擬マハラノビス距離によるファジィ c -Means クラスタリング*

大上 将之**・市橋 秀友***・三好 哲也***

Fuzzy c -Means Clustering Algorithm with Pseudo Mahalanobis Distances*

Masayuki OHUE**, Hidetomo ICHIHASHI*** and Tetsuya MIYOSHI***

1. はじめに

Gustafson and Kessel^{1),2)}によるファジィ散布行列に基づくマハラノビス距離を用いるファジィクラスタリング法は、平面や多様体状の線形クラスターを発見することができる。しかし、多くのファジィ c -Means 法¹⁾の応用の中で、あまり用いられていないのは、分散共分散行列の逆行列や、行列式の値を求める必要があり、 c -Means 法に比べて煩雑であるためと考えられる。

本研究では、擬似的なマハラノビス距離を用いて線形多様体や超楕円体状のクラスターを形成するラグランジュ関数法による反復解法を提案する。提案法は分散共分散行列を計算する程度の労力でマハラノビス距離を用いるのと同様の効果が期待できる。

2. 擬マハラノビス距離による簡便クラスタリング法

J 個のパターンを C 個のクラスターに分類する問題を考える。パターン j の I 次元特徴量ベクトルを

$$\mathbf{x}_j = (x_{j1}, x_{j2}, \dots, x_{jI}), \quad j = 1, \dots, J \quad (1)$$

パターン j のクラスター c へのメンバシップを u_{cj} , $j = 1, \dots, J$ とする。クラスター c の中心ベクトルを

$$\mathbf{v}_c = (v_{c1}, v_{c2}, \dots, v_{cI}), \quad c = 1, \dots, C \quad (2)$$

クラスター c 内の擬マハラノビス距離を定義する $I \times I$ 対称行列 $\mathbf{E}_c = (e_{cik})$ を、行列 $\mathbf{A}_c = (a_{cik})$, $a_{cik} = a_{cki}$,

$\mathbf{B}_c = (b_{cik})$, $i \neq k$ のとき, $b_{cik} = b_{cki} = 0.0$, $\mathbf{D}_c = (d_{cik})$, $d_{cik} = d_{cki}$ の和で表し、行列 $\mathbf{E}_c = (e_{cik})$ の要素を

$$e_{cik} = (1 - \delta_{ik})a_{cik} + \delta_{ik}b_{cik} + (1 - \delta_{ik})d_{cik} \quad (3)$$

とする。ただし, $a_{cik} > 0$, $b_{cii} > 0$, $d_{cik} < 0$ であり, δ_{ik} は Kronecker の記号

$$\delta_{ik} = \begin{cases} 1 & (i=k) \\ 0 & (i \neq k) \end{cases}$$

である。

クラスタリングの目的関数と制約条件を次のラグランジュ関数により表す。

$$\begin{aligned} \min L = & \sum_{j=1}^J \sum_{c=1}^C \sum_{i=1}^I \sum_{k=1}^I u_{cj} \{ (1 - \delta_{ik})a_{cik} + \delta_{ik}b_{cik} \\ & + (1 - \delta_{ik})d_{cik} \} (x_{ji} - v_{ci})(x_{jk} - v_{ck}) \\ & - \sum_{c=1}^C \lambda_c^a \left(\sum_{i=1}^I \sum_{k=1}^I a_{cik} - 1 \right) \\ & - \sum_{c=1}^C \lambda_c^b \left(\sum_{i=1}^I b_{cii} - 1 \right) \\ & - \sum_{c=1}^C \lambda_c^d \left(\sum_{i=1}^I \sum_{k=1}^I d_{cik} + 1 \right) \\ & + \lambda_0 \sum_{c=1}^C \left\{ \sum_{i=1}^I \sum_{k=1}^I (a_{cik} \log a_{cik} \right. \\ & \left. + (-d_{cik}) \log (-d_{cik})) + \sum_{i=1}^I b_{cii} \log b_{cii} \right\} \\ & - \sum_{j=1}^J \tau_j \left(\sum_{c=1}^C u_{cj} - 1 \right) \\ & + \tau_0 \sum_{j=1}^J \sum_{c=1}^C u_{cj} \log u_{cj} \end{aligned} \quad (4)$$

第1項目はクラスター c 内の行列 $\mathbf{E}_c$ に関する擬マハラノビス距離を表していて, a_{cik} , b_{cik} , d_{cik} は決定変数である。第2, 3項目はそれぞれ, 行列 $\mathbf{A}_c$, $\mathbf{B}_c$ の成分

* 原稿受付 1998年7月6日

** 大阪府立大学 大学院 工学研究科 Graduate School of Engineering, Osaka Prefecture University; 1-1 Gakuen-cho, Sakai city, Osaka 599-8531, JAPAN

*** 大阪府立大学 工学部 経営工学科 Department of Industrial Engineering, Faculty of Engineering, Osaka Prefecture University; 1-1 Gakuen-cho, Sakai city, Osaka 599-8531, JAPAN

Key Words: fuzzy clustering, fuzzy c -means, linear clusters.

