1F3-01

局所尤度推定に基くノイジーデータからの大規模分散クラスタリング

Large-scale Distributed Clustering from Noisy Data using Localized Likelihood Estimation

佐久間　淳[*1]　　小林 重信[*1]
Jun Sakuma　　Shigenobu Kobayashi

[*1]東京工業大学 大学院 総合理工学研究科
Tokyo Institute of Technology, Interidisciplinary Graduate School of Science and Engineering

This paper presents a new clustering algorithm that can be easily extended for distributed computation. Our proposal algorithm, Distributed Clustering based on Mixture Modeling (DCMM), assumes that the noise is distributed following a uniform distribution and the each cluster is estimated as a component of a mixture distribution. Based on this assumption, the estimation is conducted by each component in a serial manner, that is, the computation can be conducted on multiple CPUs separately. The effectiveness of the proposal is verified by the numerical examples using artificially generated data sets.

1. はじめに

オンラインで入手可能なデータの増加に伴い, データが潜在的に有する知識を獲得するためのツールへの要求が高まっている. クラスタリングは教示情報を用いずに, 類似度の高い複数のデータを一つのクラスタに割り当てる方法であり, データセットの大域的な構造を俯瞰することで, データセットが非明示的に有する知識の獲得を容易にする. 大規模情報源からのスケーラブルなクラスタリングアルゴリズムは活発に研究されいているが [1], たとえデータ数 N について $O(N)$ の計算量が実現されているとしても N が極めて多い場合にはアルゴリズムの並列分散化が必須である.

また実問題で利用されるデータは通常どのクラスタにも属さないであろうノイジーデータを含むが、このようなデータセットを扱うためには, 確率分布に基づくクラスタリングアルゴリズムが有効であると考えられる. 確率分布に基づくアルゴリズムは理論的背景が明確であるが, 必ずしも現実のデータに則してアルゴリズムが設計されているとは言いがたく, 大規模データにおいて時間計算量および空間計算量が問題となる. たとえばクラスタ数の決定には正則化項の導入によるモデル選択が用いられるが, クラスタ数が数百や数千に達するような状況ではこれは計算量の面からも精度の面からも適切ではない.

本稿では EM 法を利用した混合分布推定に基いて, あるクラスタに属さないデータは一様分布に従うとの仮定の下で 1 つのクラスタを推定するアルゴリズムを提案する. このアルゴリズムは, (1) 並列実行が可能なため分散化が容易である, (2) データに多量のノイズが混入している場合でもその影響を受けにくい, (3) 終了条件が必ずしも明確ではないが, クラスタ数を前もって与える必要がない, 等の利点を持つ. 提案手法の有効性は人工的に生成したデータセットを用いた計算機実験によって検証される.

2. 混合分布と EM 法

クラスタリングの対象となるデータ集合を $X = \{x_1, ..., x_N\}$, N をデータ数とする. モデルパラメータ θ で特徴づけられた確率分布を $p(x;\theta)$ とする. このときデータ集合 X が, 混合分布

$$p(x|\theta) = \sum_{j=1}^{k} \alpha_j p_j(x;\theta_j) \quad (1)$$

に基づき発生したと考える. ここで, k は混合数 (クラスタ数), α_j は $\alpha_j > 0$, $\sum_j \alpha_j = 1$ なる混合係数である. X の $p(x;\theta)$ に対する尤度 ($= \prod_{i=1}^{N} p(x_i;\theta)$) を最大にするようなパラメータを θ_{ML} とする (最尤推定). このとき, 確率モデルに基づくクラスタリングでは, データ x について最大の確率密度 $p_j(x;\theta_{ML})$ を与える j をデータ x が属するクラスタと決定する. 混合分布の最尤推定には一般に EM 法 [3] を用いる.

