論文 適応型ファジィクラスタリングによる分類と 回帰の同時分析[†]

領家 美奈^{*1} 中森 義輝^{*2}

分類と回帰を同時に分析するファジィクラスタリングについて考察する。与えられたデータ集合から高木・菅野のファジィモデルを同定する際、前件部同定のためのデータのファジィ分割と後件部同定のための回帰分析を調和的に実行する問題である。Hathaway and Bezdek の Fuzzy c-Regression Models は、与えられたデータ集合に対していくつかの回帰モデルをあてはめ、クラスタリングと回帰係数の同定を同時に行うアプローチである。本稿では、そのアプローチをファジィモデリングに利用するために、規範の改良と適応型アルゴリズムを提案する。また、数値実験により提案した手法の有効性を示すとともに、世界の人口問題に関連するファジィモデルの構築という実際問題へ応用する。

キーワード：ファジィ分割、回帰分析、ファジィクラスタリング、ファジィモデリング

1. はじめに

分類と回帰の同時分析のためのファジィクラスタリングについて、高木・菅野のファジィモデル[1]構築という文脈において考察する。ファジィクラスタリングはデータの分類のみならず、特徴抽出に利用する目的で発展してきた。たとえば、Fuzzy c-Means(FCM)[2][3]の発展形である Fuzzy c-Varieties(FCV)[4]、Adaptive Fuzzy c-Elliptotype(AFC)[5]、Fuzzy c-Regression Models(FCRM)[6][7]はクラスタの形状を操作しようとする試みである。

これらの方法が発展してきたのは、アルゴリズムの簡便さによるところが大きい。しかし、筆者らの経験では、これらの手法はファジィモデル同定のためにそれほど有効ではない。その理由は、条件変数空間の分割が都合よく実行されないという点、さらには、データに線形構造部分と非線形構造部分が混在しているとき、それらを捉えるデータ分割が困難であるという点である。本稿では、FCRM を改良することによって比較的簡単にファジィモデルを構築する手法を提案する。改良点は FCRM の規範の適切化とアルゴリズムの適応化である。

Fuzzy c-Regression Models(FCRM)は分類と回帰の同時分析のための提案である。次節において、FCRM の問題点を特にファジィモデリングに利用する立場から検討する。次に、それらの問題点をふまえて、分類と回帰という2つの目的を同時分析する新し

いファジィクラスタリング法を提案し、その有効性を数値実験により明らかにする。さらに、世界の人口問題に関連するファジィモデルを構築し、応用可能性について検討する。

2. ファジィクラスタリング

ファジィクラスタリングは初期設定の難しさを問題点として持っているが、比較的容易にデータのカバーリングを行うことができることから、ファジィモデリングに利用できることが期待される。ただし、多次元数値データのクラスタリング結果を用いてファジィモデルを構築するためには、以下の2つの要求が満たされる必要がある。

- ・前件部同定のために、条件変数空間においてデータがうまく分類されること。
- ・後件部同定のために、部分線形構造をクラスタとして捉えること。

すなわち、分類と回帰の同時分析が実行できることが要求される。

本節では、分類を目的とする代表的な手法である Fuzzy c-Means[2][3]と、回帰を主目的とする代表的な手法である Fuzzy c-Regression Models[6][7]について数値実験により問題点を検討する。

2.1 Fuzzy c-Means

データのファジィ分割行列を $U \in \mathbf{R}^{n \times c}$ とする。ただし、 n はデータ数、 c はクラスタ数である。 $u_{ik} \in U$ をデータベクトル $\mathbf{x}_i \in \mathbf{R}^s$ のクラスタ k へのメンバシップ値で、以下の条件を満たすものとする。

$$0 \leq u_{ik} \leq 1, \quad \forall i, k \quad (1)$$

[†] Simultaneous Analysis of Classification and Regression by Adaptive Fuzzy Clustering

Mina RYOKE and Yoshiteru NAKAMORI

^{*1} 大阪大学大学院基礎工学研究科

Faculty of Engineering Science, Osaka University

^{*2} 甲南大学理学部応用数学科

Department of Applied Mathematics, Konan University

$$0 < \sum_{i=1}^n u_{ik} < n, \quad \forall k \quad (2)$$

$$\sum_{k=1}^c u_{ik} = 1, \quad \forall i \quad (3)$$

Fuzzy c-Means (FCM) の代表的な規範は、クラス k の中心ベクトルを $v_k \in R_s$ として

$$J(U, v_1, \dots, v_c) = \sum_{i=1}^n \sum_{k=1}^c (u_{ik})^m \|x_i - v_k\|_{R_s}^2 \quad (4)$$

と定義される。ここに、 $m(>1)$ はスミージング・パラメータである。

分割行列 U と c 個のクラスタ中心ベクトル $v_1, \dots, v_c$ を同時に最適化することは困難であるから、FCM のアルゴリズムでは、 U を固定して J を最小化する $v_1, \dots, v_c$ を求め、次に $v_1, \dots, v_c$ を固定して J を最小化する U を求める。これを U が収束するまで繰り返す。

