クラスタ間の距離に基づくカーネルk-平均法

柳　　信一
北海道情報大学

A kernel k-means method based on distance between clusters

Shinichi YANAGI
Hokkaido Information University

平成24年11月

北海道情報大学紀要　第24巻　第１号別刷

〈論　文〉

クラスタ間の距離に基づくカーネルk-平均法

柳 信一*

A kernel k-means method based on distance between clusters

Shinichi Yanagi*

概要

本論文では，カーネルk-平均法においてガウス関数のパラメータを逐次的に計算する手法を提案する．k-平均法は代表的なクラスタリング手法であり，アルゴリズムが単純で実装が容易であるという利点のためクラスタリングを必要とする多くの研究で利用されている．カーネルk-平均法は，データ集合の高次元空間上での線形分離を試みるカーネル法をk-平均法に適用する手法であり，クラスタ同士の境界が非線形であるデータの分類が可能である．しかし，計算結果はカーネル法において距離計算に用いるガウス関数のパラメータに強く依存し，かつ，パラメータ調整が難しいという問題点がある．提案手法は，ガウス関数のパラメータの更新と従来のカーネルk-平均法の計算を交互に繰返し行う．いくつかの人工データと実データを用いて計算機実験を行い，従来のカーネルk-平均法において，入力として適切なパラメータを与えたときに対象データを正確に分類可能な場合は，ほぼ同等の性能であることを示す．

abstract

This paper presents a novel kernel k-means algorithm to compute a Gaussian parameter in the original kernel k-means algorithm repeatedly. The k-means algorithm is a well known clustering algorithm and for its simplification has been used in many works that need to identify clusters. The kernel k-means algorithm which combines a kernel method to the k-means algorithm make possible to identify clusters that are non-linearly separable in input space and can not be identified by the original k-means algorithm. However, a conclusion identifying clusters with the kernel k-means algorithm depends heavily on a chosen Gaussian parameter which used for computing distance between data, and determining an optimal Gaussian parameter is difficult. The proposed algorithm identifies clusters with the original kernel k-means algorithm and improve a Gaussian parameter repeatedly. An experimental comparison in some artificial and real data sets shows that a performance of the proposed algorithm is almost the same as the original kernel k-means algorithm which correctly identifies clusters with an appropriate Gaussian parameter.

keywords: クラスタリング，k-平均法，カーネル法，ガウス関数，パラメータ

*経営情報学部 システム情報学科 講師

1 まえがき

クラスタリングとは，データ集合をある基準によって分類することであり，教師なし学習の一つである．k-平均法[1]は代表的なクラスタリング手法であり，計算結果のクラスタ中心と各クラスタに所属するデータとの距離の総和が最小となるように目的関数を定式化する最適化手法である．最適解を得るためには各クラスタに所属するデータの組み合わせを考慮する必要があり，目的関数は局所解を持つため計算結果の初期値依存が大きい．また，通常は距離計算の方法としてユークリッド距離を使用するため，各クラスタの分離境界は線形となり，このため，データの分布形状において分離を期待するクラスタ間が線形分離可能である必要がある．このような欠点にもかかわらず，複雑な数値計算を必要としないアルゴリズムの単純さ，および，事前に必要とするパラメータがクラスタ数のみでパラメータの調整を必要としないという利点のため，クラスタリングを必要とする計算過程において利用されることが多い．

しかし，実際にはクラスタ同士が分離不可能であるようなデータの分類や，クラスタ同士の分離境界が必ずしも線形とはならないデータの分類を必要とする場面も少なくない．非線形に分離されているデータに対する処理の需要は機械学習分野では大きく，近年，カーネル法と呼ばれる一連の機械学習の手法が提案されてきた[2]．カーネル法とは線形分離不可能なデータ集合を，現在のデータの次元より高次元に非線形変換したとみなし，高次元空間上で内積計算のみを行いデータ集合の線形分離を試みる手法で，パターン認識の識別手法であるサポートベクターマシン[3]やカーネル主成分分析[4]等様々な機械学習に適用されている．

カーネルk-平均法[5]は，k-平均法にカーネル法を適用した手法であり，欠点の一つであった線形分離不可能なクラスタからなるデータへの適用が可能である．しかし，従来のk-平均法よりも局所解の問題が顕著となり，さらに，計算結果が距離計算に使用するカーネル関数のパラメータに敏感であるため，カーネル関数のパラメータの調整が困難であるという問題点がある．初期値依存問題を緩和する手法としては，グラフクラスタリングとk-平均法を併用するスペクトラルクラスタリング[6]が提案されており，スペクトラルクラスタリングはカーネルk-平均法の距離計算においてクラスタのデータの類似度を考慮した重み付きカーネルk-平均法[7]と等価となる．グラフクラスタリング[8;9]は，各データ間に適切な類似度を設定し，データ集合と類似度をグラフとみなして，クラスタリングをカット問題に帰着させる手法である．k-平均法と同様目的関数を最適化するように定式化されるため，分類結果に対する明確な評価基準を有し，さらにクラスタ同士の分離境界が非線形である場合にも対応できる．また，目的関数はレイリー商の形式で表現されるため，組合せ的な難しさは固有値問題に帰着させて緩和できる反面，実際の計算では固有値問題の数値計算を必要とし，さらに，カーネルk-平均法と同様に類似度関数のパラメータ調整が困難であるという問題点がある．他に非線形のデータに対しカーネル法を適用するためパラメータ調整の問題があるクラスタリング手法としては，高次元空間上のデータ集合を覆う超球の半径を最小化するように定式化を行うサポートベクタークラスタリング[10]がある．

上記のカーネル法を適用する手法において使用するカーネル関数は，ほとんどの場合がガウス関数であり，したがってパラメータはデータ間の距離に関するスケール調整の意味を持つ．適切なパラメータを選択する手法としては，Yuら[11]が，カーネルk-平均法によるク

ラスタリングの最適化とカーネル関数のパラメータの最適化を交互に行う手法を提案しパラメータ選択の問題を緩和しているものの，事前に解となるパラメータの有限列を用意する必要があるため依然としてパラメータ選択の困難さが残る．

本論文では，カーネルk-平均法において，クラスタ間の距離に基づきカーネル関数のパラメータを計算するアルゴリズムを提案する．いくつかの人工データと実データを用いて計算機実験を行い性能を検討する．

2 k-平均法

クラスタリングとは，対象とするn個のデータからなるデータ集合$X=\{x_i|i=1,2,\ldots,n,\forall x_i\in R^d\}$を何らかの基準により$k$個の部分集合$C_j=\{x|\forall x\in X\}$に分類することであり，この部分集合$C_j(j=1,2,\ldots,k)$をクラスタと呼ぶ．一般的にクラスタリングを必要とする場面では，データ集合のみが与えられ，クラスタ数kは未知であることが多いが，本論文では既知のものとして取り扱う．

k-平均法はクラスタリング対象である各データ$x_i\in R^d(i=1,2,\ldots,n)$と，クラスタの中心$m_j\in R^d(j=1,2,\ldots,k)$との距離の総和を最小とする最適化問題として式(1)で定式化する．

$$\begin{aligned}&\min_{m_j}\sum_{j=1}^{k}\sum_{x_i\in C_j}\parallel x_i-m_j\parallel^2,\\&\text{s.t.}\bigcup_{j=1}^{k}C_j=X,\\&\qquad C_j\cap C_l=\emptyset,\ 1\le j,l\le k,\ j\neq l.\end{aligned}\tag{1}$$

式(1)の$\|\cdot\|$はL^2ノルムである．最適化問題の解m_jは式(2)となる．

$$m_j=\frac{1}{|C_j|}\sum_{x_i\in C_j}x_i.\tag{2}$$

式(2)の$|C_j|$はC_jに属するデータの数である．式(1)の目的関数は非線形となるため局所解が存在し，さらに，可能なC_jの選択は組合せ的な探索を必要とする離散最適化であるためNP困難となる．したがって，局所解の出力を認める以下の発見的アルゴリズムを用いて計算する．

Algorithm 1 k-平均法　入力：X,k

1. 繰返し番号tを$t=1$で初期化し，k個のクラスタ中心$m_j^{(t)}(j=1,2,\ldots,k)$を初期化する．
2. 各データに対して式(3)を計算し，最も近いクラスタ中心のクラスタ$C_\alpha^{(t)}$に割り当てる．
$$\alpha=\arg\min_j\parallel x_i-m_j^{(t)}\parallel^2.\tag{3}$$
3. 2の結果から式(1)の目的関数を再計算し，変化が無ければ$C_j^{(t)}(j=1,2,\ldots,k)$を出力して計算を終了する．そうでなければ$t=t+1$として2へ戻る．

Algorithm 1 の1行目において，クラスタ中心の初期化の方法はいろいろあり，よく使用されるのは n 個の入力データの中からランダムに k 個選択し，初期クラスタ中心とする方法である．その他に，各クラスタ $C_j^{(1)}(j=1,2,\ldots,k)$ に属するデータ $x \in X$ を式 (1) の制約条件を満たすようにランダムに選択し，各初期クラスタ $C_j^{(1)}(j=1,2,\ldots,k)$ を用いて式 (2) により $m_j^{(1)}(j=1,2,\ldots,k)$ を計算する方法もある．後述の方法は次節のカーネル k-平均法で利用する．また，2行目の式 (3) の計算は，全てのデータ $x_i(i=1,2,\ldots,n)$ に対して各 $m_j(j=1,2,\ldots,k)$ との評価を行うため nk 回の距離計算を必要とし，2行目の処理が終了した時点でクラスタに変化がある場合は，式 (1) の目的関数値も変化することになる．したがって，3行目の停止条件では各 $C_j(j=1,2,\ldots,k)$ が変化しない場合には終了するという条件もよく用いられる．

k-平均法は，アルゴリズムが単純で実装しやすく，各データを必ず1つのクラスタに分類できるという利点を持つ反面，クラスタ数 k を入力とする必要があり，分類結果が初期クラスタ中心に依存するという欠点を持つ．さらに，式 (1) の定式化は暗に各クラスタが凸形状となっていることを仮定しているため，データの分布形状に関する以下のような欠点も持つ．

