研究速報

k近傍の最大距離に基づくノイズにロバストな自己組織化マップに基づくクラスタリング手法

A Clustering Method Based on a Self-Organaizing Map with Maximum Distance of k Neighbors

今村弘樹[†]，藤村誠[†]，正会員 黒田英夫[††]

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

Abstract Clustering methods, which are based on Self-Organaizing Map, can not precisely classify data when noise data is included. We describe a clustering method that can precisely classify data even when noise data are included.

キーワード：クラスタリング，自己組織化マップ，ノイズ，ロバスト性

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)〜3)]，fuzzy-c平均法[4)〜8)]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[9)10)]．

しかしこれらの手法は，本来，同じクラスに含まれるデータ間の距離は，本来，異なるクラスタに含まれるデータ間の距離よりも短いという前提でクラスタリングを行っているため，異なるクラスタ同士が接近している場合，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．

そこで筆者らは，自己組織化マップ (SOM) を用いて，SOMの各コードベクトルをコードベクトル間の距離の閾値により，分離・結合するSOMに基づくクラスタリング手法を提案した[11)12)]．この手法は，SOMのアルゴリズムに基づきコードベクトルを分離・結合させながら各クラスタにフィッティングさせ，各データに最近傍となるコードベクトルのラベルをそれぞれのデータに割り当てることにより，クラスタリングを行う．この手法により，異なるクラスタ同士が接近している場合でも良好にクラスタリングすることが可能となった．ただしこの手法は，クラスタからの外れ値であるノイズとなるデータに対するロバスト性は考慮されていないため，ノイズとなるデータが存在する場合，クラスタリング精度が著しく低下する．この課題を残したまま，この手法を，例えば，画像の領域分割に適用した場合に，本来，異なる領域が同じ領域として領域分割されたり，本来，同じ領域が異なる領域として領域分割される恐れがある．

ここでは，ノイズとなるデータに対してロバストなSOMに基づくクラスタリング手法として，k近傍の最大距離に基づくノイズにロバストな自己組織化マップに基づくクラスタリング手法を提案する．この手法は，クラスタリングする各データのk近傍の最大距離の平均と分散に基づく閾値により，ノイズとなるデータを選定し，コードベクトルをフィッティングする際に，それらのデータからの影響を除外する．これにより，ノイズに対するロバスト性が向上することが期待できる．

2. 提案手法のアルゴリズム

ここでは，提案手法のアルゴリズムを示す．図1に提案手法の処理の流れを示す．なお，提案手法のアルゴリズムは，ノイズとなるデータを除外するための閾値の決定とSOMの処理以外は，従来手法[11)]におけるアルゴリズムと同じであるので，ノイズとなるデータを除外するための閾値の決定，コードベクトルの生成とSOMの処理の個所のみ，以下に示すこととする．なお，コードベクトルの先端，末端のフラグに関する記述は省略することとする．

2008年3月26日受付，2008年7月14日再受付，2008年7月31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)

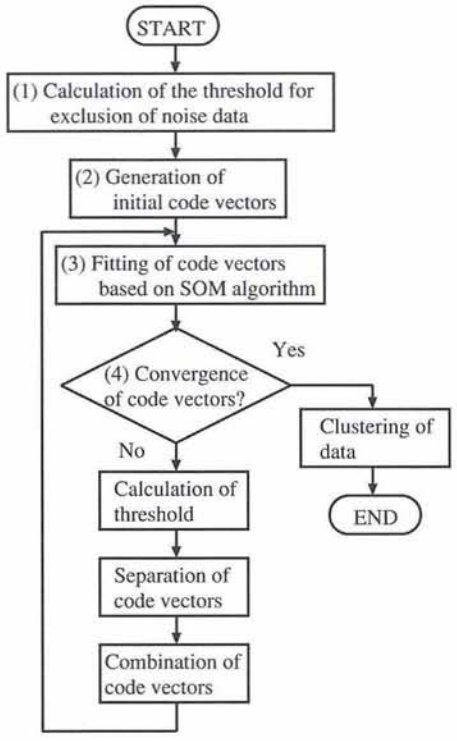

図 1　提案手法の処理の流れ図
Flow of the process in the proposed method.

