原著論文

線形ファジィクラスタリングに基づく 混合データベースの局所的な主成分分析†

上杉 亮*・本多 克宏*・市橋 秀友*・野津 亮*

Fuzzy *c*-Varieties (FCV) 法は線形ファジィクラスタリング手法の一つであり、データ分割を考慮し低次元部分空間を推定する局所的な主成分分析に適用することができる。FCV法のクラスタリング基準はデータ点とプロトタイプの線形多様体との距離からなるが、その評価基準は最小2乗近似に基づいて定義することもでき、局所的な主成分分析法の一種ともみなされる。最適尺度法は、混合データベースの多変量解析への有用なアプローチであり、線形モデルの推定に適用されている。本論文では、最適尺度法に基づく混合データベースの局所的な主成分分析の定式化を、本質的に等価であるFCV法と最小2乗近似に基づく線形ファジィクラスタリング法の2種類のモデルから提案する。提案手法では、カテゴリカル変量の数量的得点を計算するステップを繰り返しアルゴリズムに導入する。

キーワード：クラスター分析、主成分分析、カテゴリカルデータ、最適尺度

1. はじめに

局所的な線形モデルの推定は、知識発見 (Knowledge Discovery in Databases, KDD) やデータマイニングへの有用なアプローチであり、主にデータの分類と線形モデルの推定のプロセスから成る。クラスターのプロトタイプとして線形多様体を用いるBezdekらのFuzzy *c*-Varieties (FCV) 法[1,2]は、代表的な線形ファジィクラスタリング法であるが、所属度合いを表すメンバシップを考慮したファジィ散布行列の固有ベクトルとしてプロトタイプを張るベクトルを算出することから、局所的な構造を捉えながら主成分分析を施す手法[3,4]と同等とみなされ、データベースからの知識発見への応用もなされている。Fuzzy *c*-Elliptotypes (FCE) 法[1,2]はFuzzy *c*-Means (FCM) 法とFCV法の融合手法であり、これらの目的関数の重みつき線形和を新たな目的関数として用いることで楕円体状のクラスターを抽出する。線形多様体からの距離だけでなくクラスター中心からの距離も考慮したデータ分割が行われることになり、FCM法による局所領域の抽出とFCV法の主成分の推定を融合することで、局所的な主成分分析の意味合いが強い手法であるといえる。

ところで、実世界で収集されたデータを分析する際

には、しばしば名義変量、すなわち単に項目を区別するために用いられる変量が含まれることがある。量的変量と質的変量が混在する混合データベースに対するクラスタリングアプローチはこれまでに種々提案されており、名義変量も考慮したクラスタリング基準[5,6]が用いられている。これらでは、名義変量について所属カテゴリの相違に基づく距離を定義することで尺度の混在を解決しているが、量的変量と質的変量を別個に取り扱っているため、線形多様体をプロトタイプとする線形ファジィクラスタリングへの応用は困難である。そこで、本論文では、線形ファジィクラスタリング法を混合データベースへ適用するために、名義変量の数量化を考慮したアプローチを考える。

多変量解析において用いられる交互最小2乗に基づく最適尺度法[7,8,9]は、尺度基準が混在する場合に適した手法であり、モデル推定と最適変換の段階からなる。交互最適化の繰り返しステップごとに名義変量をその時点で最適な量的尺度に数量化することで、単一の数量的空間におけるモデルの最適化が施される。本多ら[10]は混合データベースのFCM法に最適尺度法を応用し、数量的空間で単一の目的関数を最小化することによりデータ分割を行う手法を提案している。クラスターごとにカテゴリを割り当てる従来法と比べて、カテゴリ間の関連がより明確となる利点がある。

本論文では、最適尺度法に基づくFCM法を線形ファジィクラスタリングに拡張することで、混合データベースの局所的な主成分分析法を提案する。線形ファジィクラスタリングにはBezdekらの標準的な

† Local Principal Component Analysis for Mixed Databases Based on Linear Fuzzy Clustering
Ryo UESUGI, Katsuhiko HONDA, Hidetomo ICHIHASHI and Akira NOTSU

* 大阪府立大学大学院 工学研究科 電気・情報系専攻 知能情報工学分野
Graduate School of Engineering, Electrical Engineering and Information Science, Osaka Prefecture University

FCV法のほかに、最小2乗近似に基づく定式化[11]も提案されており、これら二つは本質的に等価であることが示されている。そこで、これらの定式化にFCM法を融合してFCE法へと拡張した上で、最適尺度法に基づく名義変量の処理法として2種類の定式化を行う。これらの提案手法では、繰り返しアルゴリズムに名義変量のカテゴリー得点を算出するステップを組み込むことにより、単一の数量的空間で楕円体状のクラスターが抽出される。

2. 局所的な主成分分析と名義変量の最適変換

2.1 最小2乗近似に基づく局所的な主成分分析とFCV法

m 次元の n 個の標本データからなる $(n \times m)$ データ行列 $X = (x_{ij})$ が与えられたときに、 n 個の標本データを C 個のクラスターに分割する問題を考える。データ行列は適宜、変量 j に係る n 個の要素を並べた縦ベクトル x_j を用いて $X = (x_1, \dots, x_j, \dots, x_m)$ 、または標本 i に係る m 個の要素を並べた縦ベクトル $\tilde{x}_i$ を用いて $X = (\tilde{x}_1, \dots, \tilde{x}_i, \dots, \tilde{x}_n)^T$ と表す。ただし、 T は転置を表す。(以降、太字はすべて縦ベクトルを表し、行列の行方向の要素を並べた縦ベクトルは $\tilde{\cdot}$ を付して表記することとする。)

FCV法はデータ集合を C 個の線形クラスターに分割する手法であり、その目的関数はデータ点と互いに線形独立な単位ベクトル a_{ck} により張られる p 次元の線形多様体との距離を分類尺度とすることで、以下の目的関数の最小化と定式化される[1, 2]。

$$L_{fcv} = \sum_{c=1}^C \sum_{i=1}^n (u_{ci})^\theta \left\{ (\tilde{x}_i - b_c)^T (\tilde{x}_i - b_c) - \sum_{k=1}^p a_{ck}^T R_{ci} a_{ck} \right\} \quad (1)$$

$$R_{ci} = (\tilde{x}_i - b_c)(\tilde{x}_i - b_c)^T \quad (2)$$

ただし、 u_{ci} は標本データ $\tilde{x}_i$ の第 c クラスターへの帰属度を表すメンバシップであり、

$$u_{ci} \in [0, 1], \quad c = 1, \dots, C, \quad i = 1, \dots, n$$

$$\sum_{c=1}^C u_{ci} = 1, \quad i = 1, \dots, n \quad (3)$$

となる確率的制約を満たすものとする。 b_c は第 c クラスターの中心である。メンバシップのべき乗はファジィ分割を得るために用いられたもので、 θ が大きいほどおのおのの標本データの所属が明確でなくなり、あいまいなデータ分割が得られるようになる。

