Vol.2011-MPS-86 No.34
Vol.2011-BIO-27 No.34
2011/12/2

データ距離と反復法による確率的クラスタリング

北口　景子[†1]

本研究では，EM アルゴリズムの期待値段階におけるデータの各クラスタへの帰属確率を，ある種のデータ距離に基づいて与え，EM アルゴリズムと類似の反復を行うことにより確率的クラスタリングを行う手法に関して考察する．

Stochastic Clustering by a Distance and the Iterative Method

Keiko Kitaguchi[†1]

In the expectation step of EM algorithm,the membership level to each cluster is indicated by a probability induced from the likelihood.In this study,we use a certain distance instead of the likelihood and repeat some steps similar to those used in EM algorithm.

1. はじめに

混合分布問題においては EM アルゴリズムの E ステップは，尤度の比によってデータの各クラスタへの帰属確率を推定することに対応すると考えられる．本研究では，この帰属確率を尤度の比の代わりにマハラノビスの距離に基づいて与え，最適なパラメータの推定を EM アルゴリズムと同様の手順で反復をすることにより，確率的クラスタリングを行う手法を提案する．本研究では，帰属確率を制御するパラメータ λ を導入し，その値による結果の変化を数値実験で考察する．

†1 お茶の水女子大学大学院 人間文化創成科学科
Graduate school of Humanities and Sciences, Ochanomizu University

2. 混合分布問題

K 個のクラスタからなる混合分布 $f(x \mid \theta)$ は，k 番目のクラスタの確率密度関数を $f_j(x \mid \theta_j)$，混合比を $p_j(j = 1, \ldots, K)$ とするとき，

$$f(x \mid \theta) = \sum_{j=1}^{K} p_j f_j(x \mid \theta_j, p_j) \tag{2.1}$$

で表される．$\sum_{j=1}^{K} p_j = 1$ であり，θ_j は各分布を特徴付けるパラメータである．各分布が二次元正規分布 $N_2(\mu_j, A_j)$ に従うならば，各 $f_j(x \mid \theta_j)$ は

$$f_j(x \mid \theta) = \frac{1}{2\pi \mid A_j \mid^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x-\mu_j)^T {A_j}^{-1}(x-\mu_j)\} \tag{2.2}$$

である．ここで，μ_j，A_j は，それぞれの平均ベクトル，分散共分散行列である．混合分布問題では，$f(x \mid \theta)$ からの標本 $\{x_1, \ldots, x_N\}$ が与えられたとき，
パラメータ $\theta = \{\theta_1, \ldots, \theta_K, p_1, \ldots, p_K\}$ および最適なクラスタ数 K を推定する．

3. 混合分布問題と EM アルゴリズム

正解クラスの情報が与えられずに，観測データ x だけが与えられている教師なしの学習は，正解クラスを隠れ変数とする混合分布の推定と考えることができる．EM アルゴリズムでは，各クラスタの平均 μ_j，分散共分散行列 A_j に適当な初期値を与え，E ステップ，M ステップと呼ばれる 2 つのステップ行う．

(1)　E ステップ（expectation step）
データ x_i がクラスタ C_j に属する期待値（確率）を

$$z_{ij} = \frac{p_j f_j(x_i)}{\sum_{k=1}^{K} p_k f_k(x_i)} \tag{3.1}$$

と推定する．ただし f_k はクラスタ k の密度関数である．

(2)　M ステップ（maximization step）
対数尤度 $Q(\theta_1, \ldots, \theta_K, p_1, \ldots, p_K)$ が最大となるようにパラメータを推定する．正規混合分布の場合，各パラメータの更新式はそれぞれ，

$$p_j = \frac{1}{N}\sum_{i=1}^{N} z_{ij} \quad , \quad \mu_j = \frac{\sum_{i=1}^{N} z_{ij} x_i}{\sum_{i=1}^{N} z_{ij}} \quad , \quad A_j = \frac{\sum_{i=1}^{N} z_{ij}(x_i-\mu_j)(x_i-\mu_j)^T}{\sum_{i=1}^{N} z_{ij}} \tag{3.2}$$

Vol.2011-MPS-86 No.34
Vol.2011-BIO-27 No.34
2011/12/2

で考えられる．ここで，対数尤度 $Q(\theta)$ とは，

$$Q(\theta) = \sum_{i=1}^{N} \log f(x_i \mid \theta) \tag{3.3}$$

で表わされる θ の関数である．E ステップと M ステップを十分な回数繰り返すことで，クラスタ数が K 個のときの最適なパラメータ $p_j, \mu_j, A_j (j = 1, \ldots, K)$ を求める．