2.1　ノイズとなるデータを除外するための閾値の決定

まず，ノイズとなるデータを除外するための閾値の決定のアルゴリズムを以下に示す (図 1(1))．ただし，クラスタリングするデータは d 次元のベクトル $\boldsymbol{x}$ とし，データ数は n 個とする．また，閾値 Th_{dist} を決定する際に考慮する k 近傍のデータ数を $K\_NEIGHBOR$ とする．

Step1. i=1 とする．

Step2. k=1 とする．

Step3.

$$\boldsymbol{x}_i^{(k)} = \arg \min_{1 \leqq j \leqq n, j \neq i} ||\boldsymbol{x}_i - \boldsymbol{x}_j|| \tag{1}$$

を満たすデータ $\boldsymbol{x}_i^{(k)}$ を抽出する．ただし，$\boldsymbol{x}_i^{(k)}$ は，i 番目のデータ $\boldsymbol{x}_i$ に対して k 番目に近いデータを表す．

Step4. もし，k が $K\_NEIGHBOR$ ならば Step5. へ，そうでなければ，$k = k+1$ として，Step3. へ．

Step5.

$$k\_max_i = \arg \max_{2 \leqq k \leqq K\_NEIGHBOR} ||\boldsymbol{x}_i^{(k-1)} - \boldsymbol{x}_i^{(k)}|| \tag{2}$$

を抽出する．

Step6. もし，i が n なら Step7. へ，そうでなければ，$i = i+1$ として，Step2. へ．

Step7. $k\_max_i (1 \leqq i \leqq n)$ の平均 E_k と分散 V_k を算出する．

Step8. $Th_{dist} = E_k + \lambda V_k$ とする．ただし，λ は重み係数とする．

Step9. アルゴリズム終了．

ここでは，k_max_i が Th_{dist} より大きい値となったデータをノイズのデータとして扱う．また，Step8. において単位の異なる E_k と V_k を足しているが，係数の λ が単位を調整する係数も兼ねるため，問題はないと考える．

2.2　コードベクトルの生成と SOM の処理

次に，コードベクトルの生成と SOM の処理のアルゴリズムを以下に示す．

Step1. d 次元のコードベクトル $\boldsymbol{m}_p^l$ を定義域内においてランダムに生成する (図 1(2))．ただし，$\boldsymbol{m}_p^l$ は，コードベクトルのラベル $l (1 \leqq l \leqq C_l)$ における $p (1 \leqq p \leqq num_l)$ 番目におけるコードベクトルを表す．初期値として，$C_l = 1$，全てのコードベクトルのラベルを 1 とする．

Step2. 以下を $t = 1, 2, \cdots; j = 1, 2, \cdots n$ について繰り返す (図 1(3))．

Step3. 次の式

$$(l^*, p^*) = \arg \min_{1 \leqq l \leqq C_l, 1 \leqq p \leqq num_l} ||\boldsymbol{m}_p^l - \boldsymbol{x}_j|| \tag{3}$$

を満たす $\boldsymbol{m}_{p^*}^{l^*}$ を j 番目のデータ $\boldsymbol{x}_j$ に対する勝利コードベクトルとする．

Step4. 勝利コードベクトルとその周辺のコードベクトルを

$$\boldsymbol{m}_p^l = \begin{cases} \boldsymbol{m}_p^l + \alpha(t)\{\boldsymbol{x}_j - \boldsymbol{m}_p^l\}\beta ||\boldsymbol{x}_j - \boldsymbol{m}_{p^*}^{l^*}||^2 \\ \quad \text{if } (p \in p^* + N_c) \wedge (l = l^*) \\ \quad \wedge (k\_max_j < Th_{dist}) \\ \boldsymbol{m}_p^l \\ \quad \text{if } (p \notin p^* + N_c) \vee (l \neq l^*) \\ \quad \vee (k\_max_j \geqq Th_{dist}) \end{cases} \tag{4}$$