最適性の必要条件 $\partial L_{fcv} / \partial a_{ck} = 0$ から、最適なベクトル a_{ck} は以下の固有値問題を解くことで得られる。

$$S_{fc} a_{ck} = \mu_{ck} a_{ck} \quad (4)$$

ここで、 S_{fc} は以下の一般化されたファジィ散布行列である。

$$S_{fc} = \sum_{i=1}^n (u_{ci})^\theta R_{ci} \quad (5)$$

最適な a_{ck} は大きな固有値 μ_{ck} に対応する固有ベクトルとして算出されることから、クラスターごとに帰属度を考慮しながら求められるファジィ主成分ベクトルであるとみなされる。

同様に、クラスター中心およびメンバシップもおのおの最適性の必要条件 $\partial L_{fcv} / \partial b_c = 0$ および $\partial L_{fcv} / \partial u_{ci} = 0$ から求められ、繰り返しアルゴリズムによりクラスタリング結果が得られる。

FCV法の目的関数はデータ点と線形多様体との距離により定義されるが、本多ら[11]は主成分分析で用いられる最小2乗基準にメンバシップを導入することにより、データ行列の低階数近似の立場から線形ファジィクラスタリングを論じている。最小2乗基準を用いた局所的な主成分分析の目的関数は、以下のように定義される。

$$L_{lsc} = \sum_{c=1}^C \text{tr} \left\{ (X - Y_c)^T U_c^\theta (X - Y_c) \right\} \quad (6)$$

ただし、 $U_c = \text{diag}(u_{c1}, \dots, u_{cn})$ であり、 tr は対角要素の和(トレース)を表す。 $Y_c = (y_{cij})$ は第 c クラスターにおけるデータ行列 X の低階数近似行列を表し、

$$Y_c = F_c A_c^T + \mathbf{1}_n b_c^T \quad (7)$$

である。ここで、 $F_c = (\tilde{f}_{c1}, \dots, \tilde{f}_{cn})^T$ はおのおのの標本データの主成分からなる $(n \times p)$ の成分得点行列、 $A_c = (a_{c1}, \dots, a_{cp})$ は主成分ベクトルを並べた $(m \times p)$ の主成分行列である。また、 $\mathbf{1}_n$ はすべての要素が1である n 次元ベクトルである。

メンバシップを固定して考えた場合、目的関数の最適性の必要条件 $\partial L_{lsc} / \partial b_c = 0$ から、制約条件として $F_c^T U_c^\theta \mathbf{1}_n = 0$ を考えると、最適なクラスター中心は

$$b_c = (\mathbf{1}_n^T U_c^\theta \mathbf{1}_n)^{-1} X^T U_c^\theta \mathbf{1}_n \quad (8)$$

のように決まる。(8)式はFCV法のアルゴリズムにおけるクラスター中心 b_c の更新則と等しくなる。(7)式を代入することにより、(6)式は

$$L_{lsc} = \sum_{c=1}^C \left\{ \text{tr}(X_c^\top U_c^\theta X_c) - 2\text{tr}(X_c^\top U_c^\theta F_c A_c^\top) + \text{tr}(A_c F_c^\top U_c^\theta F_c A_c^\top) \right\} \quad (9)$$

のようになる。ただし、 $X_c = X - \mathbf{1}_n \mathbf{b}_c^\top$ とおいた。さらに、 $\partial L_{lsc} / \partial F_c = 0$ から、以下の関係が導かれる。

$$F_c A_c^\top A_c = X_c A_c \quad (10)$$

ここで、FCV法と同様に局所的な主成分ベクトルに $A_c^\top A_c = I$ なる制約を設けるとすると、成分得点行列は $F_c = X_c A_c$ と求まり、目的関数は以下のように書き換えられる。

$$\begin{aligned} L_{lsc} &= \sum_{c=1}^C \left\{ \text{tr}(X_c^\top U_c^\theta X_c) - \text{tr}(A_c^\top X_c^\top U_c^\theta X_c A_c) \right\} \\ &= L_{fcv} \end{aligned} \quad (11)$$

このように、(6)式はFCV法と同じ最適解を与える目的関数であるといえる。

また、(6)式のようなデータ行列の低階数近似に基づく目的関数は、要素ごとに書き表すことにより、

$$L_{lsc} = \sum_{c=1}^C \sum_{i=1}^n (u_{ci})^\theta \sum_{j=1}^m (x_{ij} - \sum_{k=1}^p f_{cik} a_{ckj} - b_{cj})^2 \quad (12)$$

のように変形できることから、要素ごとの最小2乗近似に基づく定式化であるともみなされ、要素ごとに生じる欠測値やサンプル内ノイズの処理法へと発展されている[11,12]。ただし、クラスター数が1の場合はすべてのメンバシップが1となり、最小2乗基準に基づく主成分分析の目的関数に一致する。

2.2 名義変量の最適変換

名義変量を含む多変量データの解析法として、最適尺度に基づく手法が種々提案されている[7,8,9]。これらの手法では、推定される線形モデルから得られるその時点での最適変換に最も良く適合するようにカテゴリーの数量化ベクトルを求め、数量化された名義尺度を再び交互最適化に基づくアルゴリズムに代入することでコスト関数の最小化が図られる。また、van BuurenとHeiser [13]は k -Meansクラスタリングにおいて最適尺度を考慮することにより、混合データのクラスタリングを論じている。

本多ら[10]はファジィクラスタリングにおける名義尺度の最適変換としてクラスター中心を用いた数量化ベクトルの推定を想定し、名義尺度の最適変換を伴う

交互最適化アルゴリズムを提案している。

本研究では、線形クラスタリングにおいて、最適尺度法に基づく名義変量の処理法を提案する。

3. 最適尺度法に基づく混合データベースの線形ファジィクラスタリング

FCV法をデータベースに名義変量が含まれる場合に拡張することを考える。データ点とプロトタイプとなる線形多様体との距離で表される(1)式の目的関数は、名義尺度を含むデータ点と線形多様体との距離を定義できないため、名義尺度の最適変換に基づく数量化を施すことで拡張する。