文献[5]のデータを用いて数値実験を実行した結果を図1に示す。 $c=6$, $m=1.2$ とし、分割行列 U の初期値は乱数で与えた。メンバシップ値の最大変化が 1.0×10^{-5} 以下になったとき (41 ステップ目) クラスタリングを終了した。なお、図を見やすくするために 0.6 で α -カットを実行している。FCM は説明変数と目的変数を分離しないので、ファジィモデリングに直接利用できないことは自明である。

2.2 Fuzzy c-Regression Models

前件部のデータ分割と後件部の回帰係数を同時に決定し、簡単にファジィモデルの第一近似を得る方法として、Hathaway と Bezdek [6] [7] のアプローチに注目する。

$y \in R^t$ を目的変数ベクトル、 $x \in R^s$ を説明変数ベクトルとし、与えられたデータセット $\{(x_1, y_1), \dots, (x_n, y_n)\}$ を用いて、各目的変数に対して c 個のスイッチング回帰モデル [8] :

$$y = f_k(x; \beta_k) + \varepsilon_k, \quad k=1, 2, \dots, c \quad (5)$$

を同定する。ここに、 $\beta_k \in R^{t \times (s+1)}$ (線形モデルの場合) は回帰パラメータで、 $\varepsilon_k \in R^t$ は残差ベクトルである。

$f_k(x_i; \beta_k)$ による y_i の近似誤差の尺度を

$$E_{ik}(\beta_k) = \|y_i - f_k(x_i; \beta_k)\|_{R^t}^2 \quad (6)$$

と定義し、次の評価規範を導入する。

$$J(U, \beta_1, \dots, \beta_c) = \sum_{i=1}^n \sum_{k=1}^c (u_{ik})^m E_{ik}(\beta_k) \quad (7)$$

FCRM では、 u_{ik} はモデルの値 $f_k(x_i; \beta_k)$ の y_i への適合度と解釈できる。

モデルの同定は、 U を固定し J を最小化する $\beta_1, \dots, \beta_c$ の探索と、逆に $\beta_1, \dots, \beta_c$ を固定し J を最小

図1 FCMによるクラスタリング結果

化する U の探索を繰り返すアルゴリズムによる。図2は同じく文献[5]のデータ ($t=s=1$) についての数値実験結果 (0.77 で α -カット) である。線形回帰モデルを仮定し、FCM の数値実験の場合と同じパラメータを設定した。この手法では 37 ステップで収束した。

得られた回帰直線を以下に示す。図2の a から f に対応している。

- a : $y = -0.100 - 0.074x$
- b : $y = 2.764 - 0.160x$
- c : $y = -1.092 + 0.029x$
- d : $y = -3.527 + 1.394x$
- e : $y = 1.862 + 0.091x$
- f : $y = 0.687 - 0.010x$

次式で定義する各回帰式の重み付き残差

$$\phi_k = \sum_{i=1}^n u_{ik} E_{ik}(\beta_k) \quad (8)$$

の変化の様子を図3に示す。

図2 FCRMによるクラスタリング結果

図3 FCRMによる回帰式の重み付き残差

FCRM は初期値をいかに設定するかという問題点を持っているが、得られる回帰モデルの精度は非常に高い。しかし、残差のみを考慮しているためデータの分割能力に欠ける。すなわち、各々の線形モデルは良い精度を持つが、各モデルを使用するためのルールをうまく作成できるとは限らない。したがって、FCRM の考え方をファジィモデリングに利用するためには少し工夫を要する。

3. 適応型ファジィクラスタリング

FCRM は目的変数を意識したクラスタリング法であり、ファジィモデリングにおいて、前件部変数を後件部変数から選択するときに利用できる。ここでは、その利点を活かしつつ、変数空間の高い分割能力を持つファジィクラスタリング法 Adaptive Fuzzy c-Regression Models(AFCR)を提案する。

3.1 クラスタ重心とデータとの距離

ファジィモデルにおける後件部説明変数集合を $\{x_1, x_2, \dots, x_s\}$ 、目的変数集合を $\{y_1, y_2, \dots, y_t\}$ とする。これらの変数に対する i 番目のデータを $x_{1i}, \dots, x_{si}, y_{1i}, \dots, y_{ti} (i=1, 2, \dots, n)$ とし、データベクトルを

$$x_i = (x_{1i}, \dots, x_{si})^T, \quad i=1, 2, \dots, n \quad (9)$$

$$y_i = (y_{1i}, \dots, y_{ti})^T, \quad i=1, 2, \dots, n \quad (10)$$

$$d_i = (x_i^T, y_i^T)^T, \quad i=1, 2, \dots, n \quad (11)$$

とおく。クラスタ k のファジィ重心ベクトルを次のように定義する。

$$v_k = \frac{\sum_{i=1}^n (u_{ik})^m d_i}{\sum_{i=1}^n (u_{ik})^m}, \quad k=1, 2, \dots, c \quad (12)$$