により更新する．ただし，N_c は，p^* に対する N_c 近傍を表し，k_max_j は，$\boldsymbol{x}_j$ の k 近傍におけるデータ中の最大距離を表す．また，βは重み係数を表す．

$$\alpha(t) = \frac{0.7}{1 + [t/7]} \tag{5}$$

とし，$[q]$ は，q を超えない最大の整数を表す．

Step5. 以下を $p = 1, 2, \cdots n, l = 1, 2, \cdots C_l$ について繰り返す．

Step6. 次の式

$$j^* = \arg \min_{1 \leqq j \leqq n} ||\boldsymbol{m}_p^l - \boldsymbol{x}_j|| \tag{6}$$

を満たす $\boldsymbol{x}_{j^*}$ を $\boldsymbol{m}_p^l$ に対する勝利データとする．

Step7. $\boldsymbol{m}_p^l$ とその周辺のコードベクトルを

$$\boldsymbol{m}_{pn}^l = \begin{cases} \boldsymbol{m}_{pn}^l + \alpha(t)\{\boldsymbol{x}_{j^*} - \boldsymbol{m}_{pn}^l\}\gamma ||\boldsymbol{x}_{j^*} - \boldsymbol{m}_p^l||^2 \\ \quad \text{if } (pn \in p + N_c) \\ \quad \wedge (k\_max_j < Th_{dist}) \\ \boldsymbol{m}_p^l \\ \quad \text{if } (pn \notin p + N_c) \\ \quad \vee (k\_max_j \geqq Th_{dist}) \end{cases} \tag{7}$$

により更新する．ただし，γ は重み係数を表す．
Step8.

$$dist_t = \arg \max_{\substack{1 \le l \le C_l \\ 1 \le p \le num_l \\ 1 \le j \le n}} ||\boldsymbol{m}_p^l - \boldsymbol{x}_j|| \tag{8}$$

を算出し，

$$dist_t < Th_{dist_t} \tag{9}$$

を満たす場合，アルゴリズム終了 (図 1(4))．

式 (4) と式 (7) において，$k\_max_j \geq Th_{dist}$ となるデータをノイズのデータとして処理している．これにより，コードベクトルを更新する際にノイズのデータの影響を除外している．

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．ここでは，文献 12) の手法を従来手法とし，提案手法との比較を行った．また，ここで用いる人工データは，文献 12) で用いられているデータと同等のものにノイズを付加したものとした．なお，従来手法と提案手法のパラメータは予備実験により決定した．従来手法のパラメータは，コードベクトル数は 40，Nc=3，β=0.2，γ=0.2,Th_{dist_t}=5.0×10^{-3} とし，提案手法のパラメータは，コードベクトル数は 40，Nc=3，β=0.4，γ=0.4，$K\_NEIGHBOR$=2，λ=2.5×10^{-1}，Th_{dist_t}=5.0×10^{-3} とした．

まず，図 2 に示すデータに対して，実験を行った．このデータは，中央の二つのクラスタが近接し，その周辺にノイズとなるデータが分布している．図 3 は，従来手法における繰り返し計算終了後におけるコードベクトルの状態を示している．従来手法では，コードベクトルが，ノイズとなるデータの影響を受け，クラスタリングすべきデータ以外のデータにもフィッティングしていることがわかる．図 4 は，従来手法におけるクラスタリング結果を示している．なお，クラスタリング結果では，同じクラスタには同じ記号を，異なるクラスタには異なる記号で表している．従来手法では，全体が一つのクラスタとして抽出された．これは，ノイズとなるデータにコードベクトルがフィッティングしてしまったためであると考える．また，図 5 は，提案手法における繰り返し計算終了後におけるコードベクトルの状態を示している．図 6 は，提案手法におけるクラスタリング結果を示している．なお，× 印で示されているデータは，ノイズとして処理したデータを表している．提案手法は，コードベクトルがクラスタの形状にフィットできていることが分かる．