1. 線形分離不可能なクラスタからなるデータの分類に適していない．

2. 各クラスタの分布が異なるデータの分類に適していない．

3 カーネル k-平均法

カーネル k-平均法[5] は，クラスタ同士を分離する境界が非線形となっているデータに対して k-平均法を適用する手法である．一般的に，n 個のデータからなるデータ集合 $x_1, x_2, \ldots, x_n$ の二つの部分集合は，データ $x \in R^d$ の次元数 d が $n+1$ 以上であれば，大体の場合は線形分離可能となる[12]．そこで，無限次元まで含めた，現在の空間の次元数 d よりはるかに大きい $D(\gg d)$ 次元空間上に，写像 ϕ により $x \in R^d$ を非線形写像することを考える．ここで，機械学習分野では二つのデータ x と x' の内積計算の結果のみを必要とし，内積計算をした後は，データそのもを必要としない場合がある．このことから，$\phi(x)$ 自体の計算は行わず，$\phi(x)$ と $\phi(x')$ の内積計算のみを行うことができれば計算量を削減することができる．つまり，$\phi(x)$ がヒルベルト空間の元となるような写像 ϕ を考え，そのような内積計算を行うことができる式 (4) のような関数が存在すればよい．また，式 (4) のような関数が存在し内積計算のみを必要とする場合は，写像 ϕ を定義する必要もない．

$$K(x, x') = \phi(x) \cdot \phi(x') \,. \tag{4}$$

式 (4) のような関数をカーネル関数と呼ぶ．カーネル関数は対象の類似度を表現するように設計されなければならず，$X \times X$ を定義域とする実対称関数であることとコーシー・シュワルツの不等式を満たす必要がある[3]．カーネル関数を設計できれば，カーネル関数を高次元空間上に写像された2点間の内積とし，それに基づいて L^2 ノルムを定義できる．したがって，上記の二つの条件は，距離の公理を満たす上でも必要な条件であるが，加えてカーネル関数が高次元空間における内積となっていることを保証するマーセルの定理が重要となる[13]．マーセルの定理とは，あるカーネル関数に対して，n 個のデータの各対のカーネル関数値を要素とする行列が半正定値性を満たす性質のことである．このようなカーネル関数はマーセルカーネルとも呼ばれ[2]，代表的なマーセルカーネルとして多項式カーネルやガウスカー

ネルがある．特に多くの研究では，ガウスカーネルが使用される．ガウスカーネルを式 (5) に示す．

$$K(x,x') = \exp\left(\frac{-\parallel x - x' \parallel^2}{\delta^2}\right) . \tag{5}$$

式 (5) において，δはスケールパラメータを表す．式 (4) と式 (5) より，ガウスカーネルにより誘導される写像空間上でのxとx'の距離を式 (6) のようにで定義できる．

$$\begin{aligned}\parallel \phi(x) - \phi(x') \parallel^2 &= K(x,x) - 2K(x,x') + K(x',x') \\ &= 2\left(1 - \exp\left(\frac{-\parallel x - x' \parallel^2}{\delta^2}\right)\right) .\end{aligned} \tag{6}$$

カーネルk-平均法は，各クラスタが高次元空間上で線形分離されていると仮定し，式 (6) の距離計算を用いてk-平均法を行う手法である．式 (1) の最適化問題は式 (7) のように表現される．

$$\begin{aligned}&\min_{M_j} \sum_{j=1}^{k} \sum_{x_i \in C_j} \parallel \phi(x_i) - M_j \parallel^2 , \\ &\text{s.t.} \bigcup_{j=1}^{k} C_j = X , \\ &\quad C_j \cap C_l = \emptyset , \ 1 \leq j,l \leq k , \ j \neq l .\end{aligned} \tag{7}$$

また，得られる最適化問題の解である式 (2) のクラスタ中心は式 (8) のように表現される．

$$M_j = \frac{1}{|C_j|} \sum_{x \in C_j} \phi(x) , \ M_j \in R^D (j = 1,2,\ldots,k) . \tag{8}$$

写像ϕが未定義の場合，式 (8) のクラスタ中心を計算することはできないが，クラスタ中心$M_j \in R^D$とデータ$\phi(x) \in R^D$の距離は式 (9) で計算することができる．

$$\begin{aligned}\parallel M_j - \phi(x) \parallel^2 &= \parallel M_j \parallel^2 - 2M_j \cdot \phi(x) + \phi(x) \cdot \phi(x) \\ &= \frac{1}{|C_j|^2} \sum_{x,x' \in C_j} K(x,x') - \frac{2}{|C_j|} \sum_{x' \in C_j} K(x,x') + K(x,x) .\end{aligned} \tag{9}$$

以上をまとめたカーネルk-平均法のアルゴリズムを以下に示す．

Algorithm 2 カーネルk-平均法　入力：$X, k, \delta, C_1^{(1)}, C_2^{(1)}, \ldots, C_k^{(1)}$

1　繰返し番号tを$t=1$で初期化する．
2　各データに対して各クラスタ$C_j^{(t)}$の中心との距離を式 (9) により計算し，最も近いクラスタ中心のクラスタに割り当てる．
3　2の結果から式 (7) の目的関数を再計算し，変化が無ければ$C_j^{(t)} (j = 1,2,\ldots,k)$を出力して計算を終了する．そうでなければ$t = t+1$として2へ戻る．

Algorithm 2 において，入力$C_j^{(1)} (j = 1,2,\ldots,k)$は初期クラスタであり，対象データ$x \in X$をランダムに割当てたものである．本来$C_j^{(1)} (j = 1,2,\ldots,k)$は，入力$X$に対してアルゴリズム内でランダムに$x \in X$を割当てる場合が多い．しかし，本論文では，5章の提案アルゴリズムで Algorithm 2 を再帰的に利用するため，このような表記とした．

4 ガウスカーネルのパラメーター調整の問題

本論文では最もよく利用されるカーネル関数であるガウスカーネルを用いたカーネルk-平均法に関して検討を行う．まず，通常のクラスタリングは教師ラベルなし学習であるため，クラスタリング結果に対して評価基準が必要となる．k-平均法では式 (1) の目的関数を最小とする$C_j(j=1,2,\ldots,k)$を最適なクラスタリング結果として評価する．しかし，カーネルk-平均法に拡張した場合，式 (7) はδに対して最適化していないため，δの変化に関して式 (7) の目的関数を最小とする$C_j(j=1,2,\ldots,k)$が最適とは限らない．そのため，異なるδに対するクラスタリング結果の評価が難しい．また，最適化のパラメータとしてδを加えて$M_j(1\leq j\leq k)$，$C_j(1\leq j\leq k)$，δに関して式 (7) を解くことも困難である．Yuら[11]はクラスタリングの最適化とは別にパラメータを最適化するための定式化を行い，目的関数の最適化とパラメータの最適化を交互に行いながら適切なパラメータによるクラスタリング結果を出力するアルゴリズムを提案しているが，事前にパラメータの有限列を用意する必要がある．

本論文では，適当な初期パラメータの値から逐次パラメータの更新を計算しながらカーネルk-平均法を行い，適切なパラメータによるクラスタリング結果を出力するアルゴリズムを提案する．Yuら[11]の手法との比較は今後の課題とし，従来のカーネルk-平均法に適切なパラメータを与えた場合の性能との比較を行う．

まず，2章のk-平均法の欠点である1と2を満たすデータに関して，カーネルk-平均法による分類可能性の基準を考える．あらかじめクラスタのラベルが与えられている，つまり，データ集合として正解クラスタ$C_j^t(j=1,2,\ldots,k)$が明示されている$X=C_1^t\cup C_2^t\cup\cdots\cup C_k^t$が与えられるものとする．最終的に$C_j^t(j=1,2,\ldots,k)$と等価となる$C_1,C_2,\ldots,C_k$がクラスタリング結果として得られるような適切な初期クラスタ$C_1^{(1)},C_2^{(1)},\ldots,C_k^{(1)}$を与えたときに，カーネル$k$-平均法により$C_1^t,C_2^t,\ldots,C_k^t$に完全に一致するクラスタリング結果が得られるためのパラメータδの条件について考える．

対象となるデータ集合$X\subseteq R^d$を現在の空間の次元数dより十分大きな$D(\gg d)$次元空間，具体的にはデータ数$|X|$に対して$D\geq|X|+1$であるD次元空間に非線形写像することで，写像した空間では任意の二つのデータの部分集合はほぼ線形分離可能となる[12]．したがって，カーネルk-平均法は，線形分離不可能な形状のクラスタは分離できないというk-平均法の欠点の一つを克服しているように思われる．しかし，たとえ線形分離可能であってもk-平均法の評価基準では，期待すべき結果が得られない場合もある．図1にk-平均法では分類できない線形分離可能な二つのクラスタの例を示す．図1は2次元のデータを通常のk-平均法により分類した結果である．図1において，一般的にはクラスタリングの結果として，左側と右側の塊に分類されることを期待するが，k-平均法により得られる最適なクラスタ中心の垂直二等分線，すなわち，最適なクラスタの境界線は図1に示すような境界線となる．このような結果は，二つのデータにおいて分布形状が異なっているために発生する．図1の場合は，右側のクラスタの分散が左側に比して大きく，また，二つのクラスタ間の距離が十分離れていないことが原因である．非線形変換した高次元空間においても，このような分布に依存する問題が発生しているため，線形分離可能であるにもかかわらず，カーネル関数のパラメータの値によって結果が大きく異なると考えられる．まず，簡単のため二つのクラスタ

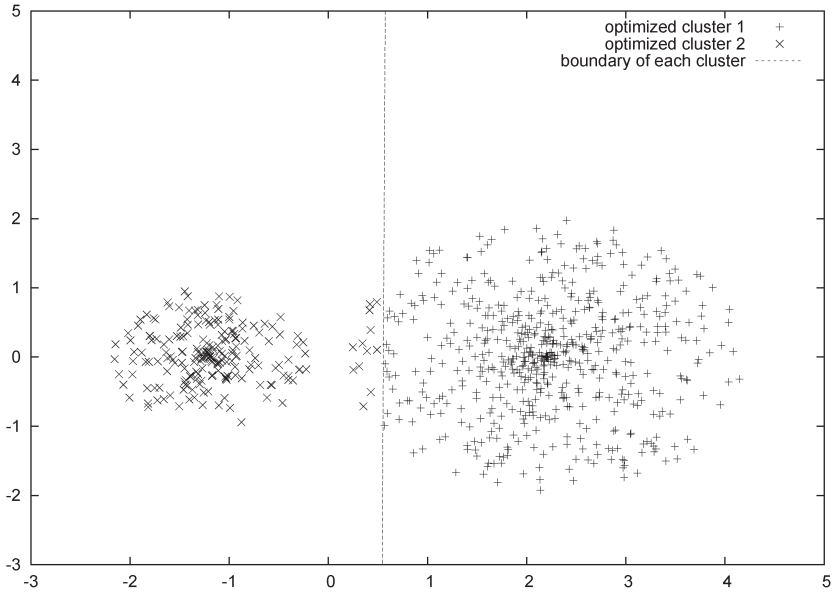

図 1: 線形分離可能な二つのクラスターを分類できない例

に限定し，正解クラスタが既知であるd次元空間上のデータ集合$X = C_a \cup C_b (C_a \cap C_b = \emptyset)$を考える．このとき，$D$次元空間上においてカーネル$k$-平均法で二つのクラスタを分離できる条件は，$D$次元空間上での$C_a$と$C_b$のクラスタ中心点$M_a \in R^D$と$M_b \in R^D$を垂直二等分する超平面が二つのクラスタを分離するようなパラメータδが存在することである．すなわち，任意の$x \in C_a, x' \in C_b$に対して式 (10) を満たすδが存在する必要がある．

$$\left((M_a - M_b)\cdot\phi(x) - \frac{||M_a||^2 + ||M_b||^2}{2}\right)\left((M_a - M_b)\cdot\phi(x') - \frac{||M_a||^2 + ||M_b||^2}{2}\right) < 0\,. \tag{10}$$

