社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－MPS－62 (21)
2006－BIO－7 (21)
2006／12／22

ファジィK－平均法の論文分類支援への応用

大川 しおり, 石榑 彩乃, 澤村 めぐみ
お茶の水女子大学大学院人間文化研究科

本研究では，教師信号を持たない非階層的分類手法であるファジィK－平均法についてのある種の改良手法と実データのクラスタリングへの応用例を提示する．ファジィK－平均法ではクラスタへの帰属度を距離による確率で表す．データ間の距離としてマハラノビス距離を採用することにより，混合分布問題のパラメータを推定を行うことも可能になる．しかし，データ数の増大と共に局所解に陥りやすいため，ファジィ度を制御するパラメータλを導入し，アニーリングを行うことで改善を試みる．また，混合分布問題においてよく用いられるEMアルゴリズムとの比較を行い，その有効性について検証する．実データへの応用としては，論文のアブストラクトの専門用語から特徴つけられた多次元ベクトルをPCAにより2次元に縮約したデータを用い，クラスタリングを施すことにより，興味の対象であるクラスタの抽出が可能であるかを調べる．

An application to a paper classification support by Fuzzy K−means

Shiori Okawa, Ayano Ishigure, and Megumi Sawamura
Ochanomizu University
Graduate School of Humanities and Sciences

In this study, we shall show one of improvement methods of the fuzzy K−means, which is non-supervised and non-hierarchical classification method, and give an application to a clustering of the multi-dimensional data of the abstracts of papers. In the fuzzy K−means, the membership level to each cluster is indicated by a probability induced from the distance. It is applicable to an estimation problem of the parameters for a mixture distribution of Gaussian laws by adopting the Mahalanobis distance. However, an increasing data brings local minimum problem, we shall try to improve it by introducing an annealing parameter λ which will controls fuzzy level.

1 はじめに

正解ラベルのついていない観測データを，その分布に従っていくつかのクラスタに分類する手法をクラスタリングと呼ぶ．正解ラベル，すなわち教師信号がないことから「教師なしの学習」,「自己組織化」などとも呼ばれる．クラスタリング手法は階層的クラスタリングと非階層的クラスタリングの2つに大きく類別される．非階層的手法の代表的手法にK-平均法やファジィK-平均法がある．K-平均法では要素がクラスターに属するか属さないか，つまり帰属度を0か1の2値的に表すの対し，ファジィK-平均法では帰属度を0から1の確率で表すことによりクラスタへの帰属にあいまいさを持たせることができる．帰属性の尺度，すなわち確率は，各要素のクラスタ中心からの距離から定めるようにする．距離としては，ユークリッド距離やマハラノビスの距離あるいは，もっと一般にL^p距離など様々な距離が考えられる．(ユークリッド距離やマハラノビスの距離はL^2クラスの距離と考えられる) この距離を分類対象により適切なものを採用することも重要である．

一方，混合分布問題の隠れたパラメータ (各分布のパラメータと混合比率) を最尤推定するEMアルゴリズムも確率的クラスタリングの代表的な手法のひとつである．そのアルゴリズムは，パラメータを

条件付き期待値で補充するEステップとそのパラメータに基づき尤度を最大化するMステップを繰り返すことで最適なパラメータに更新していく逐次的アルゴリズムである.

一般に,クラスタリングを行う際は,クラスタ数Kを既知としてK個のクラスタに分割する.しかしながら,現実的な問題ではKが未知である場合が多く,最適なクラスタ数を推定するのも大きな問題となってくる.

そこで,本研究ではKを未知のものとし,ファジィK-平均法とAICを組み合わせて最適なクラスタ数Kとその時のクラスタ中心を推定する手法を提案する.

また,実データへの応用としては,論文のアブストラクトから専門用語により特徴つけられた多次元ベクトルをPCAにより2次元に縮約したデータを用い,クラスタタリングを施すことにより,興味の対象であるクラスタの抽出が可能であるかを調べた.

2 混合分布問題と確率的クラスタリング

2.1 混合分布問題とEMアルゴリズム

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

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

で表される.もちろん$\sum_{j=1}^{K}p_j=1$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2次元正規分布$N_2(\mu_j,A_j)$に従うならば,各$f_j(x|\theta_j)$は

$$f_j(x|\theta_j)=\frac{1}{2\pi|A_j|^{\frac{1}{2}}}\times \exp\{-\frac{1}{2}(x-\mu_j)^T A_j^{-1}(x-\mu_j)\} \tag{2}$$

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.