おのおの K_j 個のカテゴリーからなる質的な観測値が n 個の標本データについて得られたとすると、第 j 変量(項目)に関する観測値は2値変量を用いて以下のような $(n \times K_j)$ ダミー変数行列 $G_j = (g_{ijk})$ で与えられる。

$$g_{ijk} = \begin{cases} 1; & \text{if object } i \text{ belongs to category } k. \\ 0; & \text{otherwise.} \end{cases}$$

$$G_j = \begin{pmatrix} g_{1j1} & \cdots & g_{1jk} & \cdots & g_{1jK_j} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ g_{ij1} & \cdots & g_{ijk} & \cdots & g_{ijK_j} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ g_{nj1} & \cdots & g_{njk} & \cdots & g_{njK_j} \end{pmatrix}$$

n 個の標本データに関する m 種類の観測値が得られた際に、はじめの $(m-q)$ 個が数量的な変量であり、残りの q 個が名義変量であるとする。

3.1 標準的なFCV法に基づく定式化

BezdekらがFCV法とFCM法の目的関数を重み付きで足し合わせることで、楕円体状のクラスターが得られるFCE法[1,2]を定式化したのに倣い、最適尺度法に基づく混合データベースの線形ファジィクラスタリングの目的関数を以下のように定義する。

$$\begin{aligned} L_{fcvcos1} &= \sum_{c=1}^C \sum_{i=1}^n (u_{ci})^\theta \left\{ \alpha \left(\sum_{j=1}^{m-q} (x_{ij} - b_{cj})^2 \right) \right. \\ &\quad + \sum_{j=m-q+1}^m (g_{ij}^\top \mathbf{q}_j - b_{cj})^2 \\ &\quad \left. - \sum_{k=1}^p \left(\sum_{j=1}^{m-q} a_{ckj} (x_{ij} - b_{cj}) \right) \right\} \end{aligned}$$

$$\begin{aligned}
& + \sum_{j=1}^{m-q} a_{ckj} (g_{ij}^\top q_j - b_{cj})^2 \Big) \\
& + (1-\alpha) \left(\sum_{j=1}^{m-q} (x_{ij} - b_{cj})^2 \right. \\
& \left. + \sum_{j=m-q+1}^m (g_{ij}^\top q_j - b_{cj})^2 \right) \Big\} \quad (13)
\end{aligned}$$

ここで、 α はFCV法とFCM法の優先度を規定するドレードオフ係数で、0のときは混合データベースのFCM法[10]と等しく、1に近づけるにつれて楕円形状のクラスターが得られるようになる。 q_j はカテゴリーの数量的得点を並べたベクトルであり、繰り返しアルゴリズムの中で推定すべき未知ベクトルである。 q_j を暫時既知として固定した場合には $g_{ij}^\top q_j$ を第 j 数量的変数 x_{ij} とみなすことで標準的なFCE法の目的関数と同等と考えることができる。

提案手法においてカテゴリーの数量的得点ベクトルは変量ごとに算出する。最適な q_j を得るために、(13)式を

$$\begin{aligned}
L_{f_{ceosI}} &= \sum_{c=1}^C \text{tr} \left((X^* - \mathbf{1}_n \mathbf{b}_c^{*\top})^\top U_c^\theta (X^* - \mathbf{1}_n \mathbf{b}_c^{*\top}) \right) \\
&+ \sum_{c=1}^C \left((G_j q_j - \mathbf{1}_n b_{cj})^\top U_c^\theta (G_j q_j - \mathbf{1}_n b_{cj}) \right) \\
&- \alpha \sum_{c=1}^C \sum_{k=1}^p \left(\mathbf{a}_{ck}^{*\top} (X^* - \mathbf{1}_n \mathbf{b}_c^{*\top})^\top U_c^\theta \right. \\
&\quad \cdot (X^* - \mathbf{1}_n \mathbf{b}_c^{*\top}) \mathbf{a}_{ck}^* \Big) \\
&- 2\alpha \sum_{c=1}^C \sum_{k=1}^p \left(\mathbf{a}_{ck}^{*\top} (X^* - \mathbf{1}_n \mathbf{b}_c^{*\top})^\top U_c^\theta \right. \\
&\quad \cdot (G_j q_j - \mathbf{1}_n b_{cj}) a_{ckj} \Big) \\
&- \alpha \sum_{c=1}^C \sum_{k=1}^p \left(a_{ckj} (G_j q_j - \mathbf{1}_n b_{cj})^\top U_c^\theta \right. \\
&\quad \cdot (G_j q_j - \mathbf{1}_n b_{cj}) a_{ckj} \Big) \quad (14)
\end{aligned}$$

のように変形する。ここで、 X^* は $(m-1)$ 個の量的変量からなる $(n \times (m-1))$ データ行列であり、 $\mathbf{b}_c^* = (b_{c1}, b_{c2}, \dots, b_{c(m-1)})^\top$ 、 $\mathbf{a}_{ck}^* = (a_{ck1}, a_{ck2}, \dots, a_{ck(m-1)})^\top$ である。また第 j 変量のダミー変量行列を $G_j = (g_{1j}, g_{2j}, \dots, g_{nj})^\top$ とおいた。(14)式は2乗基準を用いてお

り、2次導関数が常に正なので、最適性の必要条件

$$\begin{aligned}
\frac{\partial L_{f_{ceosI}}}{\partial q_j} &= 2 \sum_{c=1}^C G_j^\top U_c^\theta (G_j q_j - \mathbf{1}_n b_{cj}) \\
&- 2\alpha \sum_{c=1}^C \sum_{k=1}^p G_j^\top U_c^\theta (X^* - \mathbf{1}_n \mathbf{b}_c^{*\top}) \mathbf{a}_{ck}^* a_{ckj} \\
&- 2\alpha \sum_{c=1}^C \sum_{k=1}^p G_j^\top U_c^\theta (G_j q_j - \mathbf{1}_n b_{cj}) a_{ckj}^2 \\
&= 2 \left(G_j^\top \left(\sum_{c=1}^C \left(1 - \alpha \sum_{k=1}^p a_{ckj}^2 \right) U_c^\theta \right) G_j \right) q_j \\
&- 2 \sum_{c=1}^C G_j^\top U_c^\theta \left(\alpha \sum_{k=1}^p (X^* - \mathbf{1}_n \mathbf{b}_c^{*\top}) \mathbf{a}_{ck}^* a_{ckj} \right. \\
&\quad \left. + \left(1 - \alpha \sum_{k=1}^p a_{ckj}^2 \right) \mathbf{1}_n b_{cj} \right) \\
&= \mathbf{0} \quad (15)
\end{aligned}$$

