Vol.2012-MPS-91 No.7
Vol.2012-BIO-32 No.7
2012/12/6

画像データの学習クラスタリング

高野 太吾[†1]　佐藤 仁樹[†1]

概要：ユーザが意図するクラスターを得るためのクラスタリングパラメータ調節法を提案した．まず，ユーザが理想とするクラスターとクラスタリング結果との差を表す指標として，クラスタリング精度及びクラスター数精度を導入した．次に，クラスターの情報があらかじめ分かっている学習用データに対し，クラスタリング精度とクラスター数精度が高くなるようにクラスタリングパラメータを調節した．得られたパラメータを用いて評価用データをクラスタリングし，正しいクラスター数と高いクラスタリング精度を得た．

キーワード：学習クラスタリング，データ変換，SOM

Learning Clustering for Image Data

Abstract: We developed a method that adjusts clustering parameters to obtain ideal clusters. First, we introduced two measures, the matching accuracy and structure accuracy, that evaluate the difference between ideal clusters and the clusters obtained using the method presented in this report. Next, we developed a learning algorithm that adjusts the clustering parameters so as to increase the accuracies for the learning data. Finally, we derived the clusters for the evaluation data using the parameters that were obtained using the learning clustering. High accuracies were obtained using the learning algorithm.

Keywords: learning clustering, data conversion, SOM

1. はじめに

データのクラスタリング手法は，非階層的クラスタリングと階層的クラスタリングに分類される [1],[2]．代表的な非階層的クラスタリングである k-means 法 [1] は，階層的クラスタリングと比較して計算量が少なく，簡単な計算によってクラスターを生成できる．しかし，k-means 法は事前にクラスター数を指定する必要があるため，クラスタ数が未知のデータに対して適用できない．そこで，データ群の局所的なまとまりの良さをベイズ型情報量規準 (BIC) によって判断し，k-means 法によってデータを再帰的に分割することにより，クラスター数を決定する x-means 法が提案された [3]．x-means 法は，データの統計量に基づきクラスターを決定するため，統計的に適切なクラスターが得られる．しかし，k-means 法では初期値によって最終的なクラスタリング結果が大きく変わってしまうため，同じデータに対して何度もクラスタリングを実行し，得られた結果からユーザが最適なクラスターを選ぶ必要がある．

一方，一般的な階層的クラスタリングは，データ間の非類似度を手がかりとして樹形図を構成し，樹形図を切る高さによって，様々なクラスター構成が得られる．クラスター数を自動的に決定する必要がある場合には，階層的クラスタリングに統計的尺度を適用し，適切なクラスター数を決める方法が提案されている [4],[5]．また，SOM[6] と統計的尺度を用いてクラスター抽出するクラスター数の推定法では，SOM を用いてデータを二次元化し，統計的尺度を用いてクラスターとクラスター数を決定する [7]．そのため，結果を視覚化しやすい．また，使用する距離や各クラスターの大きさの違いに依存せずにクラスター数を推定できる．しかし，統計的尺度を用いる手法は，データの統計量に基づいてクラスターを決定するため，データ数が少ない場合には適切なクラスターが得られない可能性がある．また，統計的尺度によりデータのまとまりを判断するため，統計的尺度の観点からは良い結果であっても，得られた結果がユーザの期待する結果と一致するとは限らない．

これらの問題を解決するために，ユーザが理想とするク

[†1] 現在，公立はこだて未来大学　システム情報科学研究科
Presently with Future University Hakodate, School of Systems Information Science

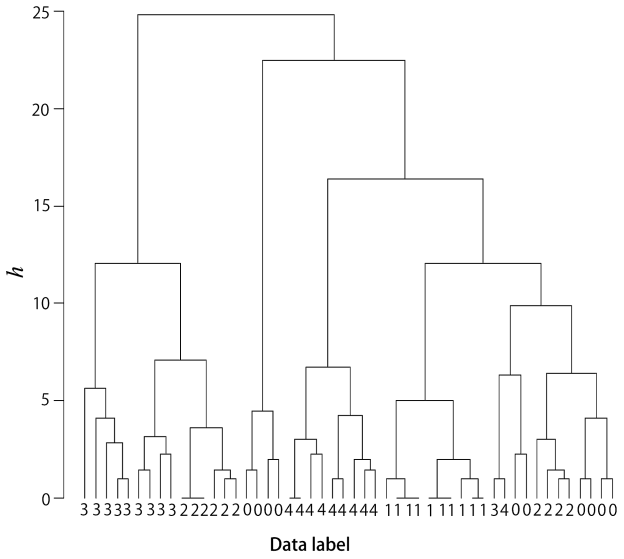

図 1 樹形図の例.
Fig. 1 Example of tree.