また、 d_i と v_k の距離を次のように定義する。

$$D_{ik} = (d_i - v_k)^T A (d_i - v_k) \quad (13)$$

ただし、 $A = (a_{ij})$ は $s+t$ 次元対角行列である。

行列 A については、距離を考慮する変数に応じて以下のような3つのパラエティを考える。

- ・タイプ1：(全変数について考える)
 $a_{ii}=1$ for all $i \in \{1, \dots, s+t\}$
- ・タイプ2：(全説明変数について考える)
 $a_{ii}=1$ for all $i \in \{1, \dots, s\}$,
 $a_{ii}=0$ for all $i \in \{s+1, \dots, s+t\}$
- ・タイプ3：(いくつかの説明変数について考える)
 $a_{ii}=1$ for some $i \in \{1, \dots, s\}$,
 $a_{ii}=0$ for all $i \in \{s+1, \dots, s+t\}$

タイプ1の距離を用いて満足できるファジィモデルが構築できない場合、タイプ2またはタイプ3の距離により検討を行う。本稿における応用例ではタイプ1で十分であった。

ところで、クラスタリングに先だって、すべての変数のデータを平均0、分散1に標準化する。ただし、線形回帰モデルの最終的な係数は元のスケールに逆変換する。なお、ファジィモデリングの目的は目的変数の説明の仕方を変えた方がよいと思われるいくつかの部分を見出すことであるから、目的変数のデータのスケールを説明変数のデータのスケールより大きくとってみることを試みてもよい。

3.2 2つの目的のバランスをとる規範

AFCR は次の規範を採用する。

$$J(U, \beta_1, \dots, \beta_c, v_1, \dots, v_c) = \sum_{i=1}^n \sum_{k=1}^c (u_{ik})^m Z_{ik} \quad (14)$$

$$Z_{ik} = \alpha_k E_{ik}(\beta_k) + (1 - \alpha_k) \eta D_{ik} \quad (15)$$

右辺第1項の $E_{ik}(\beta_k)$ はFCRMの規範と同じく、残差2乗和である。第2項の役割は変数空間の分割能力を高めることである。

本研究ではモデルベクトル $f_k(x; \beta_k)$ の各成分 f_{kj} (目的変数 y_j に対応) を $\beta_{kj0}, \dots, \beta_{kjs}$ の線形関数とする。すなわち、

$$f_{kj}(x; \beta_k) = \beta_{kj0} + \beta_{kj1}x_1 + \dots + \beta_{kjs}x_s \quad (16)$$

このとき、回帰係数ベクトル β_{kj} の推定量は次のように求めることができる。

$$\beta_{kj} = [X^T \Gamma_k X]^{-1} X^T \Gamma_k y_j \quad (17)$$

ここに、 $X \in \mathbf{R}^{n \times (s+1)}$ はデータベクトル x_i を第 $(i+1)$ 行に持つ行列(ただし、第1列は全て1.0)、 $y_j \in \mathbf{R}^n$ はデータ y_{ji} を第 i 要素とする列ベクトル、 $\Gamma_k \in \mathbf{R}^{n \times n}$ は $(u_{ik})^m$ を第 (i, i) 要素に持つ対角行列である。また、 $\beta_{kj} \in \mathbf{R}^{s+1}$ は β_{kji} を第 $(i+1)$ 要素に持つ列ベクトルである。

クラスタごとにファジィ散布行列[5]

$$F_k = \sum_{i=1}^n (u_{ik})^m (d_i - v_k) (d_i - v_k)^T \quad (18)$$

を求め、その固有値を $\lambda_{kl} (l=1, 2, \dots, s+t)$ とする。 α_k は次のように定める。

$$\alpha_k = 1 - \frac{\min_l \{\lambda_{kl}\}}{\max_l \{\lambda_{kl}\}}, \quad k=1, 2, \dots, c \quad (19)$$

α_k が 1 に近いとき、クラス k はその中で線形構造を発見している。その結果、 Z_{ik} において第 1 項の残差項が重視されることになる。逆に、 α_k が 0 に近いとき、クラス k 内では線形構造は発見されていない。その場合、データ間距離を重視してクラスタリングが実行されることになる。これは Dave[5] のアイデアである。

α_k はクラス k の形状に基づいて残差と距離のバランスをとるパラメータである。しかし、規範の第 1 項と第 2 項の平均的な大きさに無視できない差があるとき、クラスタリング結果はどちらかに引きずられたものとなる。そこで、第 1 項と第 2 項の平均的な大きさのバランスをとるパラメータ η を導入している。ただし、本研究では η の最適化を考えていない。その適切な値はデータ分布によって大きく異なると考えられるため、設定値として与え、試行を繰り返すアプローチをとる。

3.3 適応型アルゴリズム

ステップ 1: クラスタ数 c 、スムージング・パラメータ m 、終了条件パラメータ $\varepsilon (> 0)$ を設定する。また、距離のタイプ (行列 A) を決定し、残差項と距離項のバランスをとるパラメータ η 、データの初期分割行列 $U^{(0)}$ を設定する。 $r=0$ とおく。

ステップ 2: ファジィ重心ベクトルを計算する。

$$v_k^{(r)} = \frac{\sum_{i=1}^n (u_{ik}^{(r)})^m d_i}{\sum_{i=1}^n (u_{ik}^{(r)})^m}, \quad k=1, 2, \dots, c \quad (20)$$