式 (4)，式 (5)，式 (8) より，$||M||^2$および$M\cdot\phi(x)$は，ガウス関数の線形和で表現できるため，式 (10) の左辺はガウス関数の2次式までの和で表現できる．形式的には，可能なxとx'の対の選択の仕方は$|C_a||C_b|$通りあるため，式 (10) を満たすδを求めるには$|C_a||C_b|$本の不等式からなる非線形の連立不等式を解く必要がある．ここで，式 (10) の左辺は三つの変数$x \in C_a, x' \in C_b, \delta$に関する関数とみなせるので，式 (10) の左辺を$g_{ab}(x, x', \delta)$という関数に置き換え，式 (10) を満たすδの集合Δ_{ab}を式 (11) で定義する．

$$\Delta_{ab} = \left\{\, \delta \;\middle|\; g_{ab}(x, x', \delta) < 0, \forall x \in C_a, \forall x' \in C_b \right\}\,. \tag{11}$$

したがって，あらかじめ与えられた二つのクラスタC_aとC_bに関して，適切な初期クラスタを与えたときカーネルk-平均法の計算結果としてC_aとC_bを得るための必要十分条件は$\Delta_{ab} \neq \emptyset$となる．

次に，二つ以上の複数クラスタの場合を考える．与えられたデータ集合Xにあらかじめラベルが割り当てられており，k個のクラスタ$C_1, C_2, \ldots, C_k$によりXが構成されているものとする．すなわち，$X = C_1 \cup C_2 \cup \cdots \cup C_k$ かつ 任意の二つのクラスタ$C_i (i = 1, 2, \ldots, k)$と

$C_j(j=1,2,\ldots,k)$ において $C_i \cap C_j = \emptyset (i \neq j)$ とする．k 個のクラスタで構成されるデータ内の任意の二つのクラスタ C_i と C_j の Δ_{ij} は，残りのクラスタに所属するデータ $x \in X - C_i - C_j$ には依存しないため，複数のクラスタからなるデータ集合のクラスタに対しても単純に二つのクラスタ C_i と C_j に式 (11) を適用できる．したがって，各クラスタの対に関して独立に $\Delta_{ij}(1 \leq i < j \leq k)$ を定義できる．k 個のクラスタ $C_1, C_2, \ldots, C_k$ からなるデータ集合 X が与えられたとき，適切な初期クラスタを与えることで，カーネル k-平均法の計算結果として $C_1, C_2, \ldots, C_k$ を得るための必要十分条件は式 (12) となる．

$$\Delta = \bigcap_{1 \leq i < j \leq k} \Delta_{ij} \neq \emptyset \,. \tag{12}$$

与えられた $X = C_1 \cup C_2 \cup \cdots \cup C_k$ に対して，式 (12) を満たす δ は存在しない場合も考えられ，その場合，どのような初期クラスタを与えても，ガウス関数によるカーネル k-平均法では $C_1, C_2, \ldots, C_k$ に完全に一致するクラスタリング結果は得られないことになる．このように，Algorithm 2 において，初期クラスタ $C_j^{(1)}(j=1,2,\ldots,k)$ と δ の組合せの調整だけでは得られるクラスタリング結果には限界がある．

5 提案手法

Algorithm 2 を再帰的に利用しながら δ の更新を行う方法を提案する．まず，十分小さな適切なガウス関数のパラメータの初期値 δ_1 を決定する．次に，δ_1 を用いてカーネル k-平均法のアルゴリズムを実行し，5.1 節で述べる停止条件を満たしていれば計算を終了し，そうでなければガウス関数のパラメータを更新し，再びカーネル k-平均法のアルゴリズムを実行する．ガウス関数のパラメータの更新に関しては，最適化による解析ではなく，単純に増加させていくという戦略を用いる．すなわち，式 (13) を満たすように $q+1$ 番目のパラメータである δ_{q+1} の更新を行う．

$$\delta_{q+1} = \delta_q + \epsilon_q, (\epsilon_q \geq 0) \,. \tag{13}$$

以下，停止条件とパラメータの更新幅 ϵ_q の適切な値について考える．

5.1 停止条件

簡単のため，二つのクラスタ C_a と C_b の局所的な変化について考える．対象とするデータは，現在の d 次元空間より十分大きな D 次元空間 $(d \ll D)$ に非線形変換することで，どのような二つの部分集合でも大体の場合は線形分離できるが，うまく分離できない理由は図 1 に示すような分布形状に依存する問題が発生しているためと考えられる．そこで，ガウス関数のパラメータ δ の値を調整することで D 次元空間上で線形分離できるような分布を探索することを考える．D 次元空間上での 2 点間の距離は δ の値に対して単調減少となるため，δ の値を単調に増加させていく戦略においては，データ全体の分布形状を収縮させながら，二つのクラスタを分離できるようなクラスタ中心 $M_a \in R^D$ と $M_b \in R^D$ の位置を探索していくということになる．ここで，一方のクラスタは本来属するべきデータを完全に含んでおり，他方のクラスタに属するべきデータの部分集合が付加している図 1 のような状態を考える．このような状態は，パラメータの調整により，本来属するべきクラスタに分類できる可能性が高く，Algorithm 2 において，適切な初期クラスタが与えられたにもかかわらずパラ

メータが適切ではなかったため得られるクラスタリング結果の状態と考えることができる．十分小さいδに関しては，D次元空間での分布全体が膨張していると考えられるため，余分なデータを含むクラスタの方が，各データとクラスタ中心との距離の平均は大きくなると仮定する．そのようなクラスタをC_aとし，もう一方をC_bとすると上記の仮定は式 (14) を満たすことを意味する．

$$\frac{1}{|C_a|}\sum_{x\in C_a}||M_a-\phi(x)||^2 \geq \frac{1}{|C_b|}\sum_{x\in C_b}||M_b-\phi(x)||^2 . \tag{14}$$

そこで，C_aを対象クラスタとし，現在対象クラスタのクラスタ中心との距離の方が小さいデータがパラメータの更新により，非対象クラスタとのクラスタ中心との距離の方が小さくなることを考える．ここで，クラスタC_aとクラスタC_bの最小距離$d_{min}(C_a,C_b)$を式 (15) で定義する．

$$d_{min}(C_a,C_b)=\min_{x\in C_a,x'\in C_b}||x-x'||^2 . \tag{15}$$

式 (15) を満たす$x\in C_a$をx_a，$x'\in C_b$をx_bとすると，$x_a\in C_a$が移動の対象と考えているデータである．いま，ただひとつのデータx_aが本来のクラスタC_bではなくC_aに所属しており，本来のクラスタ$C_a-\{x_a\}$と$C_b\cup\{x_a\}$の間は十分離れていると仮定する．このとき，x_aの移動が発生する前のクラスタ間の中心点の距離$||M_a-M_b||^2$と移動が発生した後の距離$||M_a^{new}-M_b^{new}||^2$は大きく変化すると考えられる．そこで，$||M_a-M_b||^2$と$||M_a^{new}-M_b^{new}||^2$を導く．まず，式 (8) より，$||M_a-M_b||^2$は式 (16) となる．

$$\begin{aligned}||M_a-M_b||^2 &= ||M_a||^2-2\,M_a\cdot M_b+||M_b||^2\\ &=\frac{1}{|C_a|^2}\sum_{x,x'\in C_a}K(x,x')-\frac{2}{|C_a||C_b|}\sum_{x\in C_a,x'\in C_b}K(x,x')+\frac{1}{|C_b|^2}\sum_{x,x'\in C_b}K(x,x') .\end{aligned} \tag{16}$$

次に，上記の仮定よりM_a^{new}とM_b^{new}は式 (17)，(18) となる．

$$M_a^{new}=\frac{1}{|C_a|-1}\sum_{x\in C_a-\{x_a\}}\phi(x) , \tag{17}$$

$$M_b^{new}=\frac{1}{|C_b|+1}\sum_{x\in C_b\cup\{x_a\}}\phi(x) . \tag{18}$$

したがって，$||M_a^{new}-M_b^{new}||^2$は式 (19) となる．

$$\begin{aligned}||M_a^{new}-M_b^{new}||^2 &= ||M_a^{new}||^2-2M_a^{new}\cdot M_b^{new}+||M_b^{new}||^2\\ &= \frac{1}{(|C_a|-1)^2}\left(|C_a|^2||M_a||^2-2\sum_{x\in C_a}K(x,x_a)-K(x_a,x_a)\right)\\ &\quad -2\frac{1}{(|C_a|-1)(|C_b|+1)}\left(|C_a||C_b|(M_a\cdot M_b)+\sum_{x\in C_a}K(x,x_a)-\sum_{x\in C_b}K(x,x_a)\right)\\ &\quad +\frac{1}{(|C_b|+1)^2}\left(|C_b|^2||M_b||^2+2\sum_{x\in C_b}K(x,x_a)+K(x_a,x_a)\right) .\end{aligned} \tag{19}$$

ここで，式 (16) と式 (19) はともに同じδの値で計算することに注意する．繰り返し計算の各段階において，$||M_a-M_b||^2$と$||M_a^{new}-M_b^{new}||^2$を計算し，$||M_a^{new}-M_b^{new}||^2/||M_a-M_b||^2$の比率が，ある値をこえたとき$\delta_q$の更新を終了し，その時点でのクラスタリングの結果を最終的な計算結果とする．

5.2 パラメータの更新幅

前節と同様，二つのクラスタ C_a と C_b の局所的な変化について考える．現在の C_a と C_b の最短距離となる $x_a \in C_a$ と $x_b \in C_b$ に関して，$\delta_{q+1} = \delta_q + \epsilon_q$ のパラメータ更新により，以下のクラスタの更新が発生すると仮定する．

$$C_a^{(q+1)} = C_a^{(q)} - \{x_a\}\,, \tag{20}$$

$$C_b^{(q+1)} = C_b^{(q)} \cup \{x_a\}\,. \tag{21}$$

ここで，$C_a^{(q)}$ は q 番目のクラスタの計算結果である．式 (20)，(21) の更新が発生するような ϵ_q を推定する．これは，以下の式 (22)，(23) を満たす必要がある．

$$||M_a^{(q)} - \phi_q(x_a)||^2 \leq ||M_b^{(q)} - \phi_q(x_a)||^2\,, \tag{22}$$

$$||M_a^{(q+1)} - \phi_{q+1}(x_a)||^2 \geq ||M_b^{(q+1)} - \phi_{q+1}(x_a)||^2\,. \tag{23}$$