から、数量的得点ベクトルは以下のように更新される。

$$\begin{aligned}
q_j &= \left(G_j^\top \left(\sum_{c=1}^C \left(1 - \alpha \sum_{k=1}^p a_{ckj}^2 \right) U_c^\theta \right) G_j \right)^{-1} \\
&\cdot \left(\sum_{c=1}^C G_j^\top U_c^\theta \left(\alpha \sum_{k=1}^p (X^* - \mathbf{1}_n \mathbf{b}_c^{*\top}) \mathbf{a}_{ck}^* a_{ckj} \right. \right. \\
&\quad \left. \left. + \left(1 - \alpha \sum_{k=1}^p a_{ckj}^2 \right) \mathbf{1}_n b_{cj} \right) \right) \quad (16)
\end{aligned}$$

ここで、 $G_j^\top (\sum_{c=1}^C U_c^\theta) G_j$ は第 k カテゴリーに属する個体のメンバシップの和を第 k 対角要素にもつ $(K_j \times K_j)$ 対角行列である。 $\sum_{k=1}^p (X^* - \mathbf{1}_n \mathbf{b}_c^{*\top}) \mathbf{a}_{ck}^* a_{ckj}$ は基底ベクトル $\mathbf{a}_{ck}^*$ により張られる線形多様体への量的な観測値の射影であるので、 q_j の第 k 要素は第 k カテゴリーに属する個体の比率を重みとした b_{cj} とプロトタイプである主成分軸との距離の重みつき平均値である。ただし、唯一の解を求めるために、以下の制約条件を付加する。

$$\mathbf{1}_n^\top G_j q_j = 0 \quad (17)$$

$$\frac{1}{n} q_j^\top G_j^\top G_j q_j = \sigma^2 \quad (18)$$

(17)式は、数量化された名義変量の平均を0とする

ためのものである。また、(18)式における σ^2 は分散パラメータであり、名義変量の重要度を調整する。つまり、 σ^2 が大きいほど名義変量を重視した分析が行われる。

クラスター中心とメンバシップに関しては、 $\mathbf{g}_{ij}^\top \mathbf{q}_{ij}$ を数量的変量 x_{ij} とみなし、 $\partial L_{f_{ceos1}} / \partial \mathbf{b}_c = \mathbf{0}$ および $\partial L_{f_{ceos1}} / \partial u_{ci} = 0$ から更新則が導かれる。したがって、 $\mathbf{b}_c$ 、 $\mathbf{a}_{ck}$ 、 u_{ci} の更新を繰り返す標準的なFCE法に、 $\mathbf{q}_j$ の更新を組み込んだ繰り返しアルゴリズムによりクラスタリング結果が得られる。

3.2 最小2乗近似に基づく定式化

次に、最小2乗基準を用いた線形ファジィクラスタリングを、名義変量が含まれる混合データベースに適用する場合を考える。前節で述べた名義変量の最適変換の手法とFCE法を考慮することで、最小2乗基準を用いた線形ファジィクラスタリングの目的関数を以下のように定義する。

$$L_{f_{ceos2}} = \sum_{c=1}^C \sum_{i=1}^n (u_{ci})^\theta \left\{ \alpha \left(\sum_{j=1}^{m-q} (x_{ij} - \sum_{k=1}^p f_{cik} a_{ckj} - b_{cj})^2 \right) + \sum_{j=m-q+1}^m (\mathbf{g}_{ij}^\top \mathbf{q}_j - \sum_{k=1}^p f_{cik} a_{ckj} - b_{cj})^2 \right\} + (1-\alpha) \left(\sum_{j=1}^{m-q} (x_{ij} - b_{cj})^2 + \sum_{j=m-q+1}^m (\mathbf{g}_{ij}^\top \mathbf{q}_j - b_{cj})^2 \right) \quad (19)$$

ただし、(19)式の目的関数を最小とする解を唯一に求めるために、

$$F_c^\top U_c F_c = I, \quad c = 1, \dots, C \quad (20)$$

$$F_c^\top U_c \mathbf{1}_n = \mathbf{0}, \quad c = 1, \dots, C \quad (21)$$

および $A_c^\top A_c$ が対角行列であるという条件を付加する。最適な数量的得点ベクトル $\mathbf{q}_j$ を得るために、(19)式の目的関数を以下のように書き換える。

$$L_{f_{ceos2}} = \sum_{c=1}^C \left\{ \alpha \left(\sum_{j=1}^{m-q} ((\mathbf{x}_j - F_c \tilde{\mathbf{a}}_{cj} - \mathbf{1}_n b_{cj})^\top U_c^\theta \cdot (\mathbf{x}_j - F_c \tilde{\mathbf{a}}_{cj} - \mathbf{1}_n b_{cj})) \right) + \sum_{j=m-q+1}^m ((G_j \mathbf{q}_j - F_c \tilde{\mathbf{a}}_{cj} - \mathbf{1}_n b_{cj})^\top U_c^\theta \right.$$

$$\cdot (G_j \mathbf{q}_j - F_c \tilde{\mathbf{a}}_{cj} - \mathbf{1}_n b_{cj})) \Big) + (1-\alpha) \left(\sum_{j=1}^{m-q} ((\mathbf{x}_j - \mathbf{1}_n b_{cj})^\top U_c^\theta \cdot (G_j \mathbf{q}_j - \mathbf{1}_n b_{cj})) \right) + \sum_{j=m-q+1}^m ((G_j \mathbf{q}_j - \mathbf{1}_n b_{cj})^\top U_c^\theta \cdot (G_j \mathbf{q}_j - \mathbf{1}_n b_{cj})) \Big) \Big\} \quad (22)$$

ただし、 $A_c = (\tilde{\mathbf{a}}_{c1}, \dots, \tilde{\mathbf{a}}_{cm})^\top$ である。最適性の必要条件

$$\frac{\partial L_{f_{ceos2}}}{\partial \mathbf{q}_j} = 2\alpha \sum_{c=1}^C G_j^\top U_c^\theta (G_j \mathbf{q}_j - F_c \tilde{\mathbf{a}}_{cj} - \mathbf{1}_n b_{cj}) + 2(1-\alpha) \sum_{c=1}^C G_j^\top U_c^\theta (G_j \mathbf{q}_j - \mathbf{1}_n b_{cj}) = \mathbf{0} \quad (23)$$

から、数量的得点ベクトルは以下のように更新される。

$$\mathbf{q}_j = \left(G_j^\top \left(\sum_{c=1}^C U_c^\theta \right) G_j \right)^{-1} \cdot \left(\sum_{c=1}^C G_j^\top U_c^\theta (\alpha F_c \tilde{\mathbf{a}}_{cj} + \mathbf{1}_n b_{cj}) \right) \quad (24)$$