次に，図 7 に示すデータに対して，実験を行った．このデータは，中央のクラスタの端の密度が高く，その周辺にノイズとなるデータが分布している．図 8 は，従来手法における繰り返し計算終了後におけるコードベクトルの状態を示している．従来手法では，コードベクトルが，ノイズとなるデータの影響を受け，クラスタリングすべきデータ以外のデータにもフィッティングしていることがわかる．図 9 は，従来手法におけるクラスタリング結果を示している．従来手法では，全体が一つのクラスタとして抽出された．これは，ノイズとなるデータにコードベクトルがフィッティングしてしまったためであると考える．また図 10 は，提案手法における繰り返し計算終了後におけるコードベクトルの状態を示している．図 11 は，提案手法におけるクラスタリング結果を示している．提案手法は，コードベクトルがクラスタの形状にフィットできていることがわかる．クラスタリング結果から提案手法は，従来手法に比べ，良好にクラスタリングが行われていることがわかる．

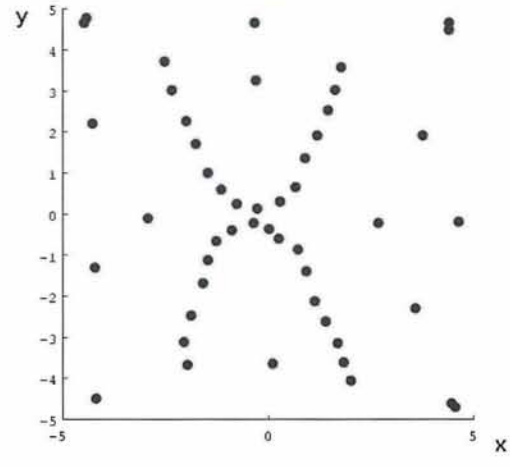

図 2　生成した人工データ#1
Generated synthesis data #1.

次に，従来手法と提案手法を用いて図 12 に示す実画像の領域分割を行った．実験に用いた実画像は，水色の背景に赤の球体（左上，右下）と青の球体（左下，右上）が存在する 32×40[pixel] のカラー画像とした．この画像の各画素の RGB 値の内，R と B の値で 2 次元のデータとし，全画素のデータをクラスタリングし，クラスタ毎ラベル値を与え，そのラベル毎に領域分割した．まず，従来手法による領域分割の結果を図 13 に示す．従来手法では，赤と青の球や球体の影の箇所が一つの領域として抽出された．これは，球体の輪郭と影の箇所のデータの影響を受けて，赤と青の球体の箇所のデータが一つのクラスタとして抽出されたためと考える．次に，提案手法の領域分割の結果を図 14 に示す．提案手法では，提案手法に比べ，背景，青の球体，赤の球体，球体の影の箇所が良好に領域分割できていることがわかる．これは，提案手法では，球体の輪郭と影の箇所のデータの影響を除外することができたためと考える．

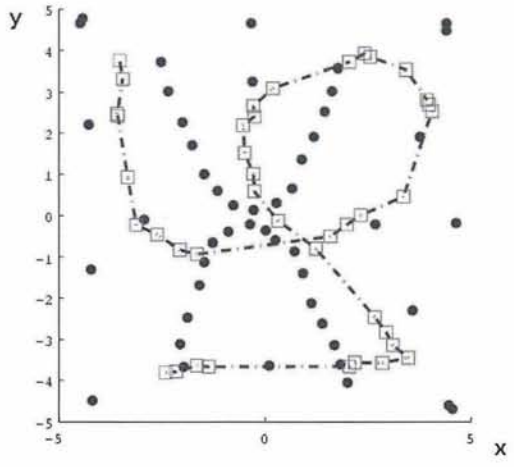

図 3　繰り返し計算終了後におけるコードベクトルの状態 (従来手法)
The state of cord vectors after final repetition calculation (the conventional method).