混合分布問題とは,$f(x|\theta)$からの標本$\{x_1,\cdots,x_N\}$が与えられたとき,パラメータ$\theta=\{\theta_1,\cdots,\theta_K,p_1,\cdots,\theta_K\}$および最適なクラスタ数$K$を推定することである.

混合分布問題では,データx_iのj番目のクラスタC_jへの帰属を示す変数z_{ij}を導入する.(真の状態はz_{ij}は0または1であるはずだが,観測データはこの情報が欠損したものであり,したがって,この変数を期待値(確率)で補うことになる)

EMアルゴリズムにおけるパラメータ推定

まず,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,その後,以下のEステップ,Mステップを必要回数繰り返す.
(各iについて,いずれかひとつのjに対してのみ$z_{ij}=1$としてMステップ,Eステップを繰り返してもよい.)

(Eステップ)

データx_iがクラスタC_jに属する期待値(確率)z_{ij}を

$$z_{ij}=\frac{p_j f_j(x_i)}{\sum_{k=1}^{K}p_k f_k(x_i)} \tag{3}$$

と推定する.

(Mステップ)

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,

$$\begin{aligned} p_j&=\frac{1}{N}\sum_{i=1}^{N}z_{ij}\\ \mu_j&=\frac{\sum_{i=1}^{N}z_{ij}x_i}{\sum_{i=1}^{N}z_{ij}}\\ A_j&=\frac{\sum_{i=1}^{N}z_{ij}(x_i-u_j)(x_i-u_j)^T}{\sum_{i=1}^{N}z_{ij}} \end{aligned} \tag{4}$$

ここで,対数尤度$Q(\theta)$とは,

$$Q(\theta)=\sum_{i=1}^{N}\log f(x_i|\theta) \tag{5}$$

で表される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 個の時の最適なパラメータ p_j, μ_j, A_j $(j = 1, \ldots, K)$ が求まる．

さらに最適な K を求めるためのモデル選択基準として AIC を用いる．AIC は 最大対数尤度 Q^* から推定した自由パラメータ数 M を用いて

$$\mathrm{AIC} = -2(Q^* - M) \tag{6}$$

で与えられ，AIC が小さい程，モデルの当てはまりが良いと判断される．

しかし，EM アルゴリズムはデータ数が多いときには，尤度の極大値に収束してしまう可能性もあるので，注意が必要となってくる．

2.2 ファジィ K-平均法

ファジィ K-平均法では，要素 x_i がクラスタ C_j に帰属する確率を各要素のクラスタ中心からの距離から定める．本研究では，クラスタ平均 μ_j の他にクラスタの大きさを分散共分散行列 A_j を用いて，ある一定のマハラノビスの距離以下にある領域として表現する手法を採る．

ファジィ $K-$ 平均法　あらかじめ帰属度を制御するパラメータ $\lambda > 0$ およびクラスタ数 K を固定する．各クラスタ平均 μ_j および分散共分散行列 A_j に適当な初期値を与えておく．

(F1) z_{ij} の推定

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の 2 乗距離 d_{ij}^2 は，以下の式

$$d_{ij}^2 = (x_i - \mu_j)^T A^{-1} (x_i - \mu_j) \tag{7}$$

で与えられる．

要素 x_i がクラスタ C_j に属する確率 z_{ij} を，

$$z_{ij} = \frac{1}{Z_i} \exp(-\frac{1}{\lambda}\, d_{ij}^2) \tag{8}$$

とする．ただし，

$$Z = \sum_{k=1}^{j} \exp(-\frac{1}{\lambda}\, d_{ik}^2) \tag{9}$$

である．このとき，変数 λ がどの分布に帰属するかの確率 (ファジィ度) を制御するパラメータになることは式からも分かるであろう．実際，λ が十分大きければ，どのクラスタの帰属も等確率に近づく．

(F2) μ_j および A_j の推定値の更新

EM アルゴリズムと同様の (4) 式よりパラメータの値を更新する．

十分な回数 $(F1)$，$(F2)$ を繰り返した時の μ_j および A_j が与えられたパラメータ $\lambda > 0$ および K に対する尤度を最大にするパラメータとなっている．ただし，推定された分布は各分布が平均ベクト μ_j および 分散共分散行列が A_j を持つ 2 次元正規分布の K 個の混合分布と考えて尤度を計算する．

(F3) 最適なパラメータ (モデル) の選択

λ および K を動かし，AIC が最小となるようなパラメータを求め，最適なモデルと推定する．

3 論文分類支援への応用

実データとして，原子分子物理学分野のジャーナル Phys. Rev. A 誌に掲載されている論文 16070 本のアブストラクトに現われる専門用語の出現頻度に基づき各論文を 3520 次元の特徴ベクトル[1] からなるデータを用いた．