[EM]

1. 初期パラメータ $\theta^{(0)}$ を設定, $t \leftarrow 0$.
2. $Q(\theta|\theta^{(t)}) = E\{\log p(X, Z|\theta)|X, \theta^{(t)}\}$ を計算 (E-step).
3. $\theta^{(t+1)} = \arg\max_\theta Q(\theta|\theta^{(t)})$ とする (M-step).
4. 収束条件を満たせば終了. そうでなければ $t \leftarrow t+1$ として 2 へ.

上記の EM 法は対数尤度の局所収束性を保証する.

3. 分散クラスタリング法の提案

3.1 混合分布推定に基づく分散クラスタリング

EM 法では E-step において $Q(\theta|\theta^{(t)})$ の計算に隠れ変数に関する条件付確率 $p(z_{ij}|x_i;\theta_j)$ が必要であるが, その計算に i と j に関する和計算が必要であるため, 少なくとも $O(kN)$ の計算量を要する. 多くのクラスタがオーバーラップするようにデータが分布する場合, これらの計算は避けがたいが, 各クラスタがある程度独立して分布している場合, $p(z_{ij}|x_i;\theta_j)$ はある j について 1 に近い値をとり, その他については 0 に近い値をとることから, N 個の全データに対する計算は実際には不要であることが多い. そこで本研究では確率モデルにおいて以下の性質を仮定したうえで条件付尤度の計算を分散化する.

Hard assignment: データ x_i が j 番目の分布から発生した確率を $p(z_{ij}|x_i)$ とする. この条件付確率について, 常に $p(z_{ij}|x_i) \in \{0,1\}$, $\sum_j p(z_{ij}|x_i) = 1$ が満たされる. これはあるデータ x_i が分布 p_j から発生する確率がつねに $\{0,1\}$ であると考えることに相当し, k-mean 法などと同等の制約である. 本論文ではこの性質を **Hard assignment** と呼ぶ.

この仮定に基づき 1 式の混合分布推定に基づく分散クラスタリング (Distributed Clustering based on Mixture Model, DCMM) のアルゴリズムの概要を **Algorithm 1** に示す.

DCMM は主に"LLEM", "混合比推定"および"クラスタの検定の 3 つのパートで構成されており, これらの詳細は次章で説明する. DCMM の特徴を以下に説明する.

連絡先: 佐久間　淳, 東京工業大学 大学院 総合理工学研究科, 神奈川県横浜市緑区長津田町 4259 G5-801, 045-924-5544, 045-924-5442, jun@fe.dis.titech.ac.jp

Algorithm 1 *DCMM*

1: X, α_0 を入力
2: $j \leftarrow 0$
3: **while** 収束条件が満たされない **do**
4: $\alpha_j \leftarrow \alpha_0$
5: **while** α_j, θ_j が収束しない **do**
6: α_j を固定し, $p_j(x;\theta_j)$ における θ_j を推定 (LLEM)
7: θ_j を固定し, α_j を推定 (混合比推定)
8: **end while**
9: $p_j(x;\theta_j)$ をクラスタとして受理するかどうかを検定
10: **if** 検定が真 **then**
11: $C_j \leftarrow X$ において密度 $p_j(x;\theta_j)$ が高い $\alpha_j N$ 個のデータ
12: $X \leftarrow X \setminus C_j$
13: $j \leftarrow j+1$
14: **end if**
15: **end while**
16: $C_j, \alpha_j, \theta_j (j = 0, ...)$ を出力

推定の分散化: Hard Assinment 仮定によって $Q(\theta|\theta^{(t)})$ の計算を各分布毎に局所化し, $\sum_j \alpha_j p_j(x;\theta_j)$ の推定を全分布の同時推定ではなく, 各分布毎に局所的に推定する (step. 5-8). あるクラスタの推定には必ずしも全データセットは不要であり, 興味のある領域に存在するデータセットについてアルゴリズムを実行すれば十分である. データセットを適宜分割することで, $O(Nk)$ の N を減少させることができる. また推定の分散化によって, $O(Nk)$ の k について, 分布毎に異なる CPU を割り当てることで $O(N)$ に減少させることができる.