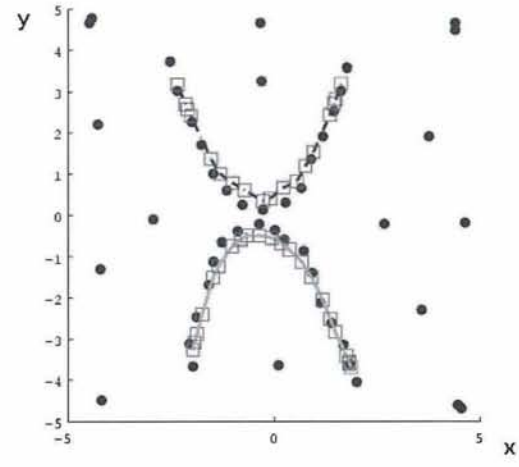

図 5　繰り返し計算終了後におけるコードベクトルの状態 (提案手法)
The state of cord vectors after final repetition calculation (the proposed method).

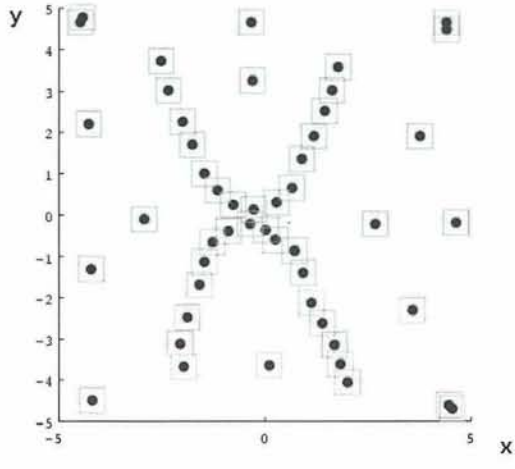

図 4　クラスタリング結果 (従来手法)
The result of clustering (the conventional method).

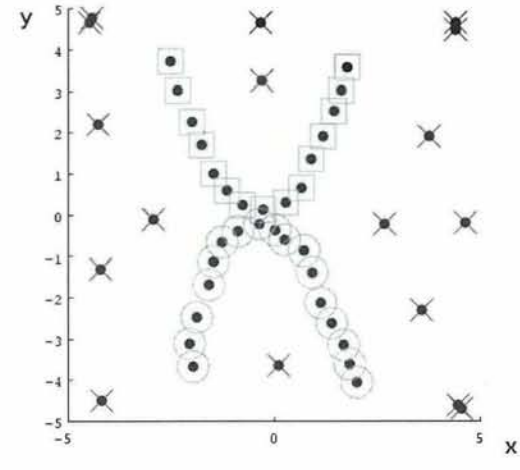

図 6　クラスタリング結果 (提案手法)
The result of clustering (the proposed method).

4. むすび

従来の分離・結合する SOM に基づくクラスタリング手法は，クラスタリングするデータにノイズとなるデータが含まれる場合，著しくクラスタリング精度が低下するという課題があった．この課題を解決するために，k 近傍の最大距離に基づく自己組織化マップに基づくクラスタリング手法を提案した．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ノイズとなるデータを含む人工データに対して，従来手法と提案手法を用いてクラスタリングを行った．実験の結果，提案手法は，従来手法に比べ，ノイズとなるデータの影響を除外して良好にクラスタリングできたと考える．また，従来手法と提案手法を用いて実画像の領域分割を行い，提案手法は，従来手法に比べ，良好に各領域を領域分割できたと考える．

〔文　献〕

1) K. Fukunaga, "Introduction to statistical pattern recognition," Academic press, Boston, 2 edition (1990)
2) R. O. Duda, P. E. Hart and D. G. Stocrk, "Pattern Clasification – Second Edition," Wiley Interscience (2002)
3) 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 (D-II), J82-D-II, 7, pp.1120-1128 (1999)
4) 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)
5) 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)
6) 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)
7) 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 (D-II), J83-D-II, 9, pp.1957-1961 (2000)
8) 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 (D-II), Vol.J85-D-II, 6, pp.1140-1143 (2002)
9) 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 (D-II), J86-D-II, 10, pp.1511-1513 (2003)