上記と同様，$M_a^{(q)} \in R^D$ はクラスタ $C_a^{(q)}$ の D 次元空間上でのクラスタ中心であり，δ_q をパラメータとするガウスカーネルにより内積計算を行うことを意味する．また，$M_a^{(q+1)}$ もクラスタ $C_a^{(q)}$ の D 次元空間上でのクラスタ中心で，δ_{q+1} をパラメータとするガウスカーネルにより内積計算を行う．$M_a^{(q)}$ と $M_a^{(q+1)}$ の違いは，あくまで内積計算時のガウスカーネルのパラメータであり，クラスタに所属するデータは変化していない．また，$\phi_q(x_a)$ はデータ x_a の D 次元空間上への写像であり，ガウスカーネルによる内積計算時はパラメータとして δ_q を使用することを意味する．式 (22)，(23) を式 (24)，(25) のように変形する．

$$||M_a^{(q)}||^2 + 2\,M_b^{(q)} \cdot \phi_q(x_a) \leq ||M_b^{(q)}||^2 + 2\,M_a^{(q)} \cdot \phi_q(x_a)\,, \tag{24}$$

$$||M_a^{(q+1)}||^2 + 2\,M_b^{(q+1)} \cdot \phi_{q+1}(x_a) \geq ||M_b^{(q+1)}||^2 + 2\,M_a^{(q+1)} \cdot \phi_{q+1}(x_a)\,. \tag{25}$$

ここで，q 番目のパラメータ δ_q に関して，D 次元空間上でのクラスタ $C^{(q)}$ の中心 $M^{(q)} \in R^D$ のノルムの 2 乗 $||M^{(q)}||^2$，および，$M^{(q)}$ と $\phi_q(x)$ との内積 $M^{(q)} \cdot \phi_q(x)$ は式 (26)，(27) のように記述できる．

$$||M^{(q)}||^2 = \frac{1}{|C|^2} \sum_{\forall\, x,x' \in C} \phi_q(x) \cdot \phi_q(x') = \frac{1}{|C|^2} \sum_{\forall\, x,x' \in C} \exp\left(-\frac{||x - x'||^2}{\delta_q^2}\right) = \frac{1}{|C|^2} \sum_{\forall\, x,x' \in C} K_q(x, x')\,, \tag{26}$$

$$M^{(q)} \cdot \phi_q(x) = \frac{1}{|C|} \sum_{\forall\, x' \in C} \phi_q(x) \cdot \phi_q(x') = \frac{1}{|C|} \sum_{\forall\, x' \in C} \exp\left(-\frac{||x - x'||^2}{\delta_q^2}\right) = \frac{1}{|C|} \sum_{\forall\, x' \in C} K_q(x, x')\,. \tag{27}$$

式 (26) と式 (27) は，いずれもカーネル関数 $K_q(x, x')$ の和で表現されており，$K_q(x, x')$ は δ_q をパラメータとしてガウスカーネルによる計算を行うことを意味する．また，$\delta_{q+1} \geq \delta_q$，かつ，ガウス関数がパラメータ δ に対して増加関数であることより式 (28) が成立する．

$$\forall\, x, x' \in R^d, \quad K_{q+1}(x, x') \geq K_q(x, x')\,. \tag{28}$$

式 (25)〜式 (28) により式 (29) を得る．

$$||M_a^{(q+1)}||^2 + 2\,M_b^{(q+1)} \cdot \phi_{q+1}(x_a) \geq ||M_b^{(q)}||^2 + 2\,M_a^{(q)} \cdot \phi_q(x_a)\,. \tag{29}$$

さらに，$\delta_q > \epsilon_q$ を条件とし，以下の近似を行う．

$$\begin{aligned} K_{q+1}(x,x') &= \exp\left(-\frac{||x-x'||^2}{{\delta_{q+1}}^2}\right) = \exp\left(-\frac{||x-x'||^2}{(\delta_q+\epsilon_q)^2}\right) = \exp\left(-\frac{||x-x'||^2}{{\delta_q}^2\left(1+\frac{\epsilon_q}{\delta_q}\right)^2}\right) \\ &\simeq \exp\left(-\frac{||x-x'||^2}{{\delta_q}^2}\left(1-2\,\frac{\epsilon_q}{\delta_q}\right)\right) \\ &= \exp\left(2\,\frac{||x-x'||^2}{{\delta_q}^3}\epsilon_q\right) K_q(x,x')\,. \end{aligned} \tag{30}$$

したがって，$d_{max}(C_a) = \max\limits_{x,x'\in C_a} ||x-x'||^2$ とすると，式 (26)，(30) より式 (31) が成立する．

$$\exp\left(2\,\frac{d_{max}(C_a)}{{\delta_q}^3}\,\epsilon_q\right) ||M_a^{(q)}||^2 \geq ||M_a^{(q+1)}||^2\,. \tag{31}$$

同様に，$d_{max}(x_a,C_b) = \max\limits_{x\in C_b} ||x-x_a||^2$ とすると，式 (27)，(30) より式 (32) が成立する．

$$\exp\left(2\,\frac{d_{max}(x_a,C_b)}{{\delta_q}^3}\,\epsilon_q\right) M_b^{(q)}\cdot\phi_q(x_a) \geq M_b^{(q+1)}\cdot\phi_{q+1}(x_a)\,. \tag{32}$$

$d_{max}(C_a,x_a,C_b) = \max\{d_{max}(C_a), d_{max}(x_a,C_b)\}$ とすると，式 (31) と式 (32) より，式 (29) は式 (33) のように記述できる．

$$\exp\left(2\,\frac{d_{max}(C_a,x_a,C_b)}{{\delta_q}^3}\,\epsilon_q\right)(||M_a^{(q)}||^2 + 2\,M_b^{(q)}\cdot\phi_q(x_a)) \geq ||M_b^{(q)}||^2 + 2\,M_a^{(q)}\cdot\phi_q(x_a)\,. \tag{33}$$

ここで，$A_q = ||M_a^{(q)}||^2 + 2\,M_b^{(q)}\cdot\phi_q(x_a)$，$B_q = ||M_b^{(q)}||^2 + 2\,M_a^{(q)}\cdot\phi_q(x_a)$ とすると，式 (24) より $\frac{B_q}{A_q} \geq 1$ となるので，式 (33) の両辺に関して対数をとり以下のように整理する．

$$\begin{aligned} &\exp\left(2\,\frac{d_{max}(C_a,x_a,C_b)}{{\delta_q}^3}\,\epsilon_q\right) \geq \frac{B_q}{A_q}\,, \\ &2\,\frac{d_{max}(C_a,x_a,C_b)}{{\delta_q}^3}\,\epsilon_q \geq \log\frac{B_q}{A_q}\,, \\ \therefore\quad &\epsilon_q \geq \frac{{\delta_q}^3}{2d_{max}(C_a,x_a,C_b)}\log\frac{B_q}{A_q}\,. \end{aligned} \tag{34}$$

式 (33) の不等式は，かなり緩い評価であり，δ_q から δ_{q+1} に変更しても，クラスタの所属データの更新は発生しない場合もある．特に初期値 δ_1 をかなり小さな値に設定する場合は，ϵ_1 も相当小さくなるため収束速度がかなり遅くなり，実用的に好ましくない．そこで，まず，クラスタ C_a に所属するデータ間の最短距離を $d_{min}(C_a)$，データ x_a とクラスタ C_b に所属するデータとの最短距離を $d_{min}(x_a,C_b)$，クラスタ C_a に所属するデータ間の距離の平均を $d_{ave}(C_a)$，データ x_a とクラスタ C_b に所属するデータとの距離の平均を $d_{ave}(x_a,C_b)$ として式 (35) のよう

に $d_1 \sim d_6$ を定義する．

$$
\begin{aligned}
d_1 &= \min\{d_{min}(C_a), d_{min}(x_a, C_b)\}\,, \\
d_2 &= \max\{d_{min}(C_a), d_{min}(x_a, C_b)\}\,, \\
d_3 &= \min\{d_{ave}(C_a), d_{ave}(x_a, C_b)\}\,, \\
d_4 &= \max\{d_{ave}(C_a), d_{ave}(x_a, C_b)\}\,, \\
d_5 &= \min\{d_{max}(C_a), d_{max}(x_a, C_b)\}\,, \\
d_6 &= \max\{d_{max}(C_a), d_{max}(x_a, C_b)\} = d_{max}(C_a, x_a, C_b)\,.
\end{aligned} \tag{35}
$$

次に，式 (34) の不等式の右辺において，$d_{max}(C_a, x_a, C_b)$ をパラメーター d_i で置き換えた式 (36) により更新幅 ϵ_q を $\epsilon_q = \epsilon_q(d_i)$ として決定する．

$$
\epsilon_q(d_i) = \frac{{\delta_q}^3}{2\,d_i} \log \frac{B_q}{A_q}\,, (i = 1, 2, \ldots, 6)\,. \tag{36}
$$

式 (36) において，d_i の値を大きくとれば $\epsilon_q(d_i)$ の値は小さくなり，d_i の値を小さくとれば $\epsilon_q(d_i)$ の値は大きくなるという関係がある．また，d_i の値を d_6 以外にした場合は，必ずしも式 (33) の不等式を満たさないため，せいぜい一つデータに関してクラスタの所属が変化するという仮定が成立しない場合も考えられる．式 (36) の d_i の選択方法に関しては，事前に安全係数 $p(0 < p \leq 0.5)$ を設定しておき，$\epsilon_q(d_i) \leq p\,\delta_q\ (i = 1, 2, \ldots, 6)$ を満たす最大の d_i を採用し $\epsilon_q(d_i)$ を更新幅とする．このように $\delta_{q+1} = \delta_q + \epsilon_q(d_i)$ と決定することで δ_q の値が小さいときの収束速度の改善を期待できる．

5.3 アルゴリズム

次に，二つ以上の複数クラスタへの拡張を考える．簡単のため対象データがカーネル k-平均法で完全に分離できる，つまり，$\Delta \neq \emptyset$ と仮定する．一般的には k 個の各クラスタの影響を考慮して最適に分割できる δ を検討する必要があるが，任意の二つのクラスタ C_i と C_j を分離できるパラメータ集合 Δ_{ij} は与えられたデータ集合において，他のクラスタに所属するデータに依存せず，また，式 (12) を満たす Δ が存在するなら，任意の二つのクラスタ C_i と C_j に関して，$\Delta \cap \Delta_{ij} \neq \emptyset$ となる．このことより，特定の二つのクラスタに着目し δ_q の値を更新していく．二つのクラスタの選択方法に関しては，現在得られている k 個の各クラスタの C_i に関して，他のクラスタ C_j とのクラスタ中心の距離を計算し，その最小値となるクラスタ番号 $min(i) = \arg\min_{j \neq i} ||M_i - M_j||^2$ を求める．つまり，$min(i)$ は i 番目のクラスタ C_i に一番近いクラスタ $C_{min(i)}$ の番号である．次に，求めた k 個のクラスタ番号の対 i と $min(i)$ に関して，クラスタ中心と所属するデータとの距離の平均 $d_m(C_i)$ を式 (37) により計算する．

$$
d_m(C_i) = \frac{1}{|C_i|} \sum_{x \in C_i} ||M_i - \phi(x)||^2 = \frac{1}{|C_i|} \sum_{x \in C_i} (||M_i||^2 - 2M_i \cdot \phi(x) + 1)\,. \tag{37}
$$