ただし、唯一の解を得るために、(17)、(18)式の制約条件を付加する。(22)式において、 $\mathbf{q}_j$ を暫時既知として固定した場合には $\mathbf{g}_{ij}^\top \mathbf{q}_j$ を第 j 数量的変量 x_{ij} とみなすことにより $\partial L_{f_{ceos2}} / \partial \tilde{\mathbf{a}}_{cj} = \mathbf{0}$ および $\partial L_{f_{ceos2}} / \partial b_{cj} = 0$ から、更新則は、

$$\tilde{\mathbf{a}}_{cj} = (F_c^\top U_c^\theta F_c)^{-1} F_c^\top U_c^\theta (\mathbf{x}_j - \mathbf{1}_n b_{cj}) \quad (25)$$

$$b_{cj} = (\mathbf{1}_n^\top U_c^\theta \mathbf{1}_n)^{-1} \mathbf{1}_n^\top U_c^\theta (\mathbf{x}_j - \alpha F_c \tilde{\mathbf{a}}_{cj}) \quad (26)$$

のように求めることができる。

同様に、 F_c および u_{ci} についても、 $\mathbf{g}_{ij}^\top \mathbf{q}_j$ を第 j 数量的変量 x_{ij} とみなし、(19)式を、

$$L_{f_{ceos2}} = \sum_{c=1}^C \sum_{i=1}^n (u_{ci})^\theta \left\{ \alpha \|\tilde{\mathbf{x}}_i - A_c \tilde{\mathbf{f}}_{ci} - \mathbf{b}_c\|^2 + (1-\alpha) \|\tilde{\mathbf{x}}_i - \mathbf{b}_c\|^2 \right\} \quad (27)$$

と変形すると, $\partial L_{f_{ceos2}} / \partial \tilde{f}_{ci} = \mathbf{0}$ および $\partial L_{f_{ceos2}} / \partial u_{ci} = 0$ より

$$\tilde{f}_{ci} = (A_c^\top A_c)^{-1} A_c^\top (\tilde{x}_i - b_c) \quad (28)$$

$$u_{ci} = \left[\sum_{l=1}^C \left(\frac{d_{ci}^2}{d_{li}^2} \right)^{\frac{1}{\sigma-1}} \right]^{-1} \quad (29)$$

が導かれる. ただし,

$$d_{ci}^2 = \alpha \|\tilde{x}_i - A_c \tilde{f}_{ci} - b_c\|^2 + (1 - \alpha) \|\tilde{x}_i - b_c\|^2 \quad (30)$$

である.

以上の更新則を用いた提案法のアルゴリズムは, 以下の通りである.

Algorithm: Fuzzy c -Elliptotypes (FCE) based on Optimal Scaling and Lower Rank Approximation

Step 1. 乱数を用いて U_c , A_c , b_c , F_c を初期化し, (3), (20), (21) 式および $A_c^\top A_c$ が対角行列となる制約条件を満たすように基準化する.

Step 2. (24) 式を用いてカテゴリーの数量的得点ベクトル q_j を更新し, (17) 式および (18) 式を満たすように基準化する.

Step 3. (25) 式を用いて A_c を更新し, $A_c^\top A_c$ が対角行列になるように変換する.

Step 4. (28) 式を用いて F_c を更新し, (20) 式および (21) 式の制約条件を満たすように基準化する.

Step 5. (26) 式を用いて b_c を更新する.

Step 6. (29) 式を用いて U_c を更新する. ただし, $d_{ci}^2 = 0$ となる c が c_i 個あれば, $u_{ci} = 1/c_i$ とする.

Step 7. すべてのパラメータが収束すれば, 終了. それ以外は Step 2 へ戻る.

4. 数値実験

提案法の有効性を検証するために, 人工データを用いた数値実験を行った. 図1に示す散布図を持つデータに提案法を適用した. 図1のデータは3次元の観測値 ($m=3$) であり, 第1 (x_1), 第2 (x_2) 変量が量的変量, 第3変量が3つのカテゴリー (●: Category1, △: Category2, ×: Category3) から成る名義変量である. 実験の目的は, 名義変量に最適変換を施しながらデータ集合を2つの線形クラスターに分割することである. このとき, FCE法とFCM法の優先度を調整するトレードオフ係数, 名義変量の数量的得点の分散パラメータをそれぞれ $\alpha=0.7$, $\sigma^2=1.0$ とした. また, ファジィ度を $\theta=1.5$ とした. このデータに2種類の提案法を適用して得られた結果を図2, 3および表1, 2に示す. 図2, 3は, 提案法によって得られた2つの

図1 人工データ1 (●:Category1, △:Category2, ×:Category3)

図2 標準的な提案法で得られたプロトタイプ(人工データ1)

図3 最小2乗近似を用いた提案法で得られたプロトタイプ(人工データ1)

クラスターにおけるプロトタイプとなる直線(線形多様体)を表しており, それぞれ x_1-x_2 , x_1-x_3 空間である. x_3 は, 数量化された名義変量を示している. 表1はおのおののクラスターにおける主成分ベクトル(プロトタイプとなる直線の基底ベクトル), 表2は名義変量に与えられたカテゴリーの数量的得点を表している. 提案手法は量的変量で構成された局所的な線形構造に適した形でカテゴリーの数量的得点が算出される

表1 提案手法により得られたファジィ主成分ベクトル

変量	標準的		最小2乗近似	
	$c=1$	$c=2$	$c=1$	$c=2$
x_1	0.747	0.636	0.747	0.636
x_2	0.300	0.410	0.300	0.410
x_3	0.593	0.654	0.593	0.654

表2 カテゴリーの数量的得点

カテゴリー	標準的	最小2乗近似
1	-1.234	-1.236
2	0.020	0.022
3	1.215	1.213

ので、数量化された名義変量を第3変量とした3次元空間上で適切なプロトタイプ直線を捉えることができた。このように、2種類の提案手法は、潜在的な線形構造を抽出しながら、名義変量に最適変換を施すために有用である。

つづいて、図4のようなデータ形状が異なる部分構造を持つ人工データに提案法を適用した。図4のデータは3次元の観測値から成るが、線形の部分集合に球状の部分集合が隣接したデータ集合となっている。また、球状の部分集合においては、それぞれのデータが不規則にカテゴリーに属している。各パラメータの設定を先ほどと同様として、得られた結果を図5に示す。図5(a)から、クラスター中心をプロトタイプとするFCM法も同時に考慮していることから、直線状($c=2$)だけでなく球状のクラスター($c=1$)も捉えることができているが、球状のクラスターでは主成分軸が意味をなさない例となっている。また、図5(b)から、直線状のクラスター($c=2$)におけるカテゴリー間の関連をよく表す最適変換が行われ、他の量的変量との間の相関が捉えられていることがわかるが、カテゴリーと量的変量に関連のない球状のクラスター($c=1$)では数量化得点 x_3 と量的変量 x_1 がほぼ無相関となっている。このように、異なる形状のクラスターを抽出したり、数量化得点と量的変量の相関の有無を調べたりすることもできる。ただし、クラスターの形状を適応的に変化させるためには、重み α を形状に合わせて適応的に変化させる適応型線形ファジィクラスタリング[14]との融合が考えられる。さらに、数量化得点と量的変量の関連がクラスターごとに異なる際には、クラスターごとに名義変量の最適変換を施すことが望ましい場合もあるといえる。これらの発展は、今後の課題とする。