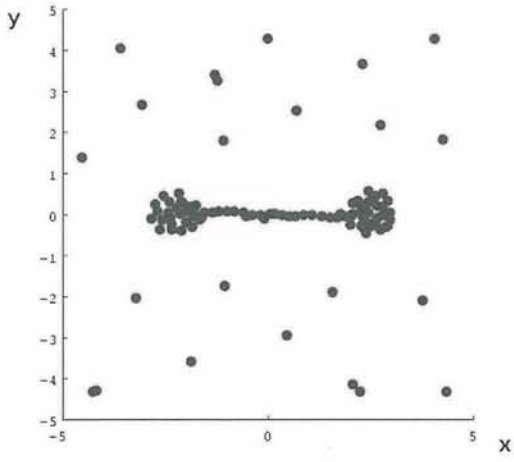

図 7　生成した人工データ#2
Generated synthesis data #2.

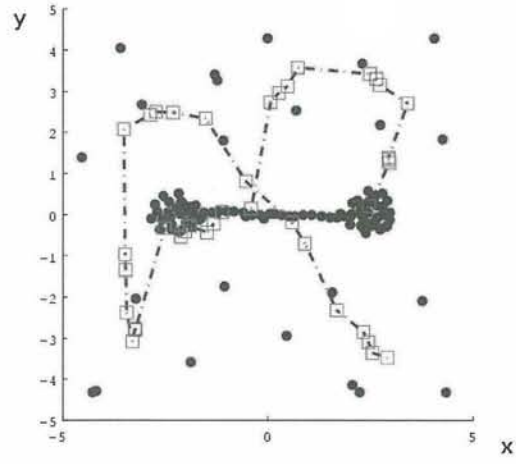

図 8　繰り返し計算終了後におけるコードベクトルの状態 (従来手法)
The state of cord vectors after final repetition calculation (the conventional method).

10) 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," 映情学会誌, 60, 4, pp.618-620 (2006)
11) 今村 弘樹, 藤村 誠, 黒田 英夫, "コードベクトルの分離・結合を考慮した自己組織化マップに基づくクラスタリング手法", 信学会 NC 研究会技報, 106, 501, pp.29-34 (2007)
12) 今村 弘樹, 藤村 誠, 黒田 英夫, "データとコードベクトルの距離による重みを考慮した自己組織化マップに基づくクラスタリング手法", 信学論 (A), 90-A, 11, pp.885-890 (2007)

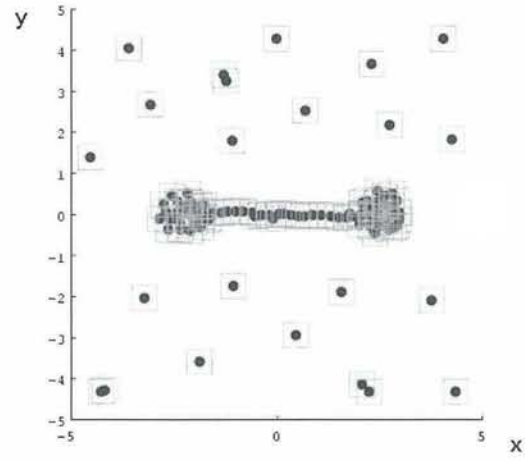

図 9　クラスタリング結果 (従来手法)
The result of clustering (the conventional method).