ステップ 3: データベクトル d_i とファジィ重心ベクトル $v_k^{(r)}$ の距離を計算する。

$$D_{ik}^{(r)} = (d_i - v_k^{(r)})^T A (d_i - v_k^{(r)}) \quad (21)$$

ステップ 4: ファジィ散布行列 $F_k^{(r)}$ の固有値 $\lambda_{kl}^{(r)}$, $\dots$, $\lambda_{k(s+t)}^{(r)}$ を求め、残差項と距離項のトレードオフ値 $\alpha_k^{(r)}$ を計算する。

$$\alpha_k^{(r)} = 1 - \frac{\min_l \{\lambda_{kl}^{(r)}\}}{\max_l \{\lambda_{kl}^{(r)}\}}, \quad k=1, 2, \dots, c \quad (22)$$

ステップ 5: $\beta_1, \dots, \beta_c$ に関する規範

$$\Psi(\beta_1, \dots, \beta_c) = J(U^{(r)}, \beta_1, \dots, \beta_c, v_1^{(r)}, \dots, v_c^{(r)}) \quad (23)$$

を最小化する回帰モデルの係数 $\beta_1^{(r)}, \dots, \beta_c^{(r)}$ を求める。

$$\beta_{kj}^{(r)} = [X^T \Gamma_k X]^{-1} X^T \Gamma_k \gamma_j \quad (24)$$

ステップ 6: U に関する規範

$$\Phi(U) = J(U, \beta_1^{(r)}, \dots, \beta_c^{(r)}, v_1^{(r)}, \dots, v_c^{(r)}) \quad (25)$$

を最小化するように分割行列を更新する： $U^{(r)} \rightarrow U^{(r+1)}$ 。

$$u_{ik}^{(r+1)} = \left[\sum_{h=1}^c \left(\frac{Z_{ik}^{(r)}}{Z_{ih}^{(r)}} \right)^{\frac{1}{m-1}} \right]^{-1} \quad (26)$$

ステップ 7: 終了条件

$$\max_{i,k} \{ |u_{ik}^{(r+1)} - u_{ik}^{(r)}| \} < \varepsilon \quad (27)$$

を満たせば終了。満たさなければ $r=r+1$ として、ステップ 2 へ戻る。

3.4 AFMR による数値実験

AFMR による数値実験結果 (0.83 で α -カット) を図 4 に示す。距離はタイプ 1 を用いている。FCM, FCRM の場合と同じ初期設定から始めて 19 ステップで収束した。なお、 $\eta=1.0$ とした。

同時に得られた回帰直線を以下に示す。図中の A から F に対応している。

$$A: y = 0.876 - 0.473x$$

$$B: y = 2.053 + 0.240x$$

$$C: y = -1.116 + 0.192x$$

$$D: y = -3.238 + 1.089x$$

$$E: y = -0.079 + 0.373x$$

$$F: y = 0.296 - 0.030x$$

各回帰式の重み付き残差

$$\phi_k = \sum_{i=1}^n u_{ik} E_{ik}(\beta_k) \quad (28)$$

の変化の様子を図 5 に示す。また、トレードオフ値 α_k の収束の様子を図 6 に示す。線形構造を捉えているク

図 4 AFMR によるクラスタリング結果

図5 AFMRによる回帰式の重み付き残差

ラスト A と E については規範の中の残差項が重視されていることがわかる。

以上のように、AFMR は線形構造とデータが密集している部分が混在していても、それらをうまく捉えることができる。

4. 応用例：人口モデル

提案した AFMR を利用して、人口変動現象を経済的・社会的要因から説明するファジィモデルを構築する。国連の人口部は、平均寿命と死亡率を用いて生存率を算出し、その生存率を基にして世界の死亡パターンを最大公約数的に示している[9]。しかし、その背景の要因を同時に説明しているものではない。

ここでは、データブック[10]のデータ(ソビエト連邦崩壊前)を用いて、男性平均寿命(歳：1980年～1985年)と死亡率(人/千人：1980年～1985年)を説明するファジィモデルを構築する。説明変数は、まず経済的要因として、1人あたりのGNP(US\$/人：1988年)、母親の知識により目的変数は影響を受けると考えられるので、識字率(%：1990年)、多産多死から少産少死の過程の位置を知るために出生率(人/千人：1980年～1985年)、最後に、人口構成を考えるため、従属人口比($\{14$ 歳以下と65歳以上の人口 $\} \div \{15$ 歳以上から64歳以下の人口 $\}$ ：1988年)を選択した。すなわち、 $t=2$, $s=4$ とした。データ数は87カ国分($n=87$)である。

予備的な検討として、目的変数それぞれに対する全域線形回帰モデルを構築した。しかし、特に出生率についての自由度調整済み決定係数が0.3482と小さくなった。出生率に対する死亡率をプロットした図7(標準化データ)を見れば、部分的な関係が存在することが理解できる。そこで、AFMRによりデータの分類と回帰モデル構築を実行する。距離はタイプ1とした。

