研究速報

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

Robust Method for Clustering Arbitrarily-shaped Clusters Based on Labeling by Ascending Order Distance between Clusters

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　The conventional robust method for clustering arbitrarily-shaped clusters takes a long time to process. To reduce the processing time, we developed a robust method for clustering arbitrarily-shaped clusters based on labeling by ascending order the distance between clusters.

キーワード：クラスタリング，　クラスタ間距離，　昇順

1. まえがき

任意形状のクラスタリング手法として，自己組織化特徴マップ (SOM) を用いる方法[1]やEMアルゴリズムを用いる手法[2]などが提案されているが，これらは，反復的な手法なので，初期値やノイズデータへの依存性が高いといった課題がある．これらの課題を克服するために，データ間連結度に基づく任意形状ファジィクラスタの抽出法[4]が提案されている．この手法は，クラスタリングする全データの連結度を算出するために，あるデータからあるデータに至る全経路を探索する必要があり，非常に計算コストが大きい．本研究では，クラスタ間距離の昇順に逐次的なクラスタリングをするアプローチをとることにより，計算コストの小さい，ノイズにロバストな任意形状クラスタリング手法を提案する．

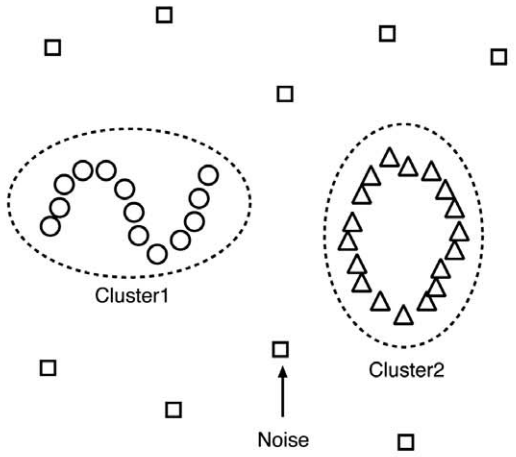

図 1　ノイズを含む場合の任意形状クラスタの例
An example of arbitarily shaped clusters in case of including noise.

2. 提案手法

2.1 提案手法の概要

まず，提案手法の概要を示す．ノイズを含む場合の任意形状クラスタ (図 1) の特徴を以下にまとめる．

- 同じクラスタにおける隣接データ間の距離は小さい．
- 異なるクラスタにおけるデータ間の距離は大きい．
- ノイズのデータと各クラスタにおけるデータ間の距離は大きい．

そこで，提案手法のアプローチとして，データ間の距離が小さい順に各データに同じラベルを与える．このときに，ラベル値の大きい方のラベルを，ラベル値の小さい方のラベルに置き換える．この処理を繰り返し，データ間の距離が一定の距離以上となった場合にラベリングを終了し，ラベルごとにクラスタリングすることとする．従来手法では，まず，各データの連結度を計算するために，あるデータからあるデータへ至る全経路を探索しなくてはならず，膨大な処理時間を必要としたが，提案手法では，逐次的にクラ

2005 年 11 月 2 日受付，2006 年 1 月 10 日再受付，2006 年 1 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)

スタリングを行うために，比較的短時間でのクラスタリングが期待できる．

2.2 提案手法のアルゴリズム

以下に，提案手法のアルゴリズムを示す．

クラスタリングするデータ $\boldsymbol{p}_i$ を i 番目の r 次元ベクトルとし，全データ数を m とする．

1.$i := 0, D_{min}^{(i)} = 0$ とし，$1, 2, \cdots, m$ の全データにそれぞれ，$1, 2, \cdots, m$ の番号でラベリングし，すべてのデータにおける処理済フラグを 0 とする．

2.$i := i + 1$ とし，

$$D_{min}^{(i)} = \arg \min_{1 \leqq j \leqq m, 1 \leqq k \leqq m, j \neq k} ||\boldsymbol{p}_j - \boldsymbol{p}_k|| \qquad (1)$$

を決定する．ただし，式 (1) における $\boldsymbol{p}_j$，$\boldsymbol{p}_k$ は，それぞれ異なるラベル値を持つものとする．

3.$i \geqq 2$，かつ，$|D_{min}^{(i)} - D_{min}^{(i-1)}| > Th_{(i)}$ ならば 5. へ．それ以外なら 4. へ．

4.異なるラベル値を持ち，データ間のユークリッド距離が $D_{min}^{(i)}$ となるデータのペアを選ぶ．これらのデータの処理済みフラグをそれぞれ 1 とし，ラベル値の大きいラベル値を a，ラベル値の小さいラベル値を b とする．全データの中で，ラベル値 a を持つデータのラベル値をすべてラベル値 b に変更し，2. へ．