クラスタ数の事前の決定が不要: DCMM ではクラスタ数のかわりに混合係数 α_0 を初期値として与える. 混合係数はそのクラスタが保持するデータ数の全データ数に対する割合である. 混合係数を固定した上でモデルで $p_j(x;\theta_j)$ を推定し (step 6), 推定パラメータ θ_j の下で α_j を推定する (step 7). Step 6 および 7 の詳細については 3.2 および 3.3 をそれぞれ参照されたい.

ノイズ対応のための密度分布に関する検定: EM 法は局所最適性しか保証しないため, Step 6 および Step 7 によって求められたクラスタが望ましい形状をしているとは限らない. そこで Step 9 ではクラスタの形状について, そのクラスタがノイズかどうか, および, 密度の形状が単峰であるかどうかをコルモゴロフ検定および二項検定により検定する.

本章では任意の分布形の混合分布 $\sum_j \alpha_j p_j$ に基づく分散クラスタリング法を説明し, 次章では分布 p_j が正規分布を取る場合の具体的なアルゴリズムについて説明する. 以下, Step 6,7 を中心に分散クラスタリング法のアルゴリズムの理論的バックグラウンドとその手順を説明する. 検定部分 (Step 9) の詳細については [4] を参照されたい.

3.2 Localized Liklelihood EM

E-step で出現する Q 関数は混合分布の場合以下のように書き下せる.

$$Q(\theta|\theta^{(t)}) = \sum_Z p(Z|X;\theta^{(t)}) \log p(X,Z;\theta) \quad (2)$$

$$= \sum_i p(z_{ij}|x_i;\theta_j^{(t)}) \log \sum_j p_j(x_j, z_{ij};\theta_j) \quad (3)$$

この計算に i と j の和計算が必要であるため, EM アルゴリズムは少なくとも $O(kN)$ の計算量を要することに注意されたい. ここで隠れ変数の条件付確率について, 常に $p(z_{ij}|x_i) = \{0,1\}$, $\sum_j p(z_{ij}|x_i) = 1$ が満たされる (Hard assignment 仮定) ならば, Q 関数は

$$Q(\theta|\theta^{(t)}) = \sum_j \sum_{x_i \in D_j} \log p_j(x_i, z_{ij};\theta_j) \quad (4)$$

$$= \Sigma_j Q_j(\theta_j|\theta_j^{(t)}). \quad (5)$$

Algorithm 2 *LLEM*

1: X を入力
2: $t \rightarrow 0$
3: $\theta_j^{(0)}$ および $\tilde{P}(Z|\Phi)^{(0)}$ を初期化
4: **while** $\theta_j^{(t)}$ が収束 **do**
5: $K(\tilde{P}(Z|\Phi)^{(t+1)}, \theta_j^{(t)}) > K(\tilde{P}(Z|\Phi)^{(t)}, \theta_j^{(t)})$ となるように $\tilde{P}(Z|\Phi)^{(t+1)}$ を決定 (E-step)
6: $K(\tilde{P}(Z|\Phi)^{(t+1)}, \theta_j^{(t+1)}) \geq K(\tilde{P}(Z|\Phi)^{(t+1)}, \theta_j^{(t)})$ となるように, $\theta_j^{(t+1)}$ を決定 (M-step)
7: $t \rightarrow t+1$
8: **end while**
9: $\theta_j^{(t)}$ および $K(\tilde{P}(Z|\Phi)^{(t)}, \theta_j^{(t)}) = 1$ なるデータを出力

のように, 各分布毎に和の形式で書くことができる. ただし, $I(i)$ はデータ x_i に対して $z_{ij} = 1$ なるインデックス j を返す関数, $D_j = \{x_i | I(i) = j\}$ である. これを**条件付対数局所尤度** (Conditional Log Localized Likelihood) と呼ぶ.