いくつかの2変数散布図を検討し、クラスタ数を3

図6 トレードオフ値の収束の様子

図7 死亡率と出生率

とした。パラメータは $m=2$, $\eta=10.0$, $\epsilon=1.0 \times 10^{-5}$ とした。クラスタリングは標準化データを用いて行った。初期分割行列は乱数で与え、157ステップで収束した。クラスタリング結果を表1から表3に示す。表中の数字は各国の各クラスタへのメンバシップ値を表す。

表1 クラスタ1の国家とメンバシップ値

国名	クラスタ1	クラスタ2	クラスタ3
フランス	0.939	0.033	0.028
ニュージーランド	0.933	0.045	0.022
ベルギー	0.931	0.035	0.033
イタリア	0.924	0.043	0.033
オーストリア	0.914	0.043	0.042
アイルランド	0.910	0.050	0.040
イギリス	0.903	0.049	0.048
ブルガリア	0.894	0.056	0.050
アラブ首長国連邦	0.894	0.078	0.028
デンマーク	0.890	0.055	0.055
スウェーデン	0.884	0.059	0.058
チェコスロバキア	0.879	0.062	0.059
ノルウェイ	0.878	0.064	0.059
フィンランド	0.872	0.070	0.058
オランダ	0.868	0.084	0.049

ルーマニア	0.854	0.081	0.064
ポルトガル	0.846	0.088	0.065
アメリカ合衆国	0.843	0.097	0.060
西ドイツ	0.835	0.084	0.081
ポーランド	0.825	0.097	0.077
ユーゴスラビア	0.823	0.113	0.065
カナダ	0.803	0.137	0.059
ハンガリー	0.786	0.111	0.103
ギリシャ	0.755	0.148	0.097
オーストラリア	0.750	0.174	0.076
東ドイツ	0.736	0.129	0.135
スイス	0.710	0.166	0.124
コスタリカ	0.695	0.233	0.072
トリニダードトバゴ	0.688	0.259	0.054
マレーシア	0.682	0.258	0.060
スペイン	0.646	0.226	0.128
アイスランド	0.639	0.242	0.119
アルゼンチン	0.595	0.285	0.120
シンガポール	0.587	0.307	0.107
日本	0.570	0.292	0.138
イスラエル	0.563	0.344	0.093
ソビエト連邦	0.550	0.252	0.198
パナマ	0.542	0.383	0.075
キューバ	0.500	0.361	0.139
韓国	0.478	0.410	0.112
中華人民共和国	0.458	0.438	0.104

表2 クラスタ2の国家とメンバシップ値

国名	クラスタ1	クラスタ2	クラスタ3
ニカラグア	0.030	0.930	0.040
リビア	0.027	0.926	0.047
トルコ	0.050	0.909	0.041
イラン	0.028	0.907	0.065
エクアドル	0.049	0.907	0.044
エジプト	0.051	0.901	0.048
ブラジル	0.068	0.898	0.034
イラク	0.060	0.894	0.046
ドミニカ王国	0.074	0.893	0.033
サウジアラビア	0.064	0.890	0.046
アルジェリア	0.038	0.889	0.073
ペルー	0.042	0.846	0.112
コロンビア	0.117	0.840	0.043
パキスタン	0.090	0.834	0.077
ホンジュラス	0.055	0.814	0.131
パラグアイ	0.153	0.798	0.049
ハイチ	0.044	0.776	0.180
フィリピン	0.111	0.766	0.123
メキシコ	0.180	0.766	0.054
グアテマラ	0.100	0.739	0.161
シリア	0.175	0.733	0.092
インド	0.056	0.732	0.212
タイ	0.158	0.712	0.131
ジャマイカ	0.201	0.691	0.108
ヨルダン	0.234	0.669	0.096
ベネズエラ	0.328	0.589	0.083
ガイアナ	0.424	0.505	0.071
チリ	0.334	0.500	0.166
スリランカ	0.450	0.466	0.084

表3 クラスタ3の国家とメンバシップ値

国名	クラスタ1	クラスタ2	クラスタ3
マダガスカル	0.003	0.011	0.986
モザンビーク	0.005	0.024	0.971
ボリビア	0.010	0.048	0.942
ナイジェリア	0.010	0.051	0.940
コンゴ	0.016	0.058	0.927
アフガニスタン	0.011	0.070	0.918
エチオピア	0.017	0.065	0.918
中央アフリカ	0.019	0.066	0.915
ネパール	0.017	0.087	0.897
スーダン	0.016	0.097	0.887
ソマリア	0.058	0.150	0.792
リベリア	0.031	0.240	0.729
ケニヤ	0.053	0.292	0.655
インドネシア	0.062	0.289	0.648
ミャンマー	0.108	0.340	0.552
モロッコ	0.077	0.460	0.463
エルサルバドル	0.255	0.360	0.385

クラスタ1に先進資本主義国と(旧)社会主義国が含まれている。クラスタ2, クラスタ3には発展途上国が含まれているが, クラスタ2には石油輸出国機構に加盟している国々の多くが含まれている。このように途上国の多様性がクラスタリングに反映された結果となった。以下において, クラスタリング結果をいくつかの2次元射影平面上で示す。なお, α -カットを0.38で実行している。

