献 [13] のクラスタリング結果を示す．文献 [13] の手法のパラメータは，前の実験と同じ値とした．提案手法では用いたコードベクトル数を 36 とし，それ以外のパラメータは，前の実験と同じ値とした．文献 [13] のクラスタリング手法に比べ，提案手法によるクラスタリングは，良好にクラスタリングできていることがわかる．ただし，最適なコートベクトル数は，前の実験と異なる数となった．したがって，適用するデータ毎に最適なコードベクトル数は異なることがわかる．また，提案手法によるクラスタリング結果では，右側のクラスタの一部に，若干エラーが生じているのがわかる．しかし，あるデータのラベルとそのデータと最短距離となるデータのラベルが異なれば，それぞれのデータの近傍データのラベルを調べ，異なるラベルの少ないほうのデータのラベルをもう一方の

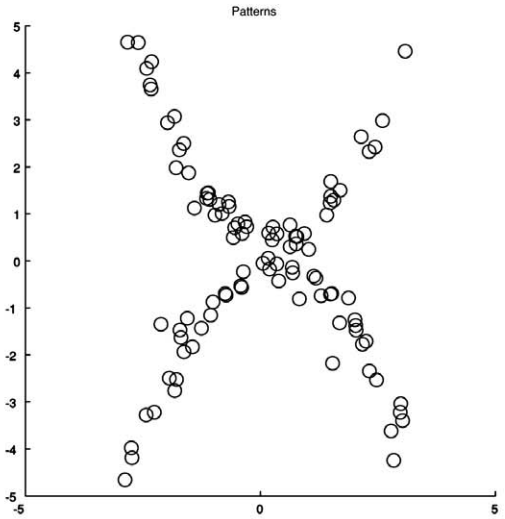

図 6　文献 [13] の手法によるクラスタリング結果#1
The clustering result#1 by the method in literature[13].

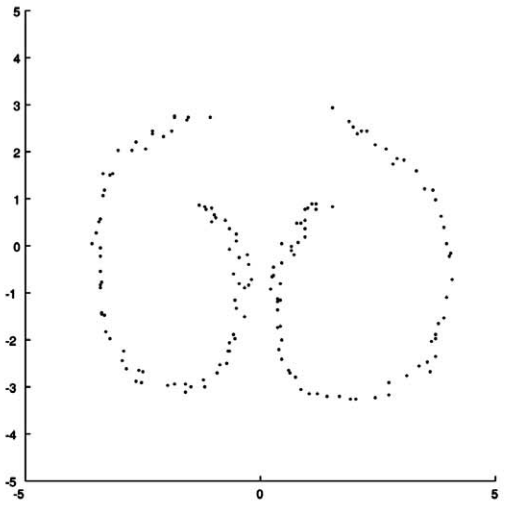

図 7　生成した人工データ#2
Generated synthesis data#2.

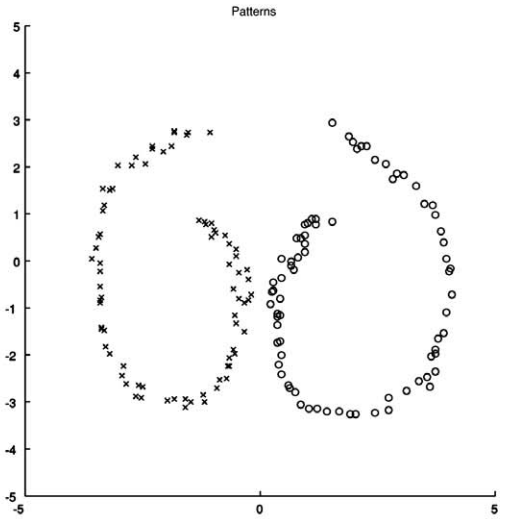

図 8　正解のクラスタリング結果#2
The correct clustering result#2.

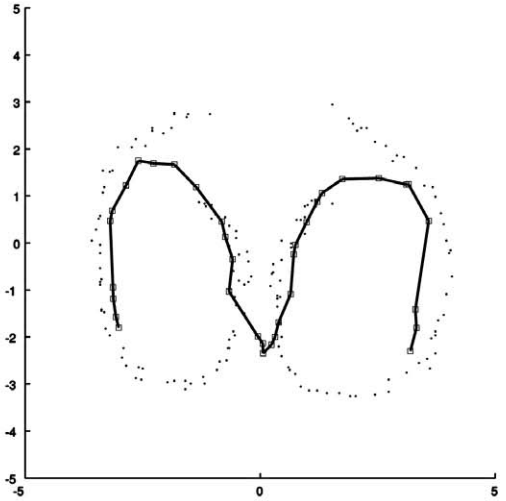

図 9　SOM によるコードベクトルのマッチング結果#2
The matching result#2 of code vploectors by SOM.