条件付対数局所尤度がコンポーネント毎に個別に計算可能である性質を用いて, 混合分布推定のための局所尤度 EM(Localized Liklihood EM, LLEM) を構築する [5]. 分布は $p_0, ..., p_{j-1}$ の順番に推定されるものとし, コンポーネント p_j を推定する LLEM のアルゴリズムを説明する. ここでは分布 p_j に hard assignment されるデータ数を $\alpha_j N = c_j$ 個と固定されるものとする. また, $\tilde{P}(Z|\Phi)$ を hard assignment 仮定を満たす関数とする.

$K(\tilde{P}(Z|\Phi), \theta_j) = \sum_Z \tilde{P}(Z|\Phi) \log p_j(X;|\theta_j)$ とすると, コンポーネント $p_j(x;|\theta_j)$ における LLEM のアルゴリズムは **Algorithm 2** のように示される.

上記の E-step, M-step の両方において, 具体的な $\tilde{P}(Z|\Phi)^{(t+1)}$ および $\theta_j^{(t+1)}$ の求め方は示されていない. 以下に, 両 step におけるその実現方法を考察する.

E-step では K の最大化は困難であるが, 増大化であれば簡単な局所探索等のヒューリスティックによって以下のように実現可能である.

1. 固定された $\theta_j^{(t)}$ において, 下式を満たすような部分集合 $D_j^{(t+1)}$ を生成

$$\sum_{x \in D_j^{(t+1)}} \log P_j(X, Z;\theta_j^{(t)}) \geq \sum_{x \in D_j^{(t)}} \log P_j(X, Z;\theta_j^{(t)})$$

2. $\tilde{P}(Z|\Phi)^{(t+1)}$ を以下のように設定:

$$\tilde{P}(z_{ij}|\Phi)^{(t+1)} = 1 \quad \text{if} \quad x_i \in D_j^{(t+1)}$$
$$\tilde{P}(z_{ij}|\Phi)^{(t+1)} = 0 \quad \text{else}$$

また M-step における $K(\tilde{P}(Z|\Phi)^{(t)}|\theta_j^{(t)})$ の増大化法は, 確率モデルおよび事前分布の具体的な形状に依存しているが, データセット $D_j^{(t)}$ についての最尤推定で最大化することができる. 上記, E-step, M-step に基づいて, LLEM 法は混合モデルのなかの一つのモデルに関する局所尤度について, 局所収束 (局所最大化) が保証される. 証明は [5] を参照されたい.

3.3 混合比推定

上述した LLEM は局所尤度に関する局所収束性を保証するが, α_j が固定値として構成されている. この値を推定するために, あるモデル p_j に着目したときに, p_j 以外のモデルについては $P_U(\ ;u)$ と同一視して,

$$p(x;\theta) = \alpha_j p_j(x;\theta_j) + (1-\alpha_j) p_U(x;u), \quad (6)$$

なる確率モデルを考えることにする。ここで，p_U はデータの領域全体にわたる一様分布，u は一様分布におけるモデルパラメータである．p_j についての LLEM が実行されていると考えれば，θ_j はすでに推定済みである．u を固定した場合，尤度は極値を持つならば，σ に関する一次元最適化によって望ましい σ を得ることができる．たとえば u を正規分布の 2σ 線[*1]を，クラスタの境界と考え，その外は一様分布であるとみなせば発散せず安定した推定が可能でる．逆に，u を決めたうえで σ を探索し，なお σ が発散するならば，それは正規分布をなさない，と結論することができる．適切な σ が決定されれば $p_j(x;\theta_j) > u$ となるようなデータを収集し，これに基づいて α_j を決定することができる．

4. 混合正規分布における DCMM

前章では任意の分布形 p_j における DCMM の構成法を説明した．本節では最も基本的な多次元正規混合分布における DCMM の具体的なアルゴリズムについて説明する．