柏木らの論文 [1] のデータでは専門家により，原子分子物理学データが掲載されている 126 件の論文をカテゴリ 1 とし，それ以外のデータをカテゴリ 0 として分類値が与えられている．

本研究においては，この分類値 1 を含むクラスタがファジィ K-平均法による分類 (教師なしの自己組織化) で抽出可能であるか否かの実験を試みた．

まず与えられたデータを前処理により 2 次元にまで縮約を行う．これは 3520 次元のベクトル値データの 2 次元への可視化に相当する．与えられたデータベクトルからなるデータ行列 (16070×3520) を作成する．この行列は非常に疎な状態にあり，多くの成分に 0 が入っている．この行列の列ベクトルでひとつにのみ値が入っているものは明らかに分類へ

[1] 導出は奈良女子大学の柏木ら [1] による

の判別への寄与はない．このように本研究では，余り多くの要素に値が入っていない列ベクトルを間引くことにより特徴ベクトルの次元の切り落としを行い，16070 × 225 のサイズのデータ行列 D を作成した．

次にこのデータ行列に PCA(主成分分析) を施した．具体的には正定値行列 D^TD の固有値分解を行い，第 1 固有値 (最大固有値) と第 2 固有値およびそれらに付随する単位固有ベクトルを求め，この 2 つの固有ベクトルに基づき 16070 個のデータの 2 次元平面への布置を行った.PCA により第 1 主成分と第 2 主成分を取り出したことになる．

もちろん，このような大幅な特徴ベクトルの次元の縮約は元々データの持っている情報の大きな損失を与えるが，人が目で見て瞬時に状況を目視するには 2 次元が最適であると考えられる．したがって，あえて 2 次元までの縮約を行った．

EM アルゴリズムによりクラスタリングを試みたが，パラメータが局所解に陥り推定が困難であった．そこで，ファジィ$K-$平均法においては，以下のように実験を行う．

(1)λ および K の推定

ファジィ$K-$平均法を独立に 20 回繰り返し，一番推定された頻度の高い λ と K を求める．

(2) アニーリング

(1) より求めた λ と K を固定し，繰り返し回数 t を温度とみなして，

$$\lambda = \lambda^* + 250/(t*t+1) \qquad (10)$$

を用いてアニーリングを行い，AIC を求める．

(3) 最適パラメータの推定

(2) を独立に 20 回繰り返し，その中で最小の AIC をとるパラメータ μ_j および A_j を最適パラメータとする．

クラスタリングを施した結果，パラメータ $\lambda^* = 2.0$，最適クラスタ数は $K = 7$ と推定された．実験結果を図 1 に示す．黒い点で示されるものが 2 次元に縮約されたデータの集合であり，白抜きの点で示されるものが，カテゴリ 1 のデータである．図からも分かるように，カテゴリ 1 に属する 126 個のデータは，互いに遠くない位置に布置されることが分かる．

図中の楕円は最小の AIC をとるパラメータ μ_j および A_j の中でカテゴリ 1 のデータを含んでいると予想されるクラスタのものを元に，それぞれ内側からマハラノビスの距離が 1, 2, 3 の距離を元に描いたものである．また，クラスタ中心は △ である．距離 1 の楕円で 32.3%，距離 2 の楕円で 50.0%，距離 3 の楕円で 65.8%，カテゴリ 1 のデータをカバーできた．

図 1 実行結果

4 まとめ

本研究では類似性のある EM アルゴリズムとファジィ$K-$平均法を用いて，最適クラスタ数 K を推定する手法を提案した．実データへの応用では局所解に陥るリスクを軽減するため，ファジィ$K-$平均法において，ファジィ度によるアニーリングを試み，収束の度合いを高めた．今後の課題は，さらに局所解を避けるアルゴリズムを提案し，実データへの有効性を高めることである．

謝辞 データの提供および有益な議論を頂きました奈良女子大学大学院の柏木裕恵氏，城和貴教授に心より感謝致します．

参考文献

1) 柏木裕恵，高田雅美，佐々木明，城和貴: アブストラクトを用いた論文分類システムの設計と実装，preprint. (2006)

2) 宮岸聖高，市橋秀友，本田克宏: K-L 情報量正則化 FCM クラスタリング法，日本ファジィ学会誌，13，4，pp. 406-417, (2001)

3) 宮本定明: クラスター分析入門，ファジィクラスタリングの理論と応用，森北出版，(1999)