5.1 節におけるアルゴリズムの停止基準の戦略に従い，$d_m(C_i)$ と $d_m(C_{min(i)})$ の差が大きな二つのクラスタを δ_q 更新のためのクラスタとして採用する．

$$
\begin{aligned}
u &= \arg\max_i \left\{ \frac{\max\{d_m(C_i), d_m(C_{min(i)})\}}{\min\{d_m(C_i), d_m(C_{min(i)})\}} \right\}\,, \\
v &= min(u)\,.
\end{aligned} \tag{38}
$$

式 (38) のvは，クラスタC_uの中心との距離が最短となる中心を持つクラスタC_vの番号である．式 (38) により，パラメータδ_qの更新のための計算対象とする二つのクラスタ番号uとvを求める．最終的に対象クラスタの番号aを式 (39) により求める．

$$a = \begin{cases} u & (d_m(C_u) \geq d_m(C_v)) \\ v & (\text{otherwise}) \end{cases} . \tag{39}$$

対象クラスタC_aと対となるクラスタC_bの番号bは，式 (39) において，aの番号として採用しなかったuとvのいずれかとする．以上をまとめたアルゴリズムを以下に示す．

Algorithm 3 提案手法　入力：X, k, δ

1 繰返し番号qを$q=1$で初期化し，$\delta_q = \delta$とする．
2 n個の対象データ$x_i \in X (i = 1, 2, \ldots, n)$をランダムに$k$個の部分集合$C_j^{(q)} \subset X (j = 1, 2, \ldots, k)$に分類し，$C_1^{(q)}, C_2^{(q)}, \ldots, C_k^{(q)}$を初期クラスタとする．
3 対象データX，k個のクラスタ$C_1^{(q)}, C_2^{(q)}, \ldots, C_k^{(q)}$，および，$\delta_q$を入力として**Algorqthm 2**を実行する．この計算により，再計算されたクラスタ$C_1^{(q)}, C_2^{(q)}, \ldots, C_k^{(q)}$が出力される．
4 終了フラグが立っていれば 9 へ進み，それ以外の場合は 5 へ進む．
5 式 (38) と (39) により，対象クラスタの対C_aとC_bを選択する．
6 式 (16) と (19) により停止条件の計算を行い，$||m_a^{new} - m_b^{new}||^2 / ||m_a - m_b||^2 < r_{low}$なら 9 へ進み，$||m_a^{new} - m_b^{new}||^2 / ||m_a - m_b||^2 > r_{up}$なら終了フラグを立てる．
7 式 (35) と (36) により，ϵ_qを計算する．
8 $\delta_q = \delta_q + \epsilon_q$，$q = q + 1$と更新し 3 へ戻る．
9 k個のクラスタ$C_1^{(q)}, C_2^{(q)}, \ldots, C_k^{(q)}$を出力し計算を終了する．

アルゴリズム 3 の 3 行目では，局所解に陥る可能性を考慮して，繰返し毎にq番目のクラスタ$C_j^{(q)} (j = 1, 2, \ldots, k)$をランダムに初期化することも有力だと思われるが，本研究では単純にq番目のクラスタリング結果を$q+1$番目の入力とした．また，停止条件に関して，収束速度を重視した式 (36) でパラメータの更新幅を決定することにより，実際にはk個のクラスタに関して必ずただ一つのデータのクラスタ所属の変更が行われるとは限らず，複数のデータによる変更が行われ，$||m_a^{new} - m_b^{new}||^2 / ||m_a - m_b||^2$の値が急激に減少する変化も考えられる．この場合は，本来クラスタC_aに所属するべきデータまでクラスタC_bに割り当てられてしまったと考え，6 行目で計算を終了するようにした．6 行目のr_{low}とr_{up}はその停止基準である．また，4 行目の終了フラグは，2 種類の停止基準の分岐を行うためのものである．

6 実験

実験は，教師ラベル付きの人工データと実データに対して行う．クラスタリングは本来教師ラベルなし学習であるが，教師ラベル付きのデータにより性能評価を行うことも多い．教師ラベル付きのデータのクラスタリング結果に対して評価を行う代表的な指標として，Adjusted Rand Index(ARI)[14] や Normalized Mutual Information(NMI)[15] がある．本論文では NMI を評価指標として用いる．あらかじめ教師ラベルにより決定されているk個の真のクラスタを$C_1^t, C_2^t, \ldots, C_k^t$とし，実際にクラスタリングすることにより得られる$k$個のクラスタを

$C_1, C_2, \ldots, C_k$ とする．また，$X_{ij} = C_i \cap C'_j (i, j = 1, 2, \ldots, k)$ とするときNMIの計算式を式 (40) に示す．

$$\mathrm{NMI} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} \frac{|X_{ij}|}{n} \log \frac{n|X_{ij}|}{|C_i||C'_j|}}{\sqrt{\left(\sum_{i=1}^{k} \frac{|C_i|}{n} \log \frac{|C_i|}{n}\right)\left(\sum_{j=1}^{k} \frac{|C'_j|}{n} \log \frac{|C'_j|}{n}\right)}} . \tag{40}$$

式 (40) において，$n = |C'_1 \cup C'_2 \cup \cdots \cup C'_k|$ はデータ数である．NMIは0と1の間の値をとり，クラスタリング結果 $C_i (i = 1, 2, \ldots, k)$ と正解クラスタ $C'_j (j = 1, 2, \ldots, k)$ が完全に一致したとき最大値の1をとる．NMIの最大値が1となることの概略は以下の通りである．まず，式 (40) の分母は，データ数 n に対するクラスタ数をそのクラスタの生起確率と考えたときの，正解クラスタと実際に得られたクラスタのエントロピーの積の平方根となる．すなわち，正解クラスタの集合を C^t，実際に得られたクラスタの集合を C とすると，各エントロピーは式 (41) となる．

$$\begin{aligned} H(C^t) &= -\sum_{j=1}^{k} \frac{|C'_j|}{n} \log \frac{|C'_j|}{n} , \\ H(C) &= -\sum_{i=1}^{k} \frac{|C_i|}{n} \log \frac{|C_i|}{n} . \end{aligned} \tag{41}$$

また，正解クラスタ C'_j に所属し，かつ，実際のクラスタリング結果として C_i に所属するデータが得られる確率を $\frac{|X_{ij}|}{n}$ とするとき，C^t と C の結合エントロピーは式 (42) となる．

$$H(C^t, C) = -\sum_{i=1}^{k}\sum_{j=1}^{k} \frac{|X_{ij}|}{n} \log \frac{|X_{ij}|}{n} . \tag{42}$$

$\frac{|C_i|}{n} = \sum_{j}^{k} \frac{|X_{ij}|}{n}$ より，式 (40) の分子は相互情報量 $I(C^t; C) = H(C^t) + H(C) - H(C^t, C)$ となる．したがって，式 (40) は式 (43) のように記述できる．

$$\mathrm{NMI} = \frac{H(C^t) + H(C) - H(C^t, C)}{\sqrt{H(C^t)H(C)}} . \tag{43}$$

正解クラスタと実際に得られたクラスタが完全に一致するときは，$H(C^t) = H(C) = H(C^t, C)$ となるので式 (43) より，式 (40) の値は1となる．

Algorithm 3 の停止基準のパラメータ r_{low} と r_{up} に関しては，それぞれ $r_{low} = 0.8$，$r_{up} = 1.2$ とした．また，式 (35) において，$d_1 \sim d_6$ のいずれかを選択するための安全係数 p に関しては，0.5と0.1の2種類に関して試行を行った．実験で使用する人工データと実データに関して，事前に各データにおける各次元の標準偏差の平均 s を計算し，$s/5$ と $s/2$ の2種類をそのデータに対する δ の初期値とした．さらに，計算が終了しない場合を考慮し最大繰返し回数を100回とし，100回の繰返しを越えても計算を終了しない場合は強制的に計算を終了するようにした．

以上の条件で，提案手法であるAlgorithm 3を用いて，各データに対して50回の試行を行い，δ の更新回数に関する平均と最小と最大，δ の更新毎のAlgorithm 2における繰返し回数の平均および1回の繰返しで停止した回数の平均，最大更新回数内に停止しなかった回数の平均，およびNMI値の平均と最小と最大を評価した．

また，比較のため，従来手法である Algorithm 2 を用いて，適切なδを5つ選択し，各データに対して 50 回の試行を行い，式 (7) の目的関数が最小となる NMI 値の平均と最小と最大を調べた．δ の選択は以下のように行う．まず，真のクラスタ $C_j^t(j=1,2,\ldots,k)$ が既知の場合，Algorithm 2 の入力の初期クラスタとして $C_j^t(j=1,2,\ldots,k)$ を与えることで，以下に示す試行により Δ の推定が可能となる．適当な計算回数 Q を決め，$\delta_1,\delta_2,\ldots,\delta_Q$ に関して，$(X,k,\delta_q,C_j^t(j=1,2,\ldots,k))$ を入力として Q 回 $(1\leq q\leq Q)$ の試行を行う．もし，$\delta_q\in\Delta$ であれば，クラスタ中心は変化していないため，Algorithm 2 は 1 回だけクラスタ中心の再計算を行い停止する．また，そのときの NMI 値は 1 となる．したがって，$\delta_Q-\delta_1$ を十分大きくとり $\delta_{q+1}-\delta_q$ を適切に設定し，δ_q に関する NMI 値を観察することで，ある程度の Δ の推定が可能となる．ここで，$\delta_1,\delta_2,\ldots,\delta_Q$ は昇順としている．各データに関して，NMI 値が 1 となる δ の範囲を推定できた場合は，その範囲から任意に 5 つの δ を選択し，NMI 値が 1 となる δ の範囲を推定できない場合は NMI 値が最も大きくなる範囲を推定し，その範囲から任意に 5 つの δ を選択する．