多次元正規分布はモデルパラメータ平均値ベクトル μ および分散共分散行列 Σ によって特徴つけられる．d 次元正規分布は下式のように記述される．

$$P_i(x;\theta) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp \quad -\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)$$

固有ベクトル $e_1,...,e_d$ について行列 $E = \{\ _1,...,\ _k\}$ とし，Σ の固有値 $a_1,...,a_d$ について対角行列 $A = \mathrm{diag}(a_1,...,a_d)$ とすると，$\Sigma = E^TAE$ となる．

前章同様，一様分布を $P_U = (x;u)$ とすると，複数の正規分布と一様分布の混合分布は

$$P_i(x;\theta) = \sum_{j=1}^{k} \alpha_j P(x;\mu_j,\Sigma_j) + \alpha_U P_U(x;u) \qquad (7)$$

となる．ただし $\sum_j \alpha_j + \alpha_U = 1$ である．本章ではデータ X が与えられたときに $\alpha_j, \mu_j, \Sigma_j$ を推定するアルゴリズムを考える．

混合正規分布における LLEM: LLEM では，混合比を固定し局所尤度を極大化するアルゴリズムである．入力の α_0 は対象とするモデルに対する混合比の初期値である． s' は $s' > \alpha_j N$ なるパラメータであり，E-step における D_j の探索において，hard assignment するデータの候補数である．実験的には $s' = 4\alpha_0 N$ 程度で十分である．

混合正規分布における混合比推定: クラスタの平均ベクトル μ および固有ベクトル $e_1,...,e_d$ を固定し，固有値 $a_1,...,a_k$ を変化させ混合比 α_j を推定する．固有値を LLEM における推定値の $r_{min}(<1)$ 倍から $r_{max}(>1)$ について局所尤度 L を 1 次元探索し，最大となる r^* を探索する．

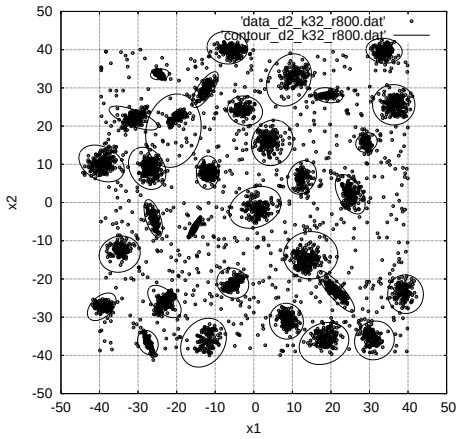
上記のアルゴリズムを組み合わせることで，正規混合分布における DCMM が構成される．クラスタ数 $k = 16$， 混入ノイズ数 $r = 800$，データ次元数 $d = 2$ において生成したデータセットに DCMM を適用した際の推定の様子を図 1 に示す．

5. 実験

7 式においてクラスタ数 k，ノイズ混入数 r およびデータ次元数 d を変化させて生成したデータについて実験を行う．1 クラスタあたりのデータ数は $[100, 200]$ における一様乱数である．

*1 正規分布から生成されたデータの 95%が 2σ 線内に存在する

図 1: クラスタ数 $k = 32$， 混入ノイズ数 $r = 800$, データ次元数 $d = 2$ において生成したデータセットに DCMM を適用した際の推定の様子