ラスターとクラスタリング結果との差を表す指標としてクラスタリング精度及びクラスター数精度を導入する．学習によりクラスタリング精度とクラスター数精度が高くなるようにクラスター数とデータ変換に関するパラメータを調節し，ユーザが意図するクラスターを得る手法を提案する．

2. 階層的クラスタリング

クラスタリングとは，分類対象の集合を内的結合と外的分離が達成されるような部分集合に分割する手法である．統計解析や多変量解析の分野ではクラスター分析とも呼ばれ，基本的なデータ解析手法として幅広く使用されている．特に，人間が分類できないほどの膨大な量のデータを分類する場合等において，クラスタリングが有効である．クラスター分析には多種多様な技法があり，大きく階層的クラスタリングと非階層的クラスタリングの 2 つに分かれる [1],[2]．本節では階層的クラスタリングについて述べる．

階層的クラスタリングは，対象間の非類似度を手がかりとして，樹形図あるいはデンドログラムと呼ばれる樹状の分類構造を目的とする．階層的クラスタリングでは，N 個のデータを入力とした場合，1〜N 個のクラスターを得る．樹形図の例を**図 1** に示す．図 1 において，縦軸 h は樹形図の高さを表す．各枝の先にはクラスタリングされるデータのラベル (図 1 の樹形図における 0,1, ... ,4) がある．樹形図を適当な高さ h_c で切断することにより，1〜N 個の任意個数のクラスターを得る．階層的クラスタリングには，Ward 法，最近隣法，最遠隣法，群平均法，McQuitty 法? などの種類がある [1],[8].

- Ward 法 (Ward's method)

クラスター内のばらつきが増えないように，クラスターを結合する方法.

- 単連結法 (Single linkage method)

２つのクラスターに属する最も近いデータ間の距離を測り，最も距離の短いクラスターを結合する方法．データの散らばり方が１つの方向に長い鎖状になっている場合に適する．

- 完全連結法 (Complete linkage method)

２つのクラスター間の距離を，それぞれのクラスターから１つずつ選んだデータ間の距離の中の最大値で定義し，クラスターを結合する．データがいくつかの集団に固まっているときに良い結果が期待できる．

- 群平均法 (Group average method)

２つのクラスターから１つずつ要素を選んで距離を求める．すべての要素の組み合わせで距離を求め，その平均を２つのクラスターの距離と定義し，距離の近いクラスターを結合する．データがいくつかの集団に固まっていたり，１つの方向に鎖状にのびている場合でも，良い結果が期待できる．

- McQuitty 法 (McQuitty method)

新しいクラスターからの距離の平均をクラスター間の距離とし，クラスターを結合する方法?.

データ間の距離 D としてユークリッド距離とマンハッタン距離を使用する．２つのデータをベクトル $\boldsymbol{a}$,$\boldsymbol{b}$ で表すと，ユークリッド距離は，$D(\boldsymbol{a},\boldsymbol{b}) = (\sum_{i=1}^{n}(a_i - b_i)^2)^{1/2}$ ，マンハッタン距離は，$D(\boldsymbol{a},\boldsymbol{b}) = \sum_{i=1}^{n}|a_i - b_i|^2$ で定義される．