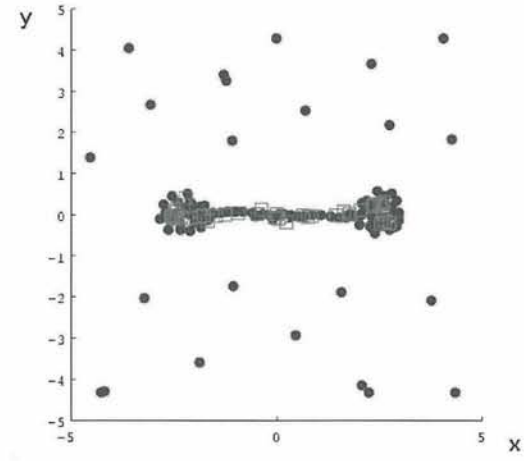

図 10　繰り返し計算終了後におけるコードベクトルの状態 (提案手法)
The state of cord vectors after final repetition calculation (the proposed method).

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, 2007 年, 長崎大学工学部情報システム工学科助教, 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, 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.

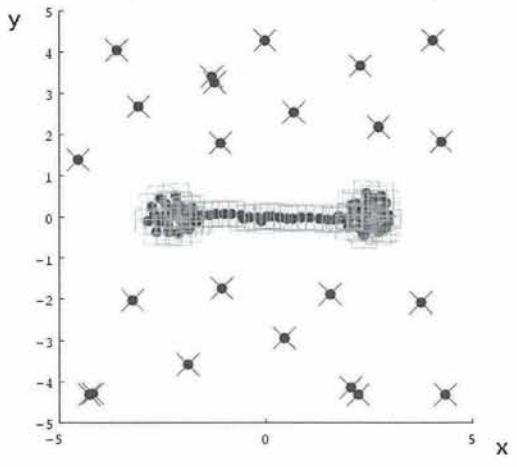

図 11　クラスタリング結果 (提案手法)
The result of clustering (the proposed method).

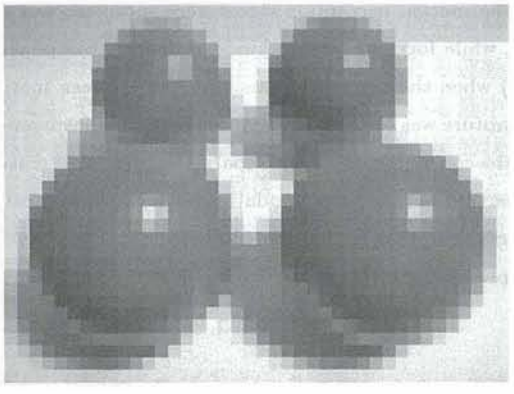

図 12　実験に用いた実画像
The actual image used in the experiment.

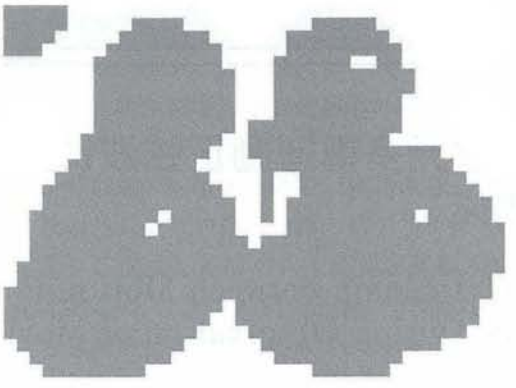

図 13　従来手法による領域分割の結果
The segmentation result by the conventional method.

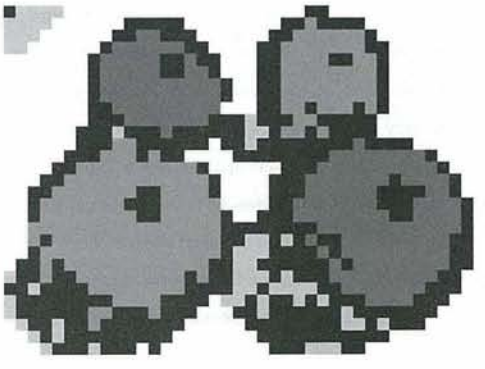

図 14　提案手法による領域分割の結果
The segmentation result by the proposed method.