5.終了．

ここで，3. における $Th_{(i)}$ は，

$$Th_{(i)} = E_{(i)} + \sigma V_{(i)} \qquad (2)$$

とする．ただし，σ は，パラメータで，クラスタリングするデータに応じて予め与えておくものとし，$E_{(i)}$，$V_{(i)}$ はそれぞれ，

$$E_{(i)} = \frac{1}{i} \sum_{t=1}^{i} D_{min}^{(t)} \qquad (3)$$

$$V_{(i)} = \frac{1}{i} \sum_{t=1}^{i} (E_{(i)} - D_{min}^{(t)})^2 \qquad (4)$$

とする．ただし，式 (3) は，i 個のデータにおける $D_{min}^{(i)}$ の平均を表し，式 (4) は，i 個のデータにおける $D_{min}^{(i)}$ の分散を表す．

3. 実　験

提案手法の有効性を評価するために，従来手法と提案手法を用いて，**図 2** に示す 180 点からなる 2 次元のデータに対してクラスタリングを行った．ここでは，従来手法として，ガウス分布状のデータにおいて，ノイズにロバストなクラスタリングが可能な手法 [3] と従来のノイズにロバストな任意形状クラスタリング手法 [4] を用いた．クラスタリング結果を**図 3** から**図 5** に示す．なお，この実験において設定した各手法のパラメータは，文献 [3] と文献 [4] における α は 1.0，クラスタ決定におけるクラスタ代表点との類似度 s の閾値は 0.5，文献 [3] における ϵ は 0.05，文献 [4] における ϵ は 0.005，提案手法における σ は 1.0 とした．図中の ×，○，＋ は，それぞれクラスタ 1 からクラスタ 3 のデータを表し，□は，ノイズデータを表す．提案手法では，処理済みフラグが 0 のデータをノイズデータとした．**表 1** に各手法の実行時間を示す．なお，表中の実行時間は，CPU が PentiumIII の 1.0GHz，メモリーが 512MB，OS が WindowXP のスペックを持つ PC 上で，数値計算アプリケーションの MATLAB を用いて実行したものとする．

文献 [3] の手法は，実行時間は，他の手法に比べ非常に短いが，クラスタリング結果は，非常に悪いことがわかる．これに比べ，文献 [4] の手法と提案手法のクラスタリング結果

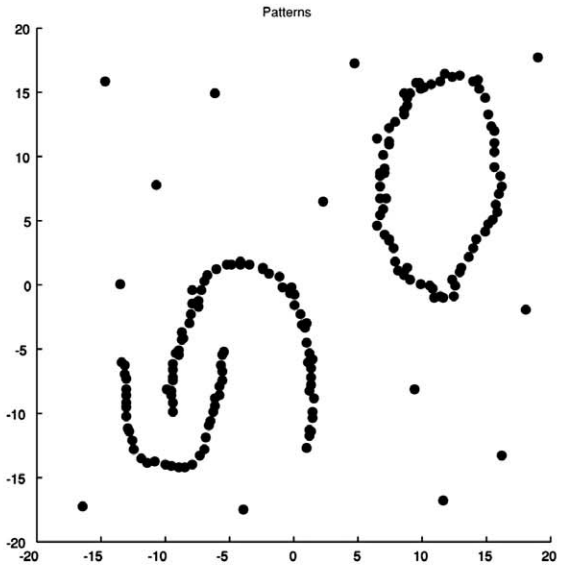

図 2 実験に用いたデータ
Data used in the experiment.

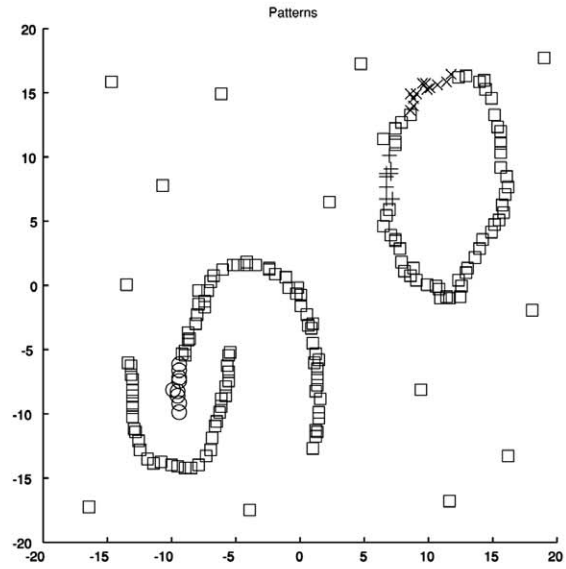

図 3 文献 [3] の手法による結果
The result by the method of reference[3].