階層的クラスタリングでは，得られた樹形図を切る高さ h_c によって得られるクラスター数が異なる．また，階層的クラスタリング手法と距離の組み合わせによって樹形図の高さや，形成されるクラスターが異なる．すなわち，正しいクラスター数を得るには，クラスター数パラメータを h として，h を適切に決める必要がある．統計的な尺度を用いて適切なクラスター数を決める方法では，得られたクラスターが必ずしもユーザが所望するクラスターと一致する保証はない．また，正しいクラスター数のクラスターが得られたとしても，必ずしも十分なクラスタリング精度が得られるとは限らない．そこで，クラスタ数精度とクラスタリング精度を改善するために，データを適切に変換する．そのパラメータをデータ変換パラメータとする．学習データに対してクラスタリングパラメータ (データ変換パラメータ及びクラスター数パラメータ) を調整する方法を提案する．

3. 学習クラスタリング

3.1 アルゴリズム

本節では，階層クラスタリング手法にクラスター数パラメータとデータ変換パラメータを導入し，これらのパラ

メータを調整する方法を述べる．

クラスタリングでは，膨大な量のデータを扱うため，すべてのデータをユーザ自身がクラスタリングできない．そこで，データの一部を抽出し，ユーザがクラスタリングすることにより，コンピュータがクラスタリングする際の手本となる学習データを作成する．次に，以下に示す学習アルゴリズムにより高いクラスタリング精度とクラスター数精度が得られるように，データ変換パラメータ及びクラスター数パラメータを調節する．学習用データの正しいクラスター数を K^*，クラスタリングによって得られたクラスター数を $\hat{K}$，クラスタリング精度を P_{C}(付録 1 参照)，クラスター数精度を P_{N} とする．P_{N} を次式で定義する．

$$P_{\mathrm{N}} = (1 - \min(\frac{|\hat{K} - K^*|}{K^*}, 1)) \times 100[\%]. \quad (1)$$

以下に本手法における学習アルゴリズムを示す．

Algorithm1. 学習

１．データ変換パラメータの調整
２．学習用データの変換
３．学習用データのクラスタリング
４．P_{N} の計算及びクラスター数パラメータ h_c の調整
５．P_{C} の計算
６．最適なパラメータの探索が完了したならば終了．
　　そうでなければステップ１へ戻る．

Algorithm1 により得られた最適なデータ変換パラメータ σ_c，s_c (3.3 参照)，及びクラスター数パラメータ h_c を用いて評価用データをクラスタリングし，Algorithm2 により評価用データのクラスタリング精度 P_{C} 及びクラスター数精度 P_{N} を計算する．

Algorithm2. 評価

１．評価用データの変換
２．評価用データのクラスタリング
３．P_{N} 及び P_{C} の計算

3.2 クラスター数パラメータの調整

図 1 から分かるように，階層的クラスタリングでは，P_{N}=100 %となる分類木の高さには幅がある．そこで，Algorithm1 のステップ 4 では，P_{N}=100 %となる分類木の高さの最大値 h_{max}，最小値 h_{min} を求め，P_{N}=100 %となる分類木の高さ，すなわち最適なクラスター数パラメータ h_c を次式により決定する．

$$h_c=(h_{max}+h_{min})/2.$$

3.3 データ変換パラメータの調整

本節では，データを変換することにより，クラスタリング精度を改善する方法を述べる．本報告では，画像データを扱う．画像の変換はガウスフィルタ [9] 及び SOM[6](付録 1) により行われる．ガウスフィルタのパラメータ σ を変えることにより，ぼかし効果を与え，線分の位置変動を吸収することにより，クラスタリング精度を改善できる [10]．また，SOM によりデータを 2 次元マップ上に配置し，データを扱いやすくする．データが配置される座標は，学習ステップ数 s によって異なる．そこで，SOM の学習ステップ数 s 及び σ をデータ変換パラメータとする．クラスタリング精度 P_{C} 及びクラスター数精度 P_{N} が最も高くなるデータ変換パラメータ σ と s を各々 σ_c，s_c とする．

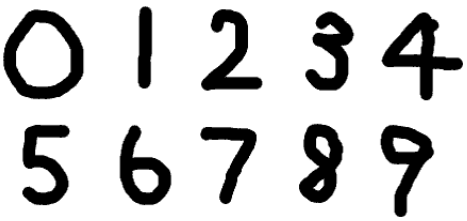

図 2　手書き文字の例.
Fig. 2　Example of characters.