6.1 人工データ

まず，図 2 に示すような 2 つのクラスタからなるデータに関する実験を行う．各クラスタのデータ数は，ともに 200 個で標準偏差は $s=0.68$ である．

図 2 のデータにおいて，入力として真のクラスタを与えたときのガウス関数のパラメータに関する NMI 値の変化を確認する．ガウス関数のパラメータを 0 ∼ 2 まで 0.01 間隔で変化させたときのガウス関数のパラメータと NMI 値の関係を図 3 に示す．図 3 において横軸がガウス関数のパラメータ δ，縦軸が NMI を示す．図 3 よりガウス関数のパラメータが 0.4 ∼ 1.2 を少し越えたあたりまでの範囲のとき，真のクラスタと計算結果のクラスタが完全に一致する可能性がある．また，ガウス関数のパラメータが 0.01 ∼ 0.03 で NMI 値が最大値の 1 となっており，0.04 で急激に減少しているが，これはガウス関数のパラメータが小さすぎてガウス関数の値が計算可能な有効桁以下となってしまい，そのため計算した距離が常に最大値である 2 を出力し，結果として最短距離を求める計算がうまく動作しなかったためである．以降このような δ の範囲を計算不能範囲と考える．

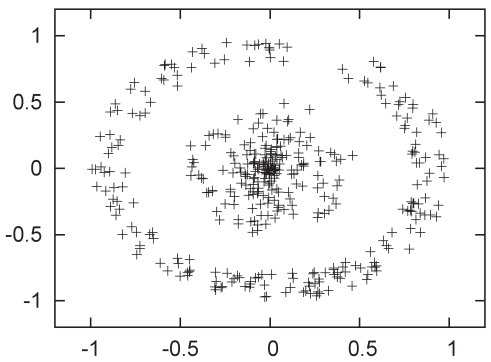

図 2: 2 つのクラスタからなるデータ

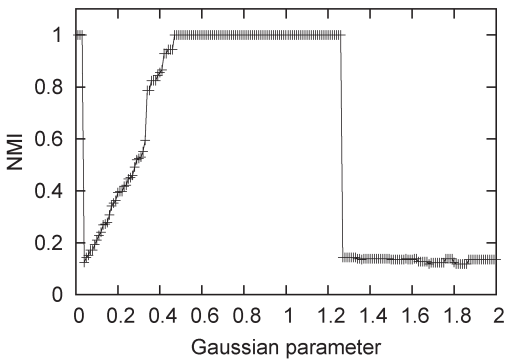

図 3: 図 2 のデータの δ と NMI の関係

表1に図2の人工データの実験結果を示す．表1の最終行の太字が従来手法の結果である．表1より，二つのクラスタからなるデータにおいてデータ数が同じである場合は，提案手法のNMIは各pと初期δの組合せに関して全て1となっており，従来手法に勝っている．δの更新回数は，$p=0.5$のときの方が少なくなっており，式(35)の中から収束速度が速いd_iを選択するという改善がうまく機能したと考えられる．$p=0.5$と$p=0.1$の場合を比較すると，各δの更新時において，Algorithm 2が1回の繰返し計算のみで停止している場合が多くなっており，クラスタの所属データの変更が行われないようなδの更新が行われている．100回の更新上限に達した試行はないため，停止条件がよく機能したと考えられる．図3より，NMI= 1となるδが広範囲であり，その範囲で停止する機会が比較的多かったことが全体的に好ましい結果となった要因であると考えられる．

表 1: 図2の人工データの実験結果

p	初期δ	δの更新回数			Algo. 2の繰返し数		未停止	NMI		
		最小	平均	最大	回数平均	1回平均		最小	平均	最大
0.5	$s/5$	18	21.46	25	3.122	9.56	0	1.000	1.000	1.000
0.5	$s/2$	5	5.62	7	3.435	3.86	0	1.000	1.000	1.000
0.1	$s/5$	62	64.28	68	2.058	37.9	0	1.000	1.000	1.000
0.1	$s/2$	37	40.88	44	1.596	31.74	0	1.000	1.000	1.000
-	-	-	-	-	-	-	-	**0.14**	**0.513**	**1.0**

次に，図4に示すようなデータ数が異なる2つのクラスタからなるデータに関する実験を行う．各クラスタのデータ数は，内側のクラスタが400個，外側のクラスタが100個で標準偏差は$s=0.68$である．上記の図2のデータと同様，図4の人工データにおいて，ガウス関数のパラメータを0.2 ~ 2.2まで0.01間隔で変化させたときのガウス関数のパラメータとNMI値の関係を図5に示す．図5より図4のデータは$\Delta=\emptyset$であると考えられる．表2に図4の人工

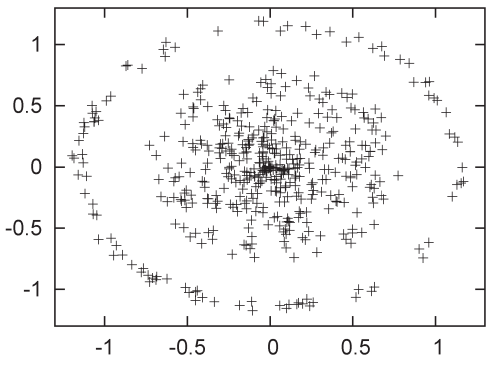

図 4: データ数が異なる2つのクラスタからなるデータ

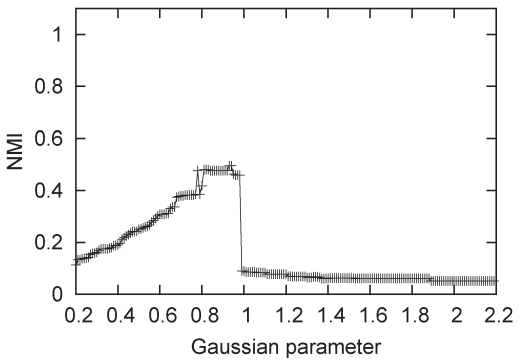

図 5: 図4のデータのδとNMIの関係

データの実験結果を示す．表2より，二つのクラスタからなるデータにおいてデータ数が異なる場合は，NMIは各pと初期δの組合せに関して，おおよそ最大値である1の半分以下の

数値となっている．未停止数が全て0であることから，停止条件は働いているものの，$\Delta = \emptyset$であると推定されるため，図4において二つのクラスタを高次元空間上で線形分離できるカーネルk平均法のクラスタ境界は存在せず，そのため停止条件である二つのクラスタが最適な分割となったときクラスタ中心の距離が最も離れるという仮定に合致しなかったと考えられる．従来手法のNMI値もかなり小さくなっていることから，従来手法が適切なクラスタリング結果を出力できないデータに関しては，提案手法の性能も悪くなると考えられる．

表2: 図4の人工データの実験結果

p	初期δ	δの更新回数			Algo. 2の繰返し数		未停止	NMI		
		最小	平均	最大	回数平均	1回平均		最小	平均	最大
0.5	$s/5$	21	30.68	71	2.27	23.28	0	0.070	0.188	0.195
0.5	$s/2$	3	3.2	4	5.623	2.98	0	0.189	0.193	0.194
0.1	$s/5$	2	47.4	69	2.23	31.26	0	0.001	0.178	0.181
0.1	$s/2$	5	8.52	9	2.46	7.94	0	0.181	0.181	0.181
-	-	-	-	-	-	-	-	**0.091**	**0.093**	**0.094**

三つ目のデータとして，図6に示すような2つのクラスタからなるデータに関する実験を行う．各クラスタのデータ数は，内側のクラスタと外側のクラスタともに200個で標準偏差は$s = 1.16$である．

図6の人工データにおいて，ガウス関数のパラメータを0 ∼ 2まで0.01間隔で変化させたときのガウス関数のパラメータとNMI値の関係を図7に示す．図7より図6のデータは$\Delta \neq \emptyset$であると考えられる．表3に図6の人工データの実験結果を示す．表3より，NMIは各pと初

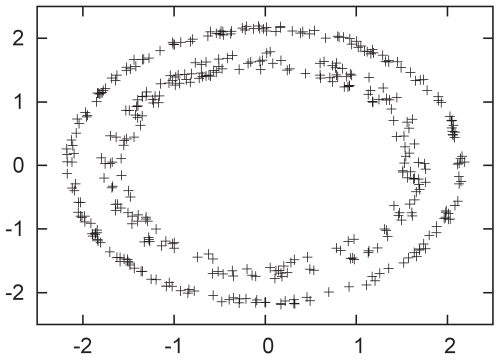

図6: リング状の2つのクラスタからなるデータ

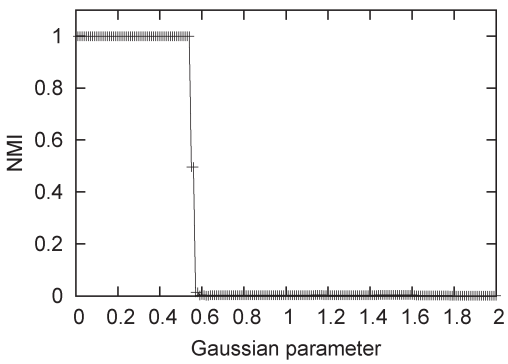

図7: 図4のデータのδとNMIの関係

期δの組合せに関して，ほとんど0となっている．図7より$\Delta \neq \emptyset$であると推定できるが，従来手法のNMI値もほぼ0となっていることより，初期クラスタへの依存度が大きいと考えられる．図7より，δの値が0.6となる近辺を境にして，NMIの値が急激に小さくなっていることから，初期クラスタが正解クラスタと僅かでも異なる場合は，クラスタリング結果は局所解になると推測できる．

表 3: 図 4 の人工データの実験結果

p	初期 δ	δ の更新回数			Algo. 2 の繰返し数		未停止	NMI		
		最小	平均	最大	回数平均	1 回平均		最小	平均	最大
0.5	$s/5$	4	78.44	100	1.744	69.38	28	0.000	0.011	0.217
0.5	$s/2$	3	64.98	100	1.768	63.6	17	0.000	0.002	0.008
0.1	$s/5$	12	93.92	100	1.423	84.84	45	0.000	0.017	0.111
0.1	$s/2$	92	99.84	100	1.133	98.06	49	0.000	0.001	0.006
-	-	-	-	-	-	-	-	**0.000**	**0.005**	**0.006**

4 つ目のデータとして，図 8 に示すような 4 クラスのデータに関する実験を行う．各クラスタのデータ数は，4 つともに 200 個で標準偏差は $s = 1.02$ である．上記と同様，図 8 の人工データにおいて，ガウス関数のパラメータを 0 ～ 2 まで 0.01 間隔で変化させたときのガウス関数のパラメータと NMI 値の関係を図 9 に示す．図 9 よりガウス関数のパラメータが 0.47 ～ 0.54 までの範囲のとき，真のクラスタと計算結果のクラスタが完全に一致する可能性がある．クラスタ数が増えた関係で図 2 のデータに比べて，NMI 値が 1 となることを期待できるパラメータの範囲が小さくなっていることが見て取れる．また，0 ～ 0.03 はは計算不能範囲である．