4. マハラノビス距離と判別分析

各データ x_i とクラスタ C_j の中心との間のマハラノビス二乗距離 d_{ij}^2 は，クラスタ平均 μ_j，分散共分散行列 A_j を用いて，以下の式で与えられる．

$$d_{ij}^2 = (x_i - \mu_j)^T A^{-1} (x_i - \mu_j) \tag{4.1}$$

データ群が正規分布に従うとき，マハラノビス二乗距離が自由度 2 のカイ二乗分布に従うことから，本研究では，データ x_i の各群への帰属確率を

$$z_{ij} = \frac{1}{Z_i} \exp\Big(-\frac{1}{2\lambda} d_{ij}^2\Big) \tag{4.2}$$

と与えることにする．ただし，

$$Z_i = \sum_{k=1}^{j} \exp\Big(-\frac{1}{2\lambda} d_{ik}^2\Big) \tag{4.3}$$

である．変数 λ は帰属確率を制御するパラメータとなり，値が大きいほど等確率になる．EM アルゴリズムにおける尤度の比にも上と同様な制御変数 λ を導入することが可能である．このとき，λ の値が小さくなるほど帰属確率は 0 又は 1 の 2 値に近づき，K-means 法と同じ動きをする．

確率的クラスタリングでは，z_{ij} はユーザーが予め選んだ確率 p 以上のときのみ x_i はクラスタ C_j に属すると考え，それ以外のデータはどちらにも属さないものと考える．

5. 実　験　例

5.1 提 案 手 法

図 1 のような, 二次正規分布に従う 2 群データについて，以下のアルゴリズムでクラスタリングを行う．1 群が 600 点, 2 群が 400 点の, 合計 1000 点のデータとなっている．

(1)　x_i の各クラスタへの帰属確率をランダムに与え，μ_j，A_j の値を定める．

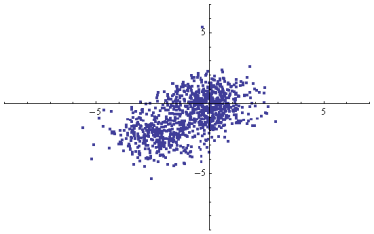

図 1　サンプルデータ

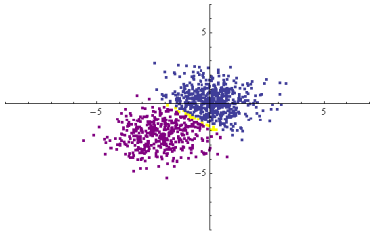

図 2　マハラノビス距離による結果

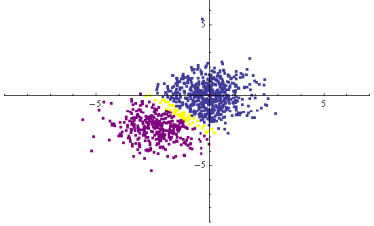

図 3　EM アルゴリズムによる結果

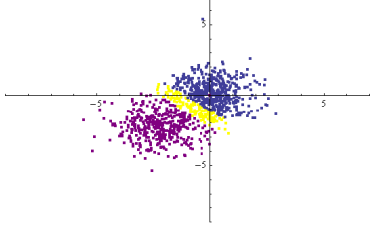

図 4　λ を導入した EM アルゴリズムによる結果

(2)　式 (4.2) により，各データ x_i のクラスタ C_j への帰属確率を求める．

(3)　式 (3.2) よりパラメータ μ_j，A_j の値を更新する．

(4)　パラメータ μ_j，A_j の値が収束するまで，(2)，(3) を十分な回数繰り返す

本実験では，帰属確率が 0.7 以上のときそのクラスタに属するものとした．

5.2 結　　果

マハラノビス距離による判別では $\lambda = 0.5$ のとき図 2 のようになった．λ の値をより大きくし，あいまい性を有すクラスタリングを得ようとしたが，最適な分布の推定ができなかった．一方，尤度の比に λ を導入し判別を行うと，図 4 のようなあいまい性を有したクラスタリングを行うことができた．

6. ま　と　め

マハラノビス距離による判別では，帰属確率は極めて 0 と 1 に近い値となり，λ の値を変えてもあいまいな結果は得られなかった．今後の課題としては，他の距離による判別や，形の異なるデータによる実験を考察したい．