4. 性能評価

本節では，3 節で提案された学習クラスタリング手法を評価する．今回の実験では，手書き文字 (0〜9) をクラスタリングする．今回は，0〜3 を学習データ，4〜9 を評価データとした．すなわち，4 クラスのデータを学習データとして，学習データとは異なる 6 種類の文字からなる評価データに対して提案手法の性能を評価する．ここで，手書き文字の画像はモノクロ，12 × 12[pixel] であり，学習データ数及び評価データ数は各々計 80 枚及び計 60 枚である．使用した手書き文字の例を**図 2** に示す．

まず，学習データのクラスター数精度 P_{N} 及びクラスタリング精度 P_{C} に対するデータ変換パラメータ σ 及び s の影響を評価した．σ 及び s は，クラスター数精度 P_{N} に影響しない．これは，学習データでは樹形図の構造に応じて，h_c を決めることにより適切なクラスター数を得られるためである．一方，クラスタリング精度 P_{C} は樹形図の構造に直接関係するため，データ変換パラメータ σ 及び s に大きく影響される．そこで，σ 及び s のクラスタリング精度に対する影響を評価し，その結果を**図 3** に示す．図 3 より，SOM 使用，未使用に関わらずデータ変換パラメータ σ に従い，クラスタリング精度 P_{C} が大きく変化することが分かる．

また，SOM で用いるデータ変換パラメータ s に対するクラスタリング精度 P_{C} を**図 4** に示す．データ変換パラメータ s の増加に伴い，クラスタリング精度 P_{C} も増加する．しかし，s がある値を超えるとクラスタリング精度 P_{C}

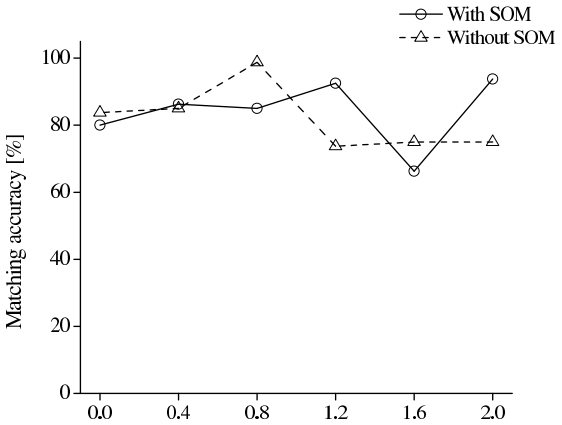

図 3 σ とクラスタリング精度の関係 (学習データ:Ward 法，Euclid 距離).

Fig. 3 Effect of σ on matching accuracies(learning data:Ward method, Euclid distance).

表 1 学習データから得られた σ_c(SOM 使用).
Table 1 Optimum values of σ_c (with SOM).

	Ward	Average	McQuitty	Complete
Euclid	1.5	0.9	1.2	1.3
Manhattan	1.5	1.1	1.7	1.2

は悪化する．すなわち s の値が小さい場合，学習不足となり，s の値が大きい場合，過学習となる．これらの結果より，データ変換パラメータの σ 及び s を最適化する必要があることがわかる．

Algorithm1 により学習データをクラスタリングした結果を**図 5**, **図 6** に示す．図 5 では，SOM 及びガウスフィルタによってデータが変換されている．図 6 では，ガウスフィルタのみによってデータが変換されている．ここで，E, M は各々 Euclid 距離と Manhattan 距離を表す．図 5 では，クラスタリング精度が 90 %以上，図 6 ではクラスタリング精度は 70 %以上である．この結果より，SOM によるデータ変換の有効性が明らかになった．クラスター数精度 P_{N} は常に 100 %となるため，図は省略する．

Algorithm1 により得られた σ_c と s_c の値を表 1, 2, 3 に示す．ここで，データ変換パラメータ σ と s の探索範囲は各々 $0 \leqq \sigma \leqq 2$，$10^2 \leqq s \leqq 10^5$ である．ここで σ=0 の場合，ガウスフィルタを用いないこととする．表 1 は SOM 及びガウスフィルタを用いたデータ変換を学習データに用いた際に得られた最適な σ の値を表す．それぞれの手法に対し，最適な σ が一つずつ得られた．表 2 はガウスフィルタのみを用いて (SOM 未使用) データ変換を行った際の最適な σ の値を表す．最適な σ が一つにならず，幅広い σ が最適と判断された．表 3 は SOM 及びガウスフィルタを用いてデータ変換を行った際に得られた最適な s の値を表す．