図4 人工データ2 (●:Category1, △:Category2, ×:Category3)

図5 提案法で得られたプロトタイプ(人工データ2)

5. POSデータへの適用

つづいて、提案手法を用いてPOSデータの解析を行った。POSとは「Point Of Sales」の略で店頭での販売情報を販売時点で把握するシステムを指す。そのシステムによって収集・蓄積されたデータがPOSデータである。ここでは、文献[15]で用いられたPOSデータに気象庁から配信されている天候データ(アメダスデータ)を加えて、提案法を適用しその有効性を検討する。

用いた変量を表3に示す。ただし、変量のうち、曜日(月曜、火曜、水曜、木曜、金曜、土曜、日曜)を最適変換すべき名義変量とみなして提案法を適用した。祝日の項目はダミー変数とし、あてはまるか否かで1または0の値を与えた。天気概況の項目は降水状況により0(晴れまたは曇り)から3(大雨)の値を与えた。日配品とは、牛乳、たまご、うどん、豆腐など工場で生産され毎日配送されるものを指し、日配品点数はPOS端末を通った客が購入した日配品の総数である。欠損している月日、および気象条件以外の社会的要因

表3 POSデータの解析に用いた変量

変量 18 次元
曜日 祝日
平均気温 9 時気温 12 時気温 15 時気温 18 時気温
相対湿度 昼天気概況 夜天気概況 日降水量
9-12 時降水量 12-15 時降水量 15-18 時降水量
来客数 (A 店) 日配品点数 (A 店)
来客数 (B 店) 日配品点数 (B 店)

による影響を大きく受けていると思われるデータ(例えば12月31日)を除いた331日分について、各数値変量ごとに平均0、分散1に基準化した。また、提案法では最小2乗近似に基づく定式化を用いた。

提案法を適用する際に、各パラメータを以下のように設定した。

クラスター数 $C: 1, 2$

求める主成分の数 $p: 1$

ファジィ度 $\theta: 1.5$

主成分分析に対する重み $\alpha: 0.7$

また、提案法の名義変量の重要度を調整するパラメータを、比較考察のために $\sigma^2=3.0, 5.0$ と適宜変化させた。

まず、POSデータ全体の構造を把握するために、クラスター数を $C=1$ として、データ分割を考慮せず名義変量の最適変換に基づく主成分分析を行った結果を表4、表5の2列目に示す。(表4では、主成分得点との相関係数である因子負荷量の絶対値が0.4以上のものを太字で示している。)因子負荷量から、2店の来客数と日配品点数は、平均気温、および9、12、15、18時の各気温との相関が高く、気温が高いほど総じて両店の客数、売上が増えるといえる。また、曜日とは負の相関があり、曜日の得点が低いほど客数、売上が増加する。表5の2列目の各曜日に与えられたカテゴリ得点に着目すると、得点の大きい順に、水曜、金曜、月曜、火曜、木曜、土曜、日曜となっている。つまり、2店の来客数・日配品数と曜日の負の相関から、水曜、金曜、月曜、火曜、木曜、土曜、日曜の順に客数、売上が増えることがわかる。このことは表6に示した両店の各曜日別平均来客数・日配品点数からも伺える。A店では木曜が最も来客数・日配品点数が多いが、両店とも合わせると日曜が最も多く、ついで土曜、木曜となっている。また逆に、水曜は両店ともに来客数・日配品点数が一番少なく、金曜、月曜という順になっている。このように、最適尺度法では、名義変量(曜日)を数値に変換することでカテゴ

表4 $C=1$ 、 $\sigma^2=3.0$ のときの因子負荷量

変量	因子負荷量
祝日	0.012
平均気温	-0.542
9 時気温	-0.536
12 時気温	-0.562
15 時気温	-0.559
18 時気温	-0.545
相対湿度	-0.032
昼天気概況	0.136
夜天気概況	-0.029
日降水量	0.080
9-12 時降水量	0.143
12-15 時降水量	0.099
15-18 時降水量	0.038
A 店来客数	-0.736
A 店日配品点数	-0.810
B 店来客数	-0.835
B 店日配品点数	-0.767
曜日	0.770

表5 各曜日に与えられたカテゴリ得点

曜日	$\sigma^2 = 3.0$		$\sigma^2 = 5.0$
	$C = 1$	$C = 2$	$C = 2$
月曜	1.307	0.202	2.815
火曜	-0.123	-0.492	-1.768
水曜	3.607	5.557	2.808
木曜	-0.855	-0.786	-1.804
金曜	1.758	0.723	2.864
土曜	-1.684	-0.969	-1.690
日曜	-2.171	-1.403	-1.795

表6 各曜日別の平均来客数・日配品点数

曜日	A 店		B 店	
	来客数	日配品点数	来客数	日配品点数
月曜	573	1545	579	1444
火曜	659	1823	639	1625
水曜	446	1175	501	1072
木曜	768	2120	618	1532
金曜	560	1455	544	1295
土曜	685	2106	707	2066
日曜	678	2108	750	2233

リー間の関係を把握することができる。

つづいて、クラスター数を $C=2$ とし、名義変量の重要度を $\sigma^2=3.0$ として、提案法を適用した結果を表5

の3列目と表7の2, 3列目および6, 7列目に示す。クラスター中心から、それぞれのクラスターは主に気温を基準にしてクラスタリングされており、クラスター1を寒冷期のクラスター、クラスター2を温暖期のクラスターとする。表7のファジィ因子負荷量を見ると、寒冷期のクラスター1では、日降水量が増えると相対湿度が高くなることがわかるが、来客数や日配品点数に関する知見は得られない。また、来客数・日配品点数と曜日との関連も低いことがわかる。温暖期のクラスター2では、クラスター数を $C=1$ として実験を行ったときと同様、2店の来客数と日配品点数は曜日と負の相関があり、曜日の得点が低いほど客数、売上が増加する傾向が見られる。表5の3列目の曜日に割り当てられた得点は、前述の実験と類似したものである。水曜、金曜、月曜、火曜、木曜、土曜、日曜の順に客数、売上が増える。これらのことは、表8からも読み取れる。表8は1, 2, 11, 12月(冬期)と6~9月(夏期)の曜日ごとの平均来客数・日配品点数を示している。各店とも水曜日の平均来客数・日配品点数が極端に少ないのは、通常水曜日が定休日、稀に営業している日があったからと考えられる。冬期と夏期の平均来客数・日配品点数を比べると、後者は曜日の得点通りの傾向が見られるのに対