図8から図10は男性平均寿命を縦軸に, それぞれ出生率, 乳児死亡率(人/出生児千人: 1980年~1985年); 1人あたりのGNPを横軸にとったものである。軸の数値は平均(0)±標準偏差である。これらの図から, 以下のような常識的なことがらが確認される。出生率が高いほど平均寿命が低い。クラスタ3からクラスタ1へ向けて, 多産多死から少産少死の状態へ移行している。また, 経済状態に対して平均寿命の伸びが頭打ちになっている。

図11から図14は死亡率を縦軸に, それぞれ出生率, 医師数(人/万人: 1988年), 乳児死亡率, 従属人口比を横軸にとったものである。一般に, 医師数が多いほど死亡率は低くなると考えられるが, クラスタ2と3は同程度の医師数に対して死亡率が大きく異なる。クラスタ3の国々は多産多死の人口構成であり, 従属人口比も大きい。

クラスタリングと同時に得られた各目的変数に対する回帰式を以下に整理する。ただし, 標準化データによる回帰モデルの係数を, 平均と分散を用いて元のスケールに変換している。

図8 男性平均寿命と出生率

図11 死亡率と出生率

図9 男性平均寿命と乳児死亡率

図12 死亡率と医師数

図10 男性平均寿命と1人あたりのGNP

図13 死亡率と乳児死亡率

図14 死亡率と従属人口比

・ クラスタ 1

$$\begin{aligned} \text{平均寿命} = & 5.867 \times 10^{+1} \\ & -1.328 \times 10^{-1} \times \text{出生率} \\ & +2.376 \times 10^{-4} \times \text{GNP} \\ & +8.246 \times 10^{+0 \pm 0} \times \text{従属人口} \\ & +6.688 \times 10^{-2} \times \text{識字率} \end{aligned}$$

$$\begin{aligned} \text{死亡率} = & 1.042 \times 10^{+1} \\ & -3.807 \times 10^{-1} \times \text{出生率} \\ & -4.677 \times 10^{-5} \times \text{GNP} \\ & +1.107 \times 10^{+1} \times \text{従属人口比} \\ & -4.960 \times 10^{-3} \times \text{識字率} \end{aligned}$$

・ クラスタ 2

$$\begin{aligned} \text{平均寿命} = & 5.963 \times 10^{+1} \\ & -9.401 \times 10^{-1} \times \text{出生率} \\ & +2.547 \times 10^{-4} \times \text{GNP} \\ & +3.907 \times 10^{+1} \times \text{従属人口比} \\ & +3.894 \times 10^{-2} \times \text{識字率} \end{aligned}$$

$$\begin{aligned} \text{死亡率} = & 1.230 \times 10^{+1} \\ & +3.747 \times 10^{-1} \times \text{出生率} \\ & +9.898 \times 10^{-5} \times \text{GNP} \\ & -1.761 \times 10^{+1} \times \text{従属人口比} \\ & -4.500 \times 10^{-2} \times \text{識字率} \end{aligned}$$

・ クラスタ 3

$$\begin{aligned} \text{平均寿命} = & 5.675 \times 10^{+1} \\ & -1.056 \times 10^{+0} \times \text{出生率} \\ & +2.249 \times 10^{-1} \times \text{GNP} \\ & +3.777 \times 10^{+1} \times \text{従属人口比} \\ & +7.498 \times 10^{-2} \times \text{識字率} \end{aligned}$$

$$\begin{aligned} \text{死亡率} = & 1.266 \times 10^{+1} \\ & +6.350 \times 10^{-1} \times \text{出生率} \\ & +2.113 \times 10^{-4} \times \text{GNP} \\ & -2.334 \times 10^{+1} \times \text{従属人口比} \\ & -5.325 \times 10^{-2} \times \text{識字率} \end{aligned}$$

ファジィモデルの前件部同定は文献[11]の方法による。前件部変数は、人口密度(人/km²:1988年),人口増加率(%:1980年~1988年),出生率,乳児死亡率,1人あたりのGNP,従属人口比,識字率(%:1990年),医師数,および栄養摂取量(カロリー/人日:1986年)から選択した。その結果,乳児死亡率,1人あたりのGNP,医師数,および栄養摂取量が選ばれた。メンバシップ関数は図15から図18のように同定された。

図15 乳児死亡率のメンバシップ関数

図16 1人あたりのGNPのメンバシップ関数

図17 医師数のメンバシップ関数

図18 栄養摂取量のメンバシップ関数

モデルの精度は、平均寿命についてはデータと推定値の相関係数が0.9524、平均2乗残差が2.853、死亡率についてはデータと推定値の相関係数が0.8906、平均2乗残差が1.991であった。なお、表4から表6はデータと推定値の対応表である。