評価用データを Algorithm2 を用いてクラスタリングし，クラスター数精度及びクラスタリング精度を評価した．ここで，クラスター数精度 $P_{\mathrm{N}} \neq 100$ %の場合，クラスタ

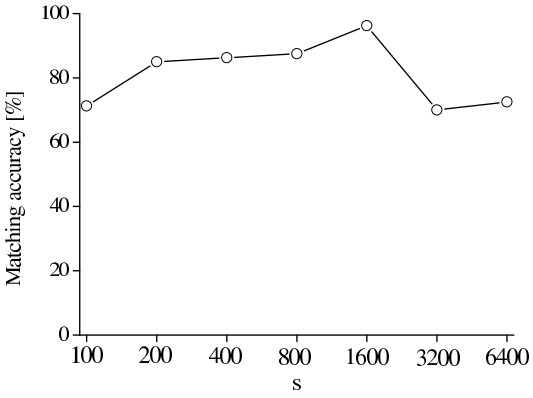

図 4 s とクラスタリング精度の関係 (学習データ:SOM 使用，Ward 法，Euclid 距離，σ=1.5).

Fig. 4 Effect of s on matching accuracies(learning data:with SOM, Ward method, Euclid distance).

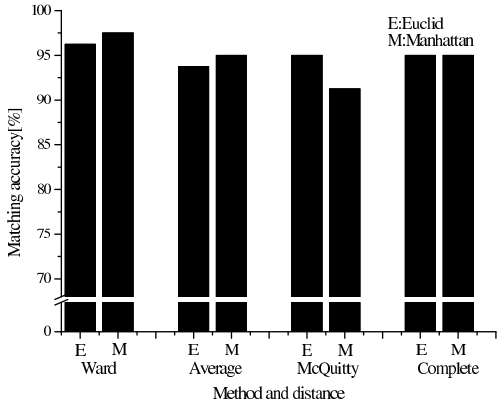

図 5 方式毎のクラスタリング精度 (学習データ:SOM 使用).

Fig. 5 Matching accuracies of each method (learning data:with SOM).

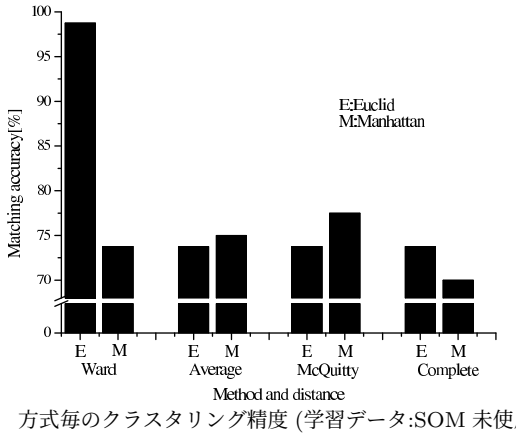

図 6 方式毎のクラスタリング精度 (学習データ:SOM 未使用).

Fig. 6 Matching accuracies of each method (learning data:without SOM).

表 2 学習データから得られた σ_c(SOM 未使用).
Table 2 Optimum values of σ_c (without SOM).

	Ward	Average	McQuitty	Complete
Euclid	0.8	0-0.7	0-0.9	0-2
Manhattan	0-0.4	1.7-1.9	0-0.7	0.6

表 3 学習データから得られた s_c(SOM 使用).
Table 3 Optimum value of s_c (with SOM).