の和が1であるという制約条件を表して、 λ_c^a, λ_c^b , $c=1, \dots, C$ はラグランジュ乗数である。第4項目は行列 D_c の成分の和が-1であるという制約条件を表して、 λ_c^d , $c=1, \dots, C$ はラグランジュ乗数である。第5項目は $a_{cik} > 0, b_{cii} > 0, d_{cik} < 0$ とするためのもので、第2から4項目より $a_{cik}, b_{cii}, -d_{cik} \in (0, 1)$ となる。第6項目はメンバシップ u_{cj} の c についての和が1であるという制約条件を表し、 $\tau_j, j=1, \dots, J$ はラグランジュ乗数である。第7項目はファジィc-Meansでのメンバシップのべき乗の代わりに用いたエントロピー項³⁾である。なお、 $\tau_0 (> 0)$ はファジィクラスターを得るための重みであり、大きいほどあいまいなクラスターとなる。

L の最適性の必要条件より

$$a_{cik} = \frac{\exp(F_{cik})}{\sum_{m=1}^I \sum_{n=1}^I \exp(F_{cmn})} \quad (5)$$

$$F_{cmn} = -\frac{1}{\lambda_0} \sum_{j=1}^J (1 - \delta_{mn}) u_{cj} (x_{jm} - v_{cm})(x_{jn} - v_{cn}) \quad (6)$$

同様に

$$b_{cii} = \frac{\exp(G_{cii})}{\sum_{m=1}^I \exp(G_{cmm})} \quad (7)$$

$$G_{cmm} = -\frac{1}{\lambda_0} \sum_{j=1}^J \delta_{mm} u_{cj} (x_{jm} - v_{cm})^2 \quad (8)$$

$$d_{cik} = \frac{-\exp(H_{cik})}{\sum_{m=1}^I \sum_{n=1}^I \exp(H_{cmn})} \quad (9)$$

$$H_{cmn} = \frac{1}{\lambda_0} \sum_{j=1}^J (1 - \delta_{mn}) u_{cj} (x_{jm} - v_{cm})(x_{jn} - v_{cn}) \quad (10)$$

クラスター中心は

$$v_{ck} = \frac{\sum_{j=1}^J u_{cj} x_{jk}}{\sum_{j=1}^J u_{cj}}, \quad k=1, \dots, I \quad (11)$$

メンバシップは

$$u_{cj} = \frac{\exp(R_{cj})}{\sum_{q=1}^C \exp(R_{qj})} \quad (12)$$

$$R_{cj} = -\frac{1}{\tau_0} \sum_{i=1}^I \sum_{k=1}^I \{ (1 - \delta_{ik}) a_{cik} + \delta_{ik} b_{cii} + (1 - \delta_{ik}) d_{cik} \} (x_{ji} - v_{ci})(x_{jk} - v_{ck}) \quad (13)$$

となる。

クラスタリングアルゴリズムを以下に示す。

ステップ 1 C を $2 \leq C \leq J$ より定める。 λ_0 と τ_0 を定める。メンバシップ $u_{cj}, j=1, \dots, J, c=1, \dots, C$ を乱数により $\sum_{c=1}^C u_{cj} = 1$ となるように定める。

ステップ 2 (11) 式より $v_c, c=1, \dots, C$ を求める。

ステップ 3 (5) 式, (7) 式, (9) 式より A_c, B_c, D_c を求める。

ステップ 4 (12) 式よりメンバシップ u_{cj} を求める。

ステップ 5 収束判定条件

$$\max_{c,j} |u_{cj} - u_{cj}^{OLD}| < \varepsilon$$

を満たせば終了。そうでなければ、ステップ 2へ。

3. 数値実験

まず、参考文献2)でとりあげられている2次元20個のパターンを用いた。ただし、以下の二つの実験におけるパターンは $[-0.5, 0.5] \times [-0.5, 0.5]$ 内に基準化されている。パターンの座標は Table 1 の1列目に示すような、二つの直交する線形構造を表している。 $\lambda_0 = 0.10$, $\tau_0 = 0.05$, $\varepsilon = 0.00001$ とした。反復回数10回で Table 1 の2列目に示すメンバシップ値 u_{1j} が得られた。Fig. 1で×はメンバシップ値が0.5以上でクラスター1に属するパターン、○はクラスター2に属するパターンを表している。以下にクリスプなクラスターの分散共分散行列 (M_1, M_2), Gustafsonら²⁾の方法により得られたファジィ分散共分散行列 (N_1, N_2), そして今回提案した手法により得られた (E_1^{-1}, E_2^{-1}) を示す。