Algorithm 3 混合正規分布における *LLEM*

1: X, α_0, s' を入力
2: μ, Σ を初期化
3: $t \leftarrow 0, s \leftarrow \alpha_0 N$
4: **while** 局所尤度が未収束 **do**
5: $x_i \in X$ を $P(x;\mu,\Sigma)$ における密度に基づきソートし，上位 s' 個を選択 ($\mathrm{List}_{s'}$)
6: $\mathrm{List}_{s'}$ から s 個の要素を密度に基づきルーレット選択し，これを List_s とする
7: $L \leftarrow \sum_{x_i \in List_s} P(x_i;\mu,\Sigma)$(局所尤度)
8: **if** $L > L_{best}$ **then**
9: List_s から $\mu^{(t)}, \Sigma^{(t)}$ を最尤推定
10: $\mu \leftarrow \mu^{(t)}, \Sigma \leftarrow \Sigma^{(t)}$
11: $L \leftarrow L_{best}$
12: **else**
13: Step 6 へ戻る
14: **end if**
15: $t \leftarrow t + 1$
16: **end while**
17: $\mu, \Sigma, List_s$ を出力

評価基準として以下に定義する accuracy と error を用いる．データにはその発生元となった分布があらわすラベルがついている (アルゴリズムにはこの情報は不可視である)．クラスタリングの結果，そのクラスタにおいて最も多くのデータが持つラベルを majority と呼ぶ．

- True Positive(TP) : あるクラスタに属するデータの中の，majority のデータ数
- False Positive(FP) : あるクラスタに属するデータの中の，非 majority のデータ数
- True Negative(TN) : あるラベルを有するデータの中の，非 majority のデータ数

このとき以下に定義する accuracy および error を評価基準とする．

- accuracy(acc) $= \frac{TP}{TP+TN}$: あるラベルが，単一のクラスタに属している率
- error(err) $= \frac{FP}{TP+FP}$: あるクラスタについて，majority ではないデータの含有率

accuracy は高いほど精度の高いクラスタを生成していることを表し，error は低いほど誤ったデータをクラスタのメンバとしてはいないことをあらわす．これに加えて，発見したクラ

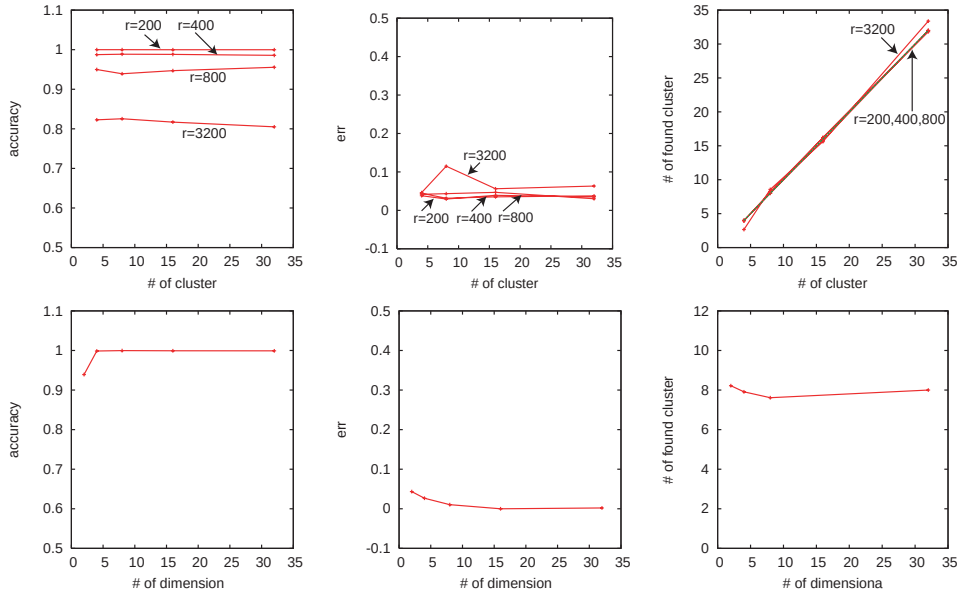


図 2: 上段は左から，クラスタ数 vs accuracy, error, 見つかったクラス数．下段は左から，次元数 vs accuracy, error, 発見クラスタ数．データはすべて 10 試行平均値．