データのラベルに置き換えるなどの処理を行うことにより，このようなエラーを削減できると考える．

4. ま と め

異なるクラスタが接近している場合においても精度良くクラスタリングするために，自己組織化マップに基づくデータの連続性を考慮したクラスタリング手法を提案した．

提案手法により，異なるクラスタが接近している場合においても，データの分布が帯状に連続的な分布であれば，精度良くクラスタリングできることを実験により示した．

今後は，動画像シーケンスのフレームを各シーン毎にクラスタリングや照明条件が変化する画像のクラスタリングなど，実際のデータに対して提案手法を適用し，その有効性を評価する．

〔文　献〕

1）B. T. Cover and P. Hart, “Nearest Neighbor Classification,” IEEE Trans. on Information Theory, IT–13, 1, pp.21–27 (1967)
2）K. Fukunaga, “Introduction to statistical pattern recognition,” Academic press, Boston, 2 edition (1990)
3）R. O. Duda, P. E. Hart and D. G. Stocrk, “Pattern Clasification – Second Edition,” Wiley Interscience (2002)
4）J. Mao and A. K. Jain, “A Self-organizing Network for Hyper-ellipsoidal Clustering (HEC),” IEEE Trans. on Neural Networks, TNN-7(1):16-29 (1996)
5）井上 光平, 浦浜 喜一, “次元削減に基づくフィルタリングによる kNN 識別の高速化,” 信学論, J85–D–II, 5, pp.950–953 (2002)
6）春日 秀雄, 山本 博章, 岡本 正行, “高速 K-means 法を用いたカラー画像の色量子化,” 信学論, J82–D–II, 7, pp.1120–1128 (1999)
7）R. L. Cannon, J. V. Dave, and J. C. Bezdek, “Efficient implementation of the fuzzy c-means clustering algorithms,” IEEE Trans. on PAMI, 8, 2, pp.248–255 (1986)
8）R. E. Hammah and J. H. Curran, “Validity Measures for the Fuzzy Cluster Analysis of Orientations,” IEEE Trans. on PAMI, 22, 12, pp.1467–1472 (2000)
9）A. Keller and F. Klawonn “Fuzzy Clustering with Weighting of Data Variables,” Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 8, 6, pp.735–746 (2000)
10）高橋 正人, 服部 和雄, “ファジー c-means 法と最近傍決定則を用いたクリスプなクラスタリング法,” 信学論, J83–D–II, 9, pp.1957–1961 (2000)
11）井上 光平, 浦浜 喜一, “緩和反復法に基づくロバストファジークラスタリング,” 信学論, J85–D–II, 6, pp.1140–1143 (2002)
12）井上光平, 浦浜喜一 “データ間連結度に基づく任意形状ファジークラスタの抽出,” 信学論, J86-D-II, 10, pp.1511-1513 (2003)

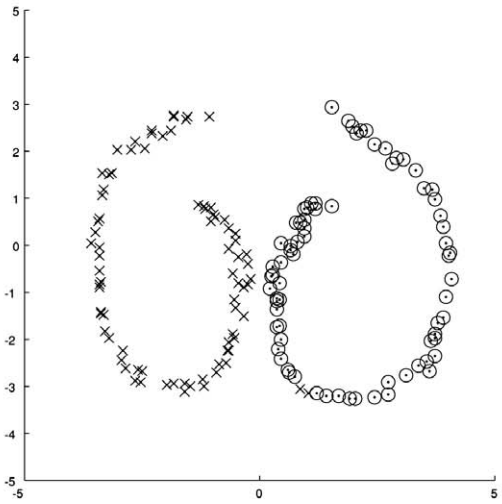

図 10 提案手法によるクラスタリング結果#2
The clustering result#2 by the proposed method.

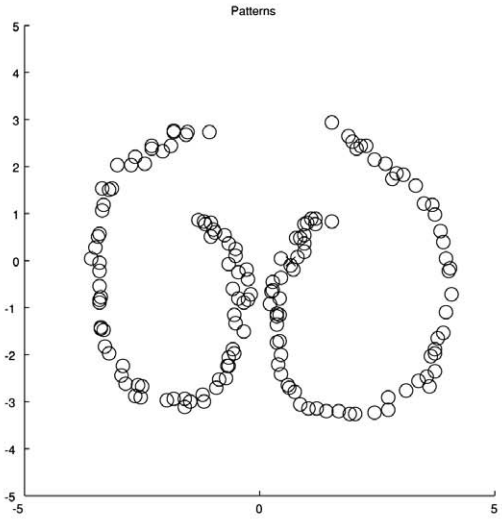

図 11 文献 [13] の手法によるクラスタリング結果#2
The clustering result#2 by the method in literature[13].

13) 今村 弘樹, 藤村 誠, 黒田 英夫, “クラスタ間距離の昇順によるラベリングに基づくノイズにロバストな任意形状クラスタリング,” 映情学誌, 60, 4, pp.618–620 (2006)

今村 弘樹 (いまむら ひろき) 1997 年，創価大学工学部情報システム学科卒業．2002 年，米国カーネギーメロン大学ロボティクス研究所訪問研究員．2003 年，北陸先端科学技術大学院大学情報科学研究科博士課程修了．同年，長崎大学工学部情報システム工学科助手，現在に至る．博士（情報科学）．画像処理，パターン認識，コンピュータグラフィックスの研究に従事．正会員．

藤村 誠 (ふじむら まこと) 1985 年，福井大学工学部卒業．同年，FHL に入社．1990 年，長崎大学工学部助手，1994 年，同講師，現在に至る．動画像の高性能符号化，画像処理などの研究に従事．

黒田 英夫 (くろだ ひでお) 1971 年，九州工業大学大学院修士課程修了．同年，日本電信電話公社電気通信研究所に入社．1989 年，長崎大学工学部大学院教授．その間，1994 年，シドニー大学客員教授，現在に至る．工学博士．画像信号高能率符号化，画像処理，CG，CV 等の研究に従事．正会員．