$$M_1 = \begin{bmatrix} 0.13 & 1.50 \\ 1.50 & 24.44 \end{bmatrix}, M_2 = \begin{bmatrix} 33.92 & -0.03 \\ -0.03 & 0.07 \end{bmatrix}$$

$$N_1 = \begin{bmatrix} 0.06 & 0.72 \\ 0.72 & 25.65 \end{bmatrix}, N_2 = \begin{bmatrix} 36.33 & 0.88 \\ 0.88 & 0.05 \end{bmatrix}$$

$$E_1^{-1} = \begin{bmatrix} 3.46 & 32.45 \\ 32.45 & 429.25 \end{bmatrix}, E_2^{-1} = \begin{bmatrix} 1407.04 & 23.94 \\ 23.94 & 1.41 \end{bmatrix}$$

Table 1 Obtained fuzzy clusters

Data x_j	Terminal memberships in fuzzy cluster 1	
(-9.75, -0.15)	0.007	0.011
(-6.44, 0.34)	0.102	0.115
(-4.69, -0.30)	0.234	0.248
(-2.04, 0.37)	0.445	0.441
(-1.24, 0.45)	0.484	0.478
(0.33, -0.08)	0.502	0.502
(5.04, -0.21)	0.241	0.279
(5.86, -0.25)	0.171	0.212
(7.54, 0.16)	0.068	0.100
(7.67, 0.24)	0.063	0.094
(-0.30, -8.07)	0.944	0.926
(0.13, -7.13)	0.897	0.873
(-0.37, -5.18)	0.745	0.725
(0.03, -3.33)	0.596	0.587
(0.35, -2.63)	0.554	0.548
(0.23, -2.68)	0.557	0.551
(-0.05, -2.00)	0.527	0.524
(0.41, 0.37)	0.510	0.508
(0.69, 4.75)	0.800	0.773
(0.74, 8.87)	0.988	0.980

Fig. 1 Gustafson's cross

すべての行列において対角要素の値がクラスター 1 と 2 の線形構造を反映している。

つぎに, Gustafson らのデータを (0,0) を中心とし 45 度回転させ, $\lambda_0 = 0.10$, $\tau_0 = 0.05$, $\varepsilon = 0.00001$ とした。反復回数 13 回で Table 1 の 3 列目のようなメンバシップ値 u_{1j} が得られた。

$$E_1^{-1} = \begin{bmatrix} 10.72 & 6.72 \\ 6.72 & 5.75 \end{bmatrix}, E_2^{-1} = \begin{bmatrix} 16.37 & -14.26 \\ -14.26 & 14.29 \end{bmatrix}$$

の二つの行列を比べると, 分散がほぼ等しく, 共分散は正負逆になっていて, それぞれのクラスターの形状をうまくとらえている。

最後に, 三つの交叉した線形構造をもった 3 次元 60 個のパターンを用いて実験を行った。Fig. 2 では三つの線形クラスターに分類できた結果を $x_1 - x_3$ 平面に射影して示している。ただし, Fig. 2 ではメンバシップ値の最も大きいクラスターにクリスプ分割して表示している。なお, 超楕円体状のクラスターに分類することもできるが, 紙面の都合で省略している。

Fig. 2 Three linear clusters in R^3

4. おわりに

本研究では, 擬似的なマハラノビス距離を用いるファジィクラスタリングのラグランジュ関数法による簡便な反復解法を提案した。マハラノビス距離は判別分析に広く用いられており, 提案法によるクラスタリングと判別分析の同時適用の有効性を検証することが今後の課題である。また, 提案法では, クラスター内でデータが特殊な分布をしていると距離を定義する行列が必ずしも正定値行列とならない。正定値行列とするための新たな拘束条件などについて検討する必要がある。

謝 辞

本研究の一部は文部省科学研究補助金・基礎研究 (C) (09838028) および平成 10 年度大阪府立大学大学院特別研究費の助成に基づいて行われたものであり, 謝意を表す。

参考文献

- 1) J. C. Bezdek: *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press (1981)
- 2) D. E. Gustafson and W. Kessel: Fuzzy clustering with a Fuzzy Covariance Matrix; *Proc. IEEE-CDC*, Vol. 2, pp. 761-766 (1979)
- 3) 宮本, 馬屋原, 向殿: ファジィc-平均法とエントロピー正則化法におけるファジィ分類関数; *日本ファジィ学会誌*, Vol. 10, No. 3, pp. 548-557 (1998)