	Ward	Average	McQuitty	Complete
Euclid	1600	6400	3200	3200
Manhattan	1600	1600	200	3200

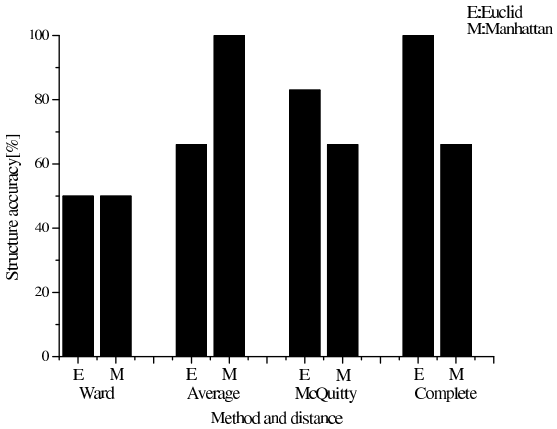

図 7 方式毎のクラスター数精度 (評価データ:SOM 使用).
Fig. 7 Structure accuracies of each method (evaluation data:with SOM).

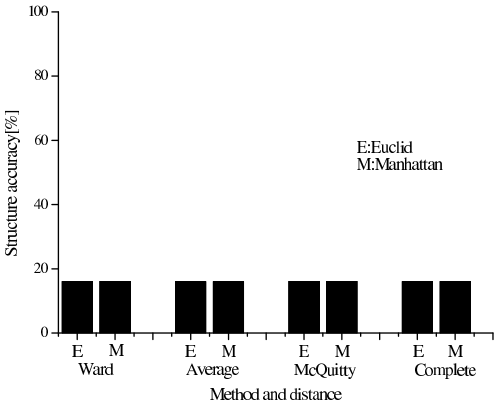

図 8 方式毎のクラスター数精度 (評価データ:SOM 未使用).
Fig. 8 Structure accuracies of each method(evaluation data:without SOM).

リング精度 P_C は 0 %とした．**図 7** に，SOM 及びガウスフィルタを用いてデータ変換を行った際のクラスター数精度，**図 8** にガウスフィルタのみ (SOM 未使用) を用いてデータ変換を行った際のクラスター数精度を示す．SOM を用いない場合は高いクラスター数が得られない．**図 9** には，SOM 及びガウスフィルタを用いてデータ変換を行った際のクラスタリング精度，**図 10** にはガウスフィルタのみ (SOM 未使用) を用いてデータ変換を行った際のクラスタリング精度を示す．図 7, 図 9 の Average 法と Complete 法では，SOM を用いた場合，評価データに対して高い精度が得られた．一方，SOM を用いない場合，意図したクラスター数とクラスタリング精度は得られなかった．

これらのシミュレーション結果から，本報告で提案した手法により，学習データとは異なる文字データからなる評

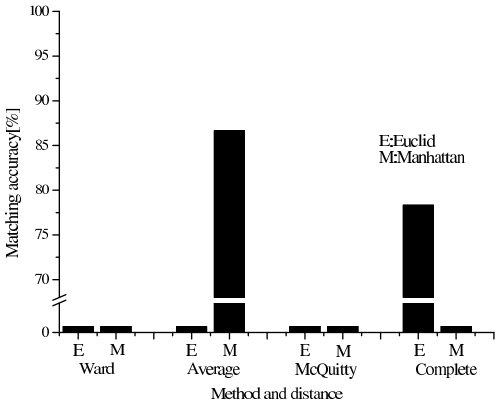

図 9 方式毎のクラスタリング精度 (評価データ:SOM 使用).
Fig. 9 Matching accuracies of each method (evaluation data:with SOM).

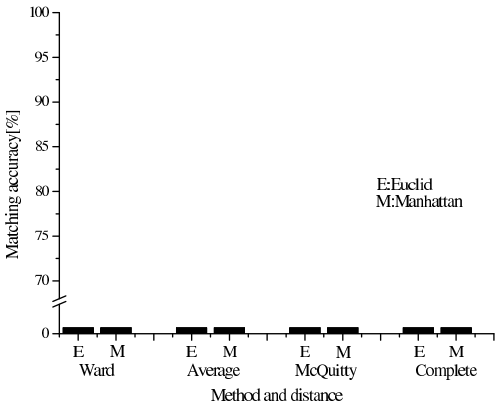

図 10 方式毎のクラスタリング精度 (評価データ:SOM 未使用).
Fig. 10 Matching accuracies of each method (evaluation data:without SOM).