表 1 各手法の実行時間
The execution times of each method.

	手法		
	文献 [3]	文献 [4]	提案手法
実行時間 [s]	2.98	3.74×10^5	2.92×10^2

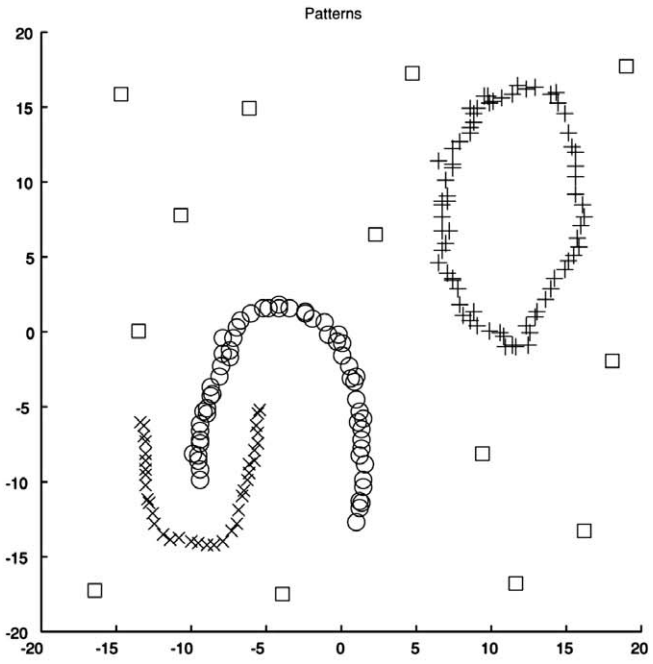

図 4　文献 [4] の手法による結果
The result by the method of reference[4].

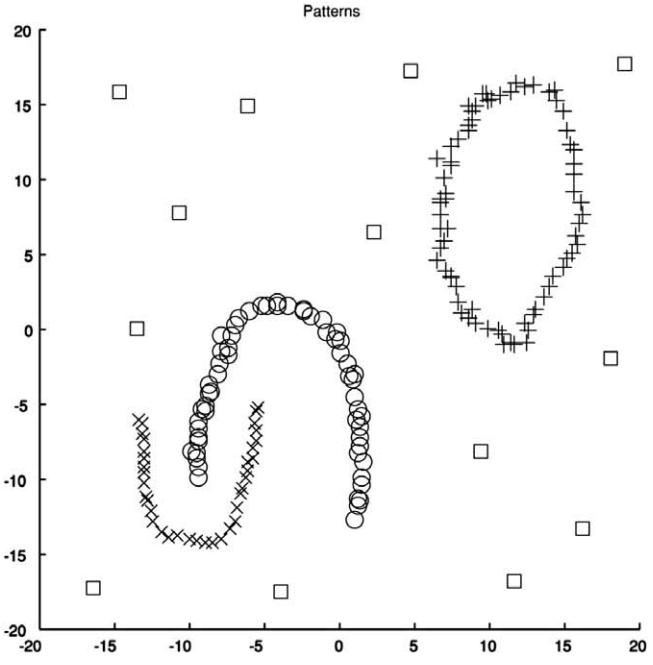

図 5　提案手法による結果
The result by the proposed method.

は，非常に良好である．しかし，文献 [4] の手法の実行時間は非常に長い．これに対して，提案手法の実行時間は比較的短いことがわかる．実験結果より，提案手法は，文献 [4] の手法に比べ，$O(n^2)$ 以上でクラスタリングに要する計算時間が短く，かつ，文献 [4] と同様に，ノイズにロバストな任意形状クラスタリングができると考える．

4. むすび

従来，莫大な処理時間を要していたノイズにロバストな任意形状のクラスタリング手法に対して，クラスタ間距離の昇順に逐次的なクラスタリングをするアプローチをとることにより，計算コストの小さいノイズにロバストな任意形状クラスタリング手法を提案した．

今後は，様々なデータに対して，提案手法の評価を行うことにより，提案手法の特性を検討する．また，提案手法を実際の画像や音声に対するクラスタリングに適用することにより，その有効性を評価する予定である．

〔文　献〕

1）L. A. Zadeh, “Fuzzy sets,” Inf. Control, 8, pp.338-353(1965)
2）J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York(1981)
3）井上光平，浦浜喜一“類似度データに基づくロバストファジークラスタリング，”信学論，Vol. J86-D-II, 8, pp.1262-1264(2003)
4）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)

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．

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，平 6 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．