表4 クラスタ1の国家と推定値

国名	平均寿命(歳)		死亡率(人/千人)	
	データ	予測値	データ	予測値
フランス	71.0	71.5	9.9	9.6
ニュージーランド	71.0	69.6	8.3	8.9
ベルギー	70.0	71.2	11.1	10.2
イタリア	71.1	70.9	9.5	10.7
オーストリア	70.4	71.5	11.4	10.3
アイルランド	70.1	70.0	9.0	10.0
イギリス	70.4	70.9	11.8	10.1
ブルガリア	68.4	68.8	11.4	10.0
アラブ首長国連邦	65.4	66.1	4.3	3.4
デンマーク	71.6	72.3	11.4	10.4
スウェーデン	73.8	72.7	11.1	10.4
チェコスロバキア	67.1	69.2	11.8	10.2
ノルウェー	72.8	73.0	9.8	10.3
フィンランド	70.1	72.0	9.6	9.6
オランダ	72.9	70.8	8.6	9.4
ルーマニア	67.0	68.0	10.3	9.6
ポルトガル	68.4	67.8	9.6	10.9
アメリカ合衆国	71.2	72.1	8.7	8.7
西ドイツ	71.2	71.9	11.5	9.9
ポーランド	66.5	67.9	10.0	9.4
ユーゴスラビア	67.7	67.2	9.1	9.3
カナダ	71.9	71.2	7.3	8.7
ハンガリー	65.6	68.3	13.9	10.7
ギリシャ	72.2	68.5	9.2	11.1
オーストラリア	72.3	70.4	7.3	9.2
東ドイツ	69.5	70.0	13.4	9.8
スイス	73.5	74.0	9.2	9.3
コスタリカ	70.5	66.7	4.1	5.4
トリニダードトバゴ	66.9	67.1	6.6	5.5
マレーシア	67.6	66.0	5.3	5.5
スペイン	72.6	69.6	7.8	10.6
アイスランド	74.0	72.6	6.8	9.2
アルゼンチン	65.5	67.7	8.0	8.2
シンガポール	68.7	67.9	5.0	8.3
日本	74.8	72.4	6.2	9.4
イスラエル	73.1	69.4	6.8	8.3
ソビエト連邦	62.9	68.2	9.7	8.1
パナマ	69.2	66.9	5.4	6.7
キューバ	72.7	67.4	6.1	8.7
韓国	67.2	67.0	6.1	7.7
中華人民共和国	66.7	65.0	6.7	8.4

表5 クラスタ2の国家と推定値

国名	平均寿命(歳)		死亡率(人/千人)	
	データ	予測値	データ	予測値
ニカラグア	58.7	58.0	9.7	9.2
リビア	56.6	57.2	10.9	10.7
トルコ	60.0	59.7	9.4	9.2
イラン	55.8	56.4	11.5	10.7
エクアドル	59.5	60.5	8.1	7.9
エジプト	56.8	58.0	9.1	10.0
ブラジル	60.9	60.6	8.4	8.8
イラク	61.5	59.1	8.7	8.9
ドミニカ王国	60.7	59.5	8.0	9.0
サウジアラビア	59.2	59.6	8.9	10.2
アルジェリア	58.5	58.9	10.7	9.7
ペルー	56.8	58.3	10.7	9.3
コロンビア	61.4	60.9	7.7	8.3
パキスタン	59.0	56.1	9.6	11.6
ホンジュラス	53.4	57.5	10.0	9.8
パラグアイ	62.8	66.0	7.2	5.1
ハイチ	51.2	44.0	14.2	19.8
フィリピン	60.2	58.0	8.4	10.9
メキシコ	62.1	60.3	7.1	8.3
グアテマラ	55.1	60.3	8.7	8.9
シリア	63.8	58.1	8.7	10.0
インド	55.6	57.0	12.6	11.1
タイ	57.6	63.0	8.0	8.1
ジャマイカ	70.3	66.5	5.4	8.8
ヨルダン	64.6	59.1	7.9	9.8
ベネズエラ	64.9	67.5	4.6	6.9
ガイアナ	65.8	65.4	7.3	6.8
チリ	61.3	67.1	6.1	8.1
スリランカ	67.8	65.9	6.2	7.7

表6 クラスタ3の国家と推定値

国名	平均寿命(歳)		死亡率(人/千人)	
	データ	予測値	データ	予測値
マダガスカル	48.9	48.6	16.5	16.6
モザンビーク	44.4	43.8	19.7	19.8
ボリビア	48.6	53.7	18.0	10.8
ナイジェリア	46.9	45.6	17.1	18.5
コンゴ	44.9	48.4	18.6	18.9
アフガニスタン	36.6	37.5	22.3	23.7
エチオピア	39.3	41.9	23.2	20.9
中央アフリカ	41.4	44.5	21.8	19.5
ネパール	50.9	46.3	18.4	18.3
スーダン	46.6	44.6	17.4	19.3
ソマリア	39.3	46.6	23.3	17.9
リベリア	47.4	43.2	17.2	20.2
ケニヤ	51.2	47.7	14.0	14.1
インドネシア	52.2	53.7	12.6	13.7
ミャンマー	55.8	61.5	11.0	8.3
モロッコ	56.6	60.4	11.3	8.4
エルサルバドル	62.6	67.3	6.0	9.2