価データを適切にクラスタリングできることが分かった．すなわち，本報告で提案したクラスタリングパラメータ調節法はユーザが意図するクラスタリングに有効である．特に，SOM 及びガウスフィルタをデータ変換に用いた場合，高いクラスター数精度及びクラスタリング精度が得られた．

5. 結論

ユーザが意図するクラスターを得るための学習クラスタリングを提案した．提案したアルゴリズムでは，ユーザが理想とするクラスターとクラスタリング結果との差を表す指標として，クラスタリング精度とクラスター数精度を導入した．これらの値が高くなるようにクラスタリングパラメータを学習することにより，未知の評価データに対して高い精度が得られた．

参考文献

[1] 神嶌敏弘：データマイニング分野のクラスタリング手法 (1) クラスタリングを使ってみよう！, 人工知能学会誌, Vol. 18, No. 1, pp. 59–65 (2003).
[2] 神嶌敏弘：データマイニング分野のクラスタリング手法 (2) 大規模データベースへの挑戦と次元の呪いの克服, 人工知能学会誌, vol. 18, No. 2, pp. 170–176 (2003).

[3] 石岡 恒憲：x-means 法改良の一提案 k-means 法の逐次繰り返しとクラスターの再併合, 計算機統計学, Vol. 18, No. 1, pp. 3–13 (2005).
[4] 新海公昭：階層クラスター分析における最適なクラスター数の決定問題, バイオメディカル・ファジィ・システム学会大会講演論文集 BMFSA, No. 21, pp. 189–190 (2008).
[5] 遠藤靖典：クラスタ数推定機能を持つ階層的ファジィクラスタリング, 電子情報通信学会論文誌 A, Vol. J79-A, No. 7, pp. 1276–1288, (1996).
[6] Kohonen,T.: 自己組織化マップ, シュプリンガーフェアラーク東京 (2005).
[7] 加藤聡：自己組織化マップと情報量規準によるクラスタ数の推定法に関する基礎的研究, 情報科学技術フォーラム講演論文集, Vol. 9, No, 2, pp. 537–538 (2010).
[8] 柳井久江：エクセル統計 実用多変量解析編, オーエムエス出版 (2005).
[9] 中川正雄：確率過程, 培風館 (2002).
[10] 福井隆文：背景伝搬法による手書き漢字認識, 電子情報通信学会, パターン認識・メディア理解, Vol. 107, No, 491, pp. 111–116 (2008).

付　　録

A.1 精度の計算

データ数を N，クラスター数を K，i をクラスター番号 (i=1，2，…．K)，得られたクラスターを C_i とする．各クラスターはそれぞれ $n_i(N{=}\sum_{i=1}^{K} n_i)$ 個のデータをもっており，それら n_i 個のデータはそれぞれ正解ラベル $L_j(j{=}1, 2, \cdots, K)$ を持つ．C_i に属するデータが持つ L_j の個数を $N_{\mathrm{L}ij}$, 各 C_i の中で最大の $N_{\mathrm{L}ij}$ を $N_{\mathrm{LMAX}i}$ として，$N_{\mathrm{LMAX}i}$ を以下の式で求める．

$$N_{\mathrm{LMAX}i} = \max(N_{\mathrm{L}i1}, N_{\mathrm{L}i2}, ..., N_{\mathrm{L}iK}) \tag{A.1}$$

ここで，各 C_i の中で最大となった $N_{\mathrm{L}ij}$ のラベルを $\hat{L}_i$ とする．また，$N_{\mathrm{L}ij}$ の個数が同数であった場合，j の大きい方を優先する．次に，各 C_i の精度を P_i として以下の式で求める．

$$P_i = \frac{N_{\mathrm{LMAX}i}}{n_i} \tag{A.2}$$

最後に，P_i を用いて精度 P_{C} を以下の式で定義する．

$$P_{\mathrm{C}} = \frac{(\sum_{i=1}^{K} P_i) \times 100}{K} [\%] \tag{A.3}$$

上記の計算式を用いて精度を計算する．$\hat{L}_i$ を比較し，重複があった場合は考慮しない．