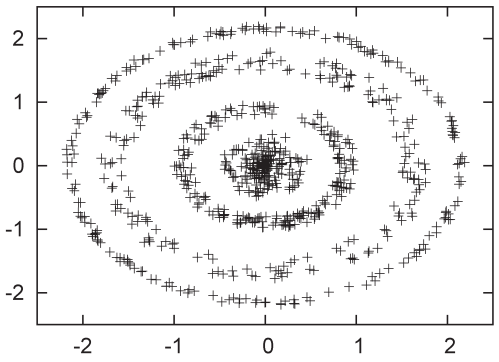

図 8: 4 つのクラスタからなるデータ

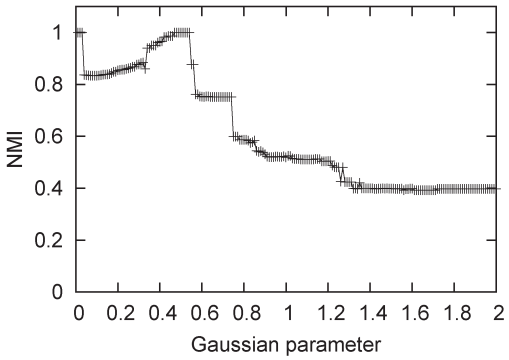

図 9: 図 8 のデータの δ と NMI の関係

表 4 に図 8 の人工データの実験結果を示す．表 4 より，4 つのクラスタからなるデータの場合は，NMI は各 p と初期 δ の組合せに関して，最小は 0.309 ～ 0.363 の間，平均は 0.517 ～ 0.546 の間，そして，最大は 0.644 ～ 0.765 の間となっていおり，正解クラスタと完全に一致することには失敗している．δ の初期値は，0.2 と 0.5 であるため，図 9 より，いずれの初期値も NMI= 1 で停止する機会はあるはずだが，NMI 値が小さくなった理由として，データの分布形状が図 2 と図 6 のデータを組み合わせた形状となっており，図 6 のデータの初期クラスタへの依存度が大きかったことから，外側の 3 つのクラスタの初期クラスタへの依存度が大きいためと考えられる．実際に，各試行のクラスタリグの結果を確認したところ，中心のクラスタ以外はほとんど正解の分類とはなっていなかった．$p = 0.1$ の場合に繰返しの最大値の条件により停止した回数が，$\delta_1 = 0.2$ の場合が 8 回，$\delta_1 = 0.5$ の場合が 1 回となっている．$p = 0.1$ で収束

速度が遅かったためと考えられる．$p=0.1$を試行した理由は，収束速度とクラスタリング精度に関するトレードオフを確認するためであるが，$p=0.5$の場合と$p=0.1$の場合のNMIを比較すると，$p=0.5$で計算してもそれ程精度は落ちていない．また，提案手法のNMIの最大値は初期δが$s/5$で$p=0.1$のとき以外は従来手法より高くなっていることから初期クラスタ次第では従来手法よりうまく分類できると考えられるが，NMIの最小値と平均は従来手法より劣っているため，初期クラスタ依存が高いデータに関しては安定性に欠けると考えられる．

表 4: 図 8 の人工データの実験結果

p	初期δ	δの更新回数			Algo. 2の繰返し数		未停止	NMI		
		最小	平均	最大	回数平均	1回平均		最小	平均	最大
0.5	$s/5$	4	23.72	57	5.246	18.4	0	0.360	0.532	0.765
0.5	$s/2$	2	5.54	30	9.269	6.94	0	0.361	0.517	0.714
0.1	$s/5$	4	54.6	100	3.453	42.6	8	0.309	0.527	0.644
0.1	$s/2$	2	12.4	100	6.359	12.18	1	0.363	0.546	0.743
-	-	-	-	-	-	-	-	**0.520**	**0.624**	**0.685**

最後のデータとして，図 10 に示すような3つのクラスタからなるデータに関する実験を行う．各クラスタのデータ数は，3つともに100個で標準偏差は$s=0.64$である．このデータの分布形状は凸状であるが，各クラスタの分散が異なるため，通常のk-平均法ではうまく分類できない．

図 10 の人工データにおいて，ガウス関数のパラメータを0～2まで0.01間隔で変化させたときのガウス関数のパラメータとNMI値の関係を図 11 に示す．図 11 より図 10 のデータは$\Delta\neq\emptyset$であると考えられる．

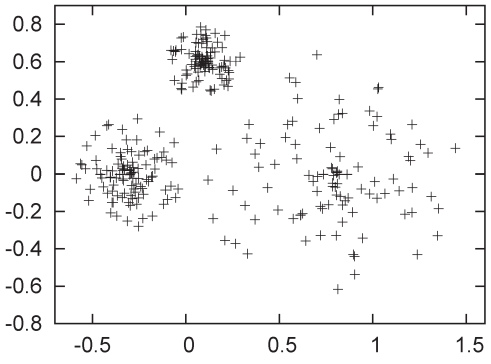

図 10: 凸状の3つのクラスタからなるデータ

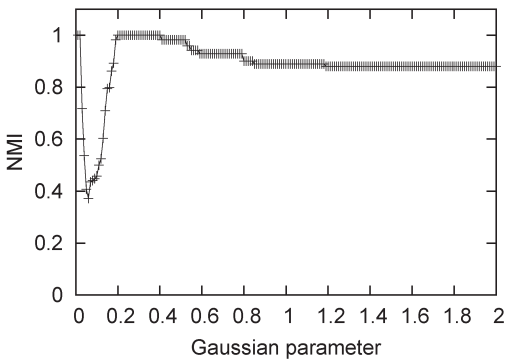

図 11: 図 10 のデータのδとNMIの関係

表 5 に図 10 の人工データの実験結果を示す．表 5 より，NMI は初期δを$s/2$としたときのNMIの平均はほぼ1に近く，最小は0.636ではあるものの高確率で正解クラスタを得られていると考えられ，従来手法にも勝っている．実際に詳細を確認したところ，pの値と初期δ

の値の組み合わせに関するNMI= 1の割合は，$(0.1, s/2)$のとき48/50，$(0.5, s/2)$のとき47/50，$(0.1, s/5)$のとき0/50，$(0.5, s/5)$のとき1/50となっていた．初期δの値が$s/5$のときは，0.939で停止することが多かったため式 (36) で強制的にd_6を採用する時期とr_{up}の検討が必要だと考えられる．

表 5: 図 10の人工データの実験結果

p	初期δ	δの更新回数			Algo. 2の繰返し数		未停止	NMI		
		最小	平均	最大	回数平均	1回平均		最小	平均	最大
0.5	$s/5$	3	15.06	34	3.172	9.8	0	0.590	0.894	1.000
0.5	$s/2$	2	2.06	5	2.848	1.04	0	0.636	0.979	1.000
0.1	$s/5$	4	20.78	40	2.975	14.06	0	0.655	0.884	0.939
0.1	$s/2$	2	2.08	6	3.093	1.06	0	0.636	0.986	1.000
-	-	-	-	-	-	-	-	**0.822**	**0.964**	**1.000**

6.2 実データ

実験はUCI machine learning repository[16] のデータベースを使用する．使用するデータは，iris, wine，satimage，pen digitの4つである．また，satimage と pen digit に関しては，元のデータのデータ数が非常に多いため，各クラスタからランダムに80個のデータを選択したものをデータとして使用した．各データの詳細を表 6に示す．表 6の各データにおけるNMIが最大となるδの区間は図 12 ∼ 15より推定した．

表 6: 実データの詳細

データ名	次元数	サンプル数	クラスタ数	標準偏差の平均	NMIが最大となるδの区間の推定
iris	4	150	3	0.97	0.8 ∼ 0.82
wine	13	178	3	5.11	265 ∼ 285
satimage	36	480	6	4.24	81 ∼ 83
pen digit	16	800	10	5.47	84 ∼ 86

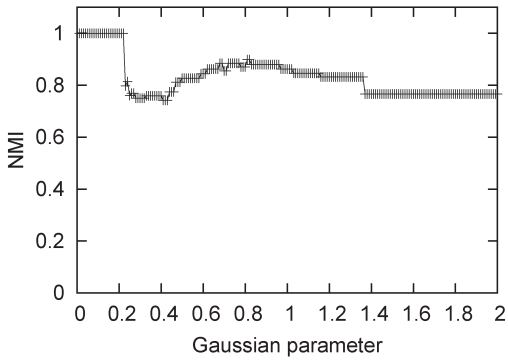

図 12: irisデータのδとNMIの関係

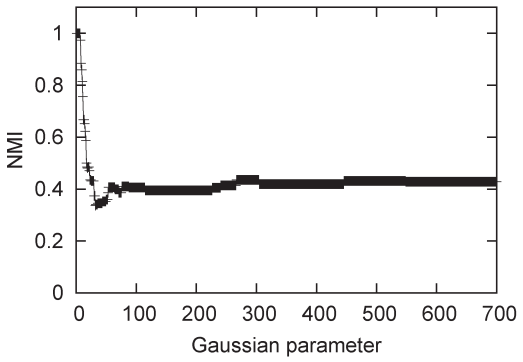

図 13: wineデータのδとNMIの関係

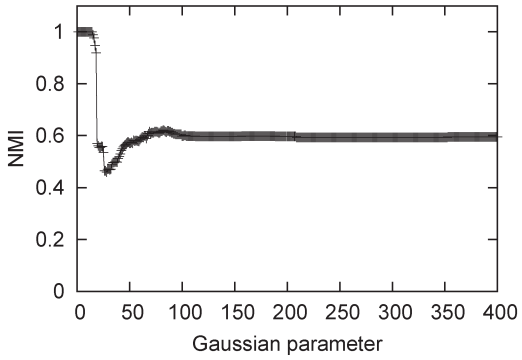

図 14: satimage データのδとNMIの関係

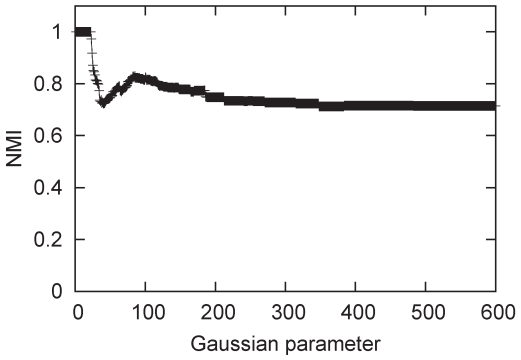

図 15: pen digit データのδとNMIの関係

図 13，図 14，図 15より，wine，satimage，pen digit に関しては，$s/2$と$s/5$がともに計算不能範囲であったため，初期パラメータδ_1を$5s$と$10s$とした．表 7に実験結果を示す．