Algorithm 4 *MixtureRatioEstimation*

1: $X, List_s$, , $^i(i = 1, ..., d), r_{min}, r_{max}, r_{step}$ を入力
2: $r_0 \leftarrow r_{min}, t \leftarrow 0$
3: **while** $r_t < r_{max}$ **do**
4: $\Sigma_t \leftarrow E^T(r_t I)E$
5: $x_i \in X$ を $P_j($; ,$\Sigma)$ における密度に基づきソートし，$P_j($; ,$\Sigma) > d$ なる x_i を List$_t$ に入れる
6: $L_t \leftarrow \sum_{x_i \in List_t} P($ $_i$; ,$\Sigma)$(局所尤度)
7: $r_{t+1} \leftarrow r_t + r_{step}, t \leftarrow t + 1$
8: **end while**
9: $t^* \leftarrow$ L_t を最大にする t
10: $r^* \leftarrow$ L_t を最大にする r_t
11: $List_t^*$ と $\Sigma^*(= E^T(r^* I)E)$ を出力

スタ数を評価基準に用いる．提案手法はクラスタ数を前もって与えないため, 発見するクラスタ数は試行によって異なる．これは真のモデルのクラスタ数に近いほど望ましい．

ロバスト性: 推定のロバスト性を見るために，クラスタ数 k およびノイズ混入数 r を変化させて実験を行う．この実験では $k = \{4, 8, 16, 32\}$, $r = \{0, 200, 800, 3200\}$ としてデータ次元数を $d = 2$ とした．評価基準として accuracy, error および発見クラスタ数を用いる．図 2(上) は異なるノイズ混入数について, 真のクラスタ数に対する accuracy, error, 発見クラスタ数をプロットしたものである．ノイズは一様にデータに混入するため, r が大きいほど accuracy は悪化する．ただし提案手法では確率モデルが正確に推定されているため, error および発見クラスタ数はノイズの影響を受けていないことがわかる．またデータの真のクラスタ数が増えても, これらの評価基準においては性能が悪化しないことも確認できる．

スケーラビリティ: アルゴリズムのスケーラビリティを見るために，ノイズ混入数 $r = 800$, クラスタ数 $k = 8$ として, 次元数 d を $d = \{2, 4, 8, 16, 32\}$ と変化させて実験を行った (図 2(下))．評価基準は前節と同様である．次元数についても, 精度・発見クラスタ数ともに正確な推定を行っており, 高次元データにおいても性能の悪化がないことを確認できた．

6. おわりに

本稿ではノイズモデルとして一様分布を想定した確率モデルにおいて, 混合分布に基づくクラスタリングアルゴリズム DCMM を提案した．DCMM はノイズに対するロバスト性が強く, また次元数に対する性能のスケーラビリティにも優れていることを実験により明らかにした．今後はオンメモリでの処理が困難なギガデータに対してグリッド計算機を利用したクラスタリングを試みる予定である．また XML などの半構造化データや文書データ, 木構造で記述されたルール, 関係グラフ, などのさまざまな非構造化データにおけるクラスタリングに拡張する予定である．

参考文献

[1] Domingos, P. and Hulten, G.: A General Method for Scaling Up Machne Learning Algorithms and its Application to Clustering, *Proc. of the 18th Int'l Conf. on Machine Learning*, pp. 106-113 (2001)

[2] Kollios, G., Gunopulos, D, Koudas, N., Berchtold, S.: Efficient Biased Sampling for Approximate Clustering and Outlier Detection in Large Datasets, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15, No. 5, (2003).

[3] Depmster, A. P., Laird, N.M. and Rubin, D. B. (1977), Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Statist. Soc. B*, 39:1-38 (1977).

[4] 佐久間, 小林, 局所尤度推定に基づくノイジーデータからの分散クラスタリング, 第 35 回システム工学部会研究会, (2005).

[5] 佐久間, 小林, 依存性を有する事前分布における Localized MAP-EM, 情報論的学習理論ワークショップ, pp. 289-294 (2003).