して、前者はその傾向がなく、各曜日ごとの来客数・日配品点数の差も小さい。したがって、温暖期のクラスターでは曜日と来客数・日配品点数との相関を強調した主成分が得られたのに対して、寒冷期のクラスターでは強調されていない。

次に、クラスター数は先ほどと同様 $C=2$ とし、名義変量の重要度を $\sigma^2=5.0$ として、提案法を適用した結果を表7の4, 5列目および8, 9列目に示す。表7のクラスター中心から、来客数・日配品点数の多いクラスター1, 少ないクラスター2となっている。これは、曜日の変数の重要度を過剰に与えたために、曜日と関連のある来客数および日配品点数のみが強調されてFCMクラスタリングが行われたことを示している。そのために、クラスターごとの来客数・日配品点数および曜日の変数の分散が小さくなり、表7の8, 9列目からこれらとの関連を軽視した主成分が得られていることがわかる。このように、重要度 σ^2 を変化させることで、分析において重視する変数の選別が行われる。

以上のように、数値変量だけでなく名義変量も考慮することで、従来は不可能であった同時分析が可能となる。また、 σ^2 を順に変化させることで、名義変量と他の変量との関連の強さを考察することができる。

表7 $C=2$ のときのクラスター中心およびファジィ因子負荷量

変数	クラスター中心				ファジィ因子負荷量			
	$\sigma^2 = 3.0$		$\sigma^2 = 5.0$		$\sigma^2 = 3.0$		$\sigma^2 = 5.0$	
	$c = 1$	$c = 2$	$c = 1$	$c = 2$	$c = 1$	$c = 2$	$c = 1$	$c = 2$
祝日	-0.025	-0.035	0.011	-0.067	0.032	0.007	-0.024	0.034
平均気温	-0.759	0.701	0.004	-0.019	0.472	0.272	0.727	0.589
9時気温	-0.753	0.696	0.001	-0.018	0.476	0.263	0.720	0.583
12時気温	-0.767	0.712	-0.011	-0.003	0.421	0.282	0.723	0.579
15時気温	-0.765	0.713	-0.003	-0.006	0.385	0.280	0.718	0.574
18時気温	-0.758	0.703	0.006	-0.016	0.414	0.274	0.716	0.579
相対湿度	-0.120	0.044	0.045	-0.097	0.534	0.051	0.246	0.215
昼天気概況	0.059	-0.140	0.083	-0.128	0.595	-0.055	0.083	0.101
夜天気概況	-0.056	-0.010	0.009	-0.050	0.374	0.089	0.137	0.098
日降水量	0.068	-0.141	0.039	-0.106	0.613	-0.060	0.111	0.148
9-12時降水量	0.071	-0.108	0.075	-0.098	0.425	-0.086	0.030	0.104
12-15時降水量	0.059	-0.084	0.024	-0.061	0.446	-0.070	0.047	0.095
15-18時降水量	0.045	-0.089	-0.009	-0.050	0.398	-0.015	0.075	0.053
A店来客数	-0.142	0.130	-0.702	0.469	-0.279	0.616	0.247	0.117
A店日配品点数	-0.170	0.150	-0.759	0.503	-0.238	0.630	0.257	0.161
B店来客数	-0.208	0.191	-0.718	0.481	-0.241	0.618	0.261	0.204
B店日配品点数	-0.057	0.045	-0.736	0.484	-0.213	0.607	0.106	0.126
曜日	-0.309	0.230	2.588	-1.691	0.229	-0.713	-0.002	-0.107

表 8 冬期・夏期の各曜日別の平均来客数・日配品点数

曜日	冬期				夏期			
	A 店		B 店		A 店		B 店	
	来客数	日配品点数	来客数	日配品点数	来客数	日配品点数	来客数	日配品点数
月曜	556	1473	564	1419	575	1539	592	1424
火曜	628	1728	612	1587	695	1929	669	1681
水曜	456	1136	476	1046	466	1237	531	1147
木曜	717	1946	606	1532	801	2252	631	1533
金曜	675	2043	674	2046	569	1489	547	1291
土曜	675	2043	674	2046	702	2164	718	2016
日曜	639	1970	692	2111	705	2214	786	2316

6. おわりに

本研究では、主成分分析とファジィクラスタリングの同時分析における最適尺度法に基づく名義変量の処理法を2つのモデルの立場から定式化した。名義変量は繰り返しアルゴリズムの中で、FCEクラスタリングに適した数量的得点に逐次的に変換される。したがって、単一の目的関数の最小化に基づいて行われるデータ分割は、単一の数量的データ空間において実行されるといえる。

Bezdekらによる標準的なFCE法に基づく定式化では、名義変量を同時に2つ以上扱うことができないため、互いに関連する名義変量を順に最適変換が施される。一方、最小2乗近似に基づく混合データベースの線形クラスタリングでは、交互最小2乗の原理に従い複数の名義変量に対して互いに独立に最適変換が施されるという利点がある。また、繰り返しアルゴリズムは固有値計算を必要としないことから、理解が容易であるという利点がある。

さらに、最小2乗近似に基づく定式化は、全ての変量が質的なデータで表されるとき、局所的な等質性分析[16]と類似性がある。等質性分析は数量化分析3類とも関連の強い分析法であるが、カテゴリーの数量的得点の行列の階数を1とする制約条件の下で質的データのための主成分分析法ともみなされる。そして、すべての変量が質的なデータで表されるとき、提案法は局所的な等質性分析で同様の制約を用いた場合と類似性があるが、単一の数量的空間を構成するという点で相違がある。

これらの提案手法では、従来手法と同様のファジィ度を制御するパラメータ θ の決定に加え、名義変量の重要度を制御するパラメータ σ^2 を決定する必要がある。本論文の数値実験では、 σ^2 の値を変化させて結果の比較を行いながら名義変量と他の変量の関連を考

察したが、 σ^2 の設定の指標の作成は今後の課題である。また、提案手法の他の拡張として、クラスターの形状を適応的に変化させるアプローチ[14]との融合や、クラスターごとに最適変換を施す定式化などが考えられる。