5. おわりに

Hathaway らの手法を改良して、データのファジィ分割と回帰分析を同時に実行するためのファジィクラスタリング法を提案した。また、数値実験と応用例を用いて、ファジィモデル構築に対して有効であることを確認した。クラスタリング過程において、クラスごとに説明変数を選択すること、さらにはメンバーシップ関数を同時に同定することが今後の課題である。

参 考 文 献

- [1] T.Takagi and M.Sugeno: Fuzzy Identification of Systems and Its Applications to Modeling and Control. *IEEE Trans. on Systems, Man and Cybernetics*, Vol.SMC-15, No.1, pp.116-132, 1985.
- [2] J.Dunn: A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *J. Cybernetics*, Vol.3, pp.32-57, 1974.
- [3] J.C.Bezdek: *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.
- [4] J.C.Bezdek et al.: Detection and Characterization of Cluster Substructure II. Fuzzy c-Varieties and Convex Combinations Thereof. *SIAM J. Appl. Math.*, Vol.40, No.2, pp.358-372, 1981.
- [5] R.N.dave: An Adaptive Fuzzy c-Elliptotype Clustering Algorithm. *Proc. NAFIPS 90: Quater Century of Fuzziness*, Vol.I, pp.9-12, 1990.
- [6] J.C.Bezdek and R.J.Hathaway: Generalized Regression and Clustering. *Proc. of Int. Conf. on Fuzzy Logic and Neural Networks*, pp.575-578, Fukuoka, Japan, July 20-24, 1990.
- [7] R.J.Hathaway and J.C.Bezdek: Switching Regression Models and Fuzzy Clustering. *IEEE Trans. on Fuzzy Systems*, Vol.1, No.3, pp.195-204, 1993.
- [8] D.W.Hosmer, Jr.: Maximum Likelihood Estimates of the Parameters of a Mixture of Two Regression Lines. *Communications in Statistics*, Vol.3, No.10, pp.995-1005, 1974.
- [9] 河野稠果: 世界の人口. 東京大学出版会, 1986.
- [10] 朝日新聞社: 朝日現代用語1991別冊付録「知恵庫」. 1991.
- [11] 中森義輝, 領家美奈: 探索的ファジィモデリング. 日本ファジィ学会誌, 第5巻, 第6号, pp.1323-1336, 1993.

(1994年10月20日 受付)

(1995年3月1日 再受付)

[問い合わせ先]

〒560 大阪府豊中市待兼山町1-3

大阪大学大学院基礎工学研究科

領家 美奈

TEL : 06-850-6378

FAX : 06-850-6341

Email : ryoke@tamlab.sys.es.osaka-u. ac.jp

著 者 紹 介

領家 美奈 (りょうけ みな)

大阪大学大学院基礎工学研究科

1992年3月 甲南大学理学部応用数学科卒業。同年4月より松下電工中央研究所勤務。空気浄化システムの研究開発に従事。1993年4月 甲南大学大学院自然科学研究科修士課程情報・システム科学専攻に入学。1995年3月 修了。同年4月 大阪大学大学院基礎工学研究科物理系専攻に入学し現在に至る。一般化回帰とファジィクラスタリング、ファジィモデリングに興味を持つ。日本ファジィ学会の会員。

中森 義輝 (なかもり よしてる)

甲南大学理学部応用数学科

1979年 京都大学大学院工学研究科数理工学専攻博士課程修了。1981年より甲南大学理学部応用数学科勤務。現在 教授。自然および人間システムの観測と制御に関する研究に従事。最近はシステムモデリング、シミュレーションにおける、対話的かつ知的支援の方法論に興味を持つ。1984年10月より1985年11月まで、国際応用システム解析研究所(オーストリア)研究員として、地域的な水資源の管理・計画問題の研究に従事。1986年4月より、環境庁国立環境研究所客員研究員として、都市環境のモデリングと制御問題の研究に従事。1992年9月より、大連理工大学大学客員教授兼務。日本ファジィ学会、計測自動制御学会、環境科学会、IEEEなどの会員。

Simultaneous Analysis of Classification and Regression by Adaptive Fuzzy Clustering

by

Mina RYOKE and Yoshiteru NAKAMORI

Abstract :

This paper considers the simultaneous analysis of classification and regression in the context of the fuzzy modeling. The fuzzy c-regression models (FCRM) by Hathaway and Bezdek is an extension of the classical fuzzy clustering to obtain a fuzzy partition of the given data and at the same time a number of regression models corresponding to the data partition. However, the direct application of this method to the fuzzy modeling does not necessarily bring a good result. After explaining this fact using a numerical example, we try to revise the FCRM in two aspects : modification of the criterion and adaptation of the algorithm. An application to building of a fuzzy model related to the world population is presented to show how the proposed method works.

Keywords : Fuzzy Partition, Regression Analysis, Fuzzy Clustering, Fuzzy Modeling

Contact Address : **Mina RYOKE**

Faculty of Engineering Science, Osaka University
1-3 Machikaneyama-Cho, toyonaka, Osaka 560, Japan
TEL : 06-850-6378
FAX : 06-850-6341
Email : ryoke@tamlab.sys.es.osaka-u.ac.jp