表 7: 実データの実験結果

データ名	p	初期δ	δの更新回数			Algo. 2の繰返し数		未停止	NMI		
			最小	平均	最大	回数平均	1回平均		最小	平均	最大
iris	0.5	$s/5$	2	60.86	100	1.845	51.02	22	0.002	0.397	0.778
	0.5	$s/2$	2	6.98	58	3.910	4.16	0	0.408	0.648	0.786
	0.1	$s/5$	5	75.34	100	1.561	62.2	24	0.002	0.402	0.75
	0.1	$s/2$	2	16.26	100	2.809	13.62	2	0.388	0.601	0.784
	-	-	-	-	-	-	-	-	**0.778**	**0.778**	**0.778**
wine	0.5	$5s$	8	88.38	100	1.243	85.1	36	0.010	0.167	0.401
	0.5	$10s$	2	48.76	100	1.866	46.26	16	0.041	0.222	0.434
	0.1	$5s$	2	90.08	100	1.299	85.72	38	0.010	0.142	0.379
	0.1	$10s$	2	51.72	100	1.913	49.96	15	0.033	0.217	0.406
	-	-	-	-	-	-	-	-	**0.419**	**0.432**	**0.435**
sat image	0.5	$5s$	13	45.84	100	3.296	30.86	3	0.143	0.573	0.671
	0.5	$10s$	3	34.6	100	3.212	27.94	5	0.454	0.579	0.672
	0.1	$5s$	24	67.82	100	2.732	44.64	9	0.361	0.564	0.662
	0.1	$10s$	2	33.68	100	3.681	25.44	2	0.470	0.574	0.677
	-	-	-	-	-	-	-	-	**0.615**	**0.621**	**0.628**
pen digit	0.5	$5s$	12	40.52	100	5.094	25.86	2	0.348	0.711	0.777
	0.5	$10s$	5	25.94	100	4.884	20.62	1	0.683	0.733	0.799
	0.1	$5s$	7	56.68	100	4.407	38	6	0.380	0.707	0.788
	0.1	$10s$	9	39.38	100	3.813	28.18	4	0.705	0.745	0.784
	-	-	-	-	-	-	-	-	**0.765**	**0.790**	**0.809**

実データのNMI値は全て1以下であると推定されるため，Algorithm 2では，例え最適な初期クラスタとδを与えても，正解クラスタに完全に一致するクラスタを出力することは不可

能であると考えられる．表7より，全体的に繰返し回数の上限に達している割合が高い．初期δ_1の値が小さい場合に顕著であり，特にwineは初期δ_1が$5s$のときには，半分以上の試行が停止基準を満たさないまま計算を終了している．NMIの値に関しては，最大値は従来手法のNMI値にほぼ近い値となっているものの，平均のNMI値は従来手法と比較して全体的に小さくなっており，$\Delta = \emptyset$と推定されるデータにおいては，従来手法に適切なパラメータを設定したときの性能より，やや不安定になると考えられる．

7 考察

まず，停止条件に関して，実験では，$\Delta \neq \emptyset$，かつ，初期クラスタへの依存度が低いと思われるデータに関しては期待通りの結果が得られている．逆に，初期クラスタへの依存度が高いデータに関しては，図6の人工データの実験結果から，$\Delta \neq \emptyset$であってもAlgorithm 2によるカーネルk-平均法が局所解に収束する可能性が高くなるため，でたらめなデータ分類によるクラスタの中心間の距離を評価していることとなり意味をなさない．その意味で，Algorithm 2が高確率で最適解を出力することができるという条件が，Algorithm 3の性能を保証することになると考える．また，分離不可能な場合，性能はr_{low}とr_{up}に依存することとなる．r_{up}とr_{low}に関しては，事前にいくつかのデータに関して予備実験を行い，それぞれの値を固定値で決定したが，本来はクラスタの分布形状，クラスタ数，データ間の距離，クラスタ間の距離などの影響があると考えられる．実際に，wineにおいては，ほとんど停止条件が働いていない．表7におけるwineのNMIの結果は，ほとんど最大更新数である100に依存していると考えられる．$\Delta = \emptyset$のデータであれば，それなりに尤もらしい分類結果が得られればよく，多次元データで直接目視できない場合は，式(7)の目的関数を評価基準とするのが普通である．しかし，式(7)はδの変化に対する最適化を保証していないため比較による結果の採用ができず，停止基準が手法の性能を左右することになる．クラスタ間の距離や各クラスタの形状は事前に知ることができないが，データの分布状況，各データ間の距離，入力クラスタ数などの事前に知ることができる情報を停止基準であるr_{up}とr_{low}に反映させるような検討が必要である．

次に，更新幅に関して，q番目の繰返しにおける更新幅ϵ_qは式(34)において${\delta_q}^3$に比例した値で下からおさえられているため，$\delta_q \ll 1$の場合は相当小さな値となる．この場合，δ_qの更新に対して，クラスタの更新が行われない頻度が多くなる可能性が高く，収束に影響する．それを避けるため式(35)に示す$d_1 \sim d_6$の中で$d_i \leq p\delta_q$ $(i = 1, 2, \ldots, 6)$を満たす最大のd_iを選択するという条件を設定した．しかし，停止基準を満たしたときに複数個のデータに関してクラスタの所属変更が発生する場合は，一度に複数のデータのクラスタ所属の変更が行われることによりクラスタ間の距離が最大となる分類では停止せず，クラスタ間距離が減少方向に転じる場合が考えられる．対象データに関して$\Delta \neq \emptyset$であり，かつ，図2の人工データのように$\delta \in \Delta$がほぼ連続的で広範囲に分布する場合，このような危険は緩和されると考えられるが，逆に，図10のデータのように，$\Delta \neq \emptyset$ではあるものの，$\delta \in \Delta$の分布範囲が狭い場合は上記の事象が発生することも考えられる．仮に，二つのクラスタC_aとC_b間において，データ集合$X_a = \{x_{a1}, x_{a2}, \ldots, x_{al}\}$のクラスタ所属の変更が発生したとする．つまり，$C_a^{new} = C_a - X_a$，$C_b^{new} = C_b^{new} \cup X_a$となったとする．一般性を失わず$\| m_a - x_{a_1} \|^2 \leq \| m_a - x_{a_2} \|^2 \leq \cdots \leq \| m_a - x_{a_l} \|^2$とする．このとき，クラスタ中心の再計算

を行う前に，各$x \in X_a$に関して，一つずつ所属クラスタの変更を行い，変更毎に停止基準を検査する方法が考えられる．すなわち，$C_a^{new(i)} = C_a^{new(i-1)} - \{x_{ai}\}$，$C_b^{new(i)} = C_b^{new(i-1)} \cup \{x_{ai}\}$とし，$l$個の$\| m_a^{new(i)} - m_b^{new(i)} \|^2 / \| m_a^{new(i-1)} - m_b^{new(i-1)} \|^2$に関して逐次計算を行う．ただし，$1 \leq i \leq l$，$C_a^{new(0)} = C_a$，$C_b^{new(0)} = C_b$である．いずれかの$i$において停止基準を満たしていたなら，$i$個以下のデータの更新が行われるような$d_i$を選択するようにする．この改善の詳しい検証は今後の課題とする．式 (36) のd_iの決め方に関しては，式 (29) の厳密な解析ではなくパラメータpの値により更新幅を選択する戦略としたため，pの決め方が収束速度とクラスタの分類結果に大きく影響することになる．実験では，$p = 0.5$と$p = 0.1$に関して確認したが，NMI値の結果にそれ程差が出ていないことと，δの更新回数に関して$p = 0.5$が勝っていることから，$p = 0.5$とすることで問題ないと考えられる．停止条件を満たした時点で更新幅ϵ_qが最も小さくなるd_6を選択するように決めたことが，収束地点近辺で更新幅が大きくなりすぎることを抑制することに役立ったと考えられる．

最後に計算量に関しては，従来手法であるカーネルk-平均法は$\mathcal{O}(tkn^3d)$で提案手法は$\mathcal{O}(qtkn^3d)$となる．提案手法は従来手法のq倍のオーダではあるものの，連続量であるδに関して任意の一点を選択することは困難な作業であることを考慮すると，Algorithm 2が高確率で正確な分類をできるデータに関しては，概ねうまく分類できるδを上記の計算量で求めることができることは大きな欠点とはならない．δの初期値はある程度計算不能範囲とならないものを選択する必要があるものの，それらは，入力であるn個のデータからなる$n \times n$個の全ての対に関して，式 (5) が0.0を出力しなければよいという比較的計算し易い基準があるため，ある程度推定できると考えられる．

8 むすび

本論文では，カーネルk-平均法において，代表的なカーネル関数であるガウスカーネルのパラメータを逐次計算更新しながらカーネルk-平均法の計算を行い，適切なパラメータを探索するアルゴリズムを提案した．いくつかの人工データと実データを用いて，従来手法であるカーネルk-平均法との比較実験を行い，カーネルk-平均法が高い確率で最適解を出力できるデータ，すなわち，初期クラスタへの依存度が低いデータに関しては提案手法が有効であることを示した．また，カーネルk-平均法がうまく分類できないデータに関しては，カーネルk-平均法と同等の性能であることを確認した．今後の課題として，7章で考察した提案手法の改善，および，Yuら[11]の手法との比較がある．

参考文献

[1] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.

[2] K.R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf. An introduction to kernel-based learning algorithms. *Neural Networks, IEEE Transactions on*, Vol. 12, No. 2, pp. 181–201, 2001.

[3] N. Vapnik Vladimir. *Statistical learning theory*. Wiley-Interscience, 1998.

[4] B. Schölkopf, A. Smola, and K.R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, Vol. 10, No. 5, pp. 1299–1319, 1998.

[5] M. Girolami. Mercer kernel-based clustering in feature space. *Neural Networks, IEEE Transactions on*, Vol. 13, No. 3, pp. 780–784, 2002.

[6] A.Y. Ng, M.I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, Vol. 2, pp. 849–856, 2002.

[7] I.S. Dhillon, Y. Guan, and B. Kulis. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 551–556. ACM, 2004.

[8] J. Shi and J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 22, No. 8, pp. 888–905, 2000.

[9] C.H.Q. Ding, X. He, H. Zha, M. Gu, and H.D. Simon. A min-max cut algorithm for graph partitioning and data clustering. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pp. 107–114. IEEE, 2001.

[10] A. Ben-Hur, D. Horn, H.T. Siegelmann, and V. Vapnik. Support vector clustering. *The Journal of Machine Learning Research*, Vol. 2, pp. 125–137, 2002.

[11] S. Yu, L.C. Tranchevent, B. De Moor, and Y. Moreau. Optimized data fusion for kernel k-means clustering. *Kernel-based Data Fusion for Machine Learning*, pp. 89–107, 2011.

[12] 赤穂昭太郎, 津田宏治. サポートベクターマシン. 数理科学, pp. 52–58, 2000.

[13] L. Máté. *Hilbert space methods in science and engineering*. Hilger, 1989.

[14] L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, Vol. 2, No. 1, pp. 193–218, 1985.

[15] A. Strehl and J. Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, Vol. 3, pp. 583–617, 2003.

[16] A. Frank and A. Asuncion. UCI machine learning repository. http://archive.ics.uci.edu/ml, 2010.