謝辞

本研究の一部は科学研究費補助金・若手研究(B)(17700240)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] J. C. Bezdek, C. Coray, R. Gunderson and J. Watson: Detection and characterization of cluster substructure 2. fuzzy c -varieties and convex combinations thereof; *SIAM J. Appl. Math.*, vol. 40, no. 2, pp.358-372, 1981.
- [3] N. Kambhatla and T. K. Leen, "Dimension reduction by local principal component analysis," *Neural Computation*, vol. 9, no. 7, pp.1493-1516, 1997.
- [4] G. E. Hinton, P. Dayan and M. Revow, "Modeling the manifolds of images of handwritten digits," *IEEE Trans. on Neural Networks*, vol. 8 no. 1, pp.65-74, 1997.
- [5] C. Döring, C. Borgelt and R. Kruse, "Fuzzy clustering of quantitative and qualitative data," *Proc. Conf. North American Fuzzy Information Processing Society 2004*, pp.84-89, 2004.
- [6] M.-S. Yang, P.-Y. Hwang and D.-H. Chen, "Fuzzy clustering algorithm for mixed feature variables," *Fuzzy Sets and Systems*, vol. 141, no. 2, pp.301-317, 2004.
- [7] F. W. Young, J. de Leeuw and Y. Takane, "Multiple and canonical regression with a mix of quantitative and qualitative variables: an alternating least squares method with optimal scaling features," *Psychometrika*, vol. 41, pp.505-530, 1976.
- [8] Y. Takane, F. W. Young and J. de Leeuw, "Nonmetric

- individual differences multidimensional scaling: an alternating least squares method with optimal scaling features," *Psychometrika*, vol. 42, pp.7-67, 1977.
- [9] F. W. Young, Y. Takane and J. de Leeuw, "Principal components of mixed measurement level multivariate data: an alternating least squares method with optimal scaling features," *Psychometrika*, vol. 43, pp.279-281, 1978.
- [10] 本多克宏, 上杉亮, 市橋秀友: 混合データベースのFCMクラスターリング, 日本知能情報ファジィ学会誌, vol. 18, no. 4, pp.598-608, 2006.
- [11] 本多克宏, 杉浦伸和, 市橋秀友, 荒木昭一, 久津見洋: 最小2乗基準を用いたFuzzy c -Varieties法における欠測値の処理法, 日本ファジィ学会誌, vol. 13, no. 6, pp.680-688, 2001.
- [12] 本多克宏, 杉浦伸和, 市橋秀友: ロバストな線形ファジィクラスターリング法の提案と協調フィルタリングシステムへの応用, システム制御情報学会論文誌, vol. 16, no. 11, pp.597-605, 2003.
- [13] S. van Buuren and W. J. Heiser, "Clustering N objects into K groups under optimal scaling of variables," *Psychometrika*, vol. 54, no. 4, pp.699-706, 1989.
- [14] R. N. Davé, "An adaptive fuzzy c -elliptotype clustering algorithm," *Proc. of the North American Fuzzy Information Processing Society: Quater Century of Fuzziness*, vol. 1, pp.9-12, 1990.
- [15] 本多克宏, 山川あす香, 市橋秀友, 三好哲也, 奥山哲史: ファジィクラスターリングと回帰と主成分の同時分析法, システム制御情報学会論文誌, vol. 13, no. 5, pp.236-243, 2000.
- [16] K. Honda, Y. Nakamura and H. Ichihashi, "Simultaneous application of fuzzy clustering and quantification with incomplete categorical data," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 8, no. 4, pp.397-402, 2004.
- (2006年12月4日 受付)
(2007年4月24日 採録)
- [問い合わせ先]
〒599-8531 大阪府堺市中区学園町1-1
大阪府立大学大学院工学研究科
電気・情報系専攻知能情報工学分野
本多 克宏
TEL: 072-254-9355
FAX: 072-254-9915
E-mail: honda@cs.osakafu-u.ac.jp

著者紹介

うえすぎ りょう
上杉 亮 [非会員]

2007年大阪府立大学大学院工学研究科博士前期課程電気・情報系専攻修了。同年シャープ(株)入社。在学中はファジィクラスターリング, ファジィモデリングの研究に従事。

ほんだ かつひろ
本多 克宏 [正会員]

1999年大阪府立大学大学院工学研究科博士前期課程電気・情報系専攻修了。同年日本電信電話(株)入社, 同年大阪府立大学工学部経営工学科助手, 2000年同大学院工学研究科助手, 2007年同助教, 現在に至る。博士(工学)。ファジィクラスターリングによるデータマイニングやニューラルネットワークの研究に従事。IEEE, 日本知能情報ファジィ学会, システム制御情報学会, 日本経営工学会の会員。

いちはし ひでとも
市橋 秀友 [正会員]

1971年大阪府立大学工学部経営工学科卒業。同年松下電器産業(株)入社, 1981年大阪府立大学工学部経営工学科助手, 1987年同講師, 1989年同助教授, 1993年同教授, 2000年同大学院工学研究科教授, 現在に至る。工学博士。ファジィクラスターリングやニューラルネットワークなどでのデータ解析法, その知的システムや人間情報システムへの応用研究に従事。IEEE, 日本知能情報ファジィ学会, システム制御情報学会, 電子情報通信学会, 日本経営工学会などの会員。

のづ あきら
野津 亮 [非会員]

2005年3月京都大学大学院情報科学研究科システム科学専攻博士後期課程修了。同年4月大阪府立大学大学院工学研究科助手, 2007年同助教, 現在に至る。博士(情報学)。認知モデル, マルチエージェントシステムの研究に従事。計測自動制御学会, ヒューマンインタフェース学会の会員。

Local Principal Component Analysis for Mixed Databases Based on Linear Fuzzy Clustering

by

Ryo UESUGI, Katsuhiko HONDA, Hidetomo ICHIHASHI and Akira NOTSU

Abstract :

Fuzzy c -Varieties (FCV) is a tool for linear fuzzy clustering and is also applicable to local principal component analysis (local PCA), in which each low-dimensional subspace is estimated considering data partition. Although the clustering criterion in FCV is distances between data points and prototypical linear varieties, the criterion can also be defined based on least square approximation. Optimal scaling is a useful approach to multivariate analysis for mixed databases and has been applied to linear model estimation. This paper proposes two formulations of local PCA for mixed databases based on optimal scaling, in which a conventional FCV and linear fuzzy clustering using least square approximation are enhanced. The proposed algorithms include a step of calculating numerical scores of categorical variables in addition to the ordinary alternative optimization.

Keywords : Cluster analysis, Principal component analysis, Categorical data, Optimal scaling

Contact Address : **Katsuhiko HONDA**

Graduate School of Engineering, Electrical Engineering and Information Science, Osaka Prefecture University

1-1 Gakuen-cho, Naka-ku, Sakai, Osaka 599-8531, JAPAN

TEL : 072-254-9355

FAX : 072-254-9915

E-mail : honda@cs.osakafu-u.ac.jp