数理解析研究所講究録
第1589巻 2008年 176-181

Approach of Multi Objective Optimization for Fuzzy c-means Analysis (Fuzzy c-means モデルに対する多目的最適化アプローチ)

金沢学院大学・基礎教育機構　春名　亮 (Ryo Haruna)
Organization of Core Curriculum Studies
金沢学院大学・経営情報学部　桑野裕昭 (Hiroaki Kuwano)
Faculty of Business Adiministration and Information Science
Kanazawa Gakuin University

1　はじめに

齋藤らは文献 [6] で「我々は日常生活あるいは実務や研究などの職業活動において，多種多様な分類 (クラス分け) の作業を，無意識的であれ意識的であれ行っている」と述べており，分類は人間の営む最も根源的な作業の1つと考えられる．従って，非常に古くから広範な領域に分類という作業が用いられている．学術研究領域において，分類に関わる研究が最も早く始まった生物学では，系統分類，数値分類学が行われている．また，統計学においては，判別分析，クラスタリングなどが知られている．そのクラスタリングはクラスター分析とも呼ばれ，データ集合を幾つかのクラスター (群，クラス) に分類することであり，その手法を大別すると，階層的手法と非階層的手法が知られている．

階層的手法とは，与えられたデータ集合の各データが1つのクラスターとなっている状態を初期状態とし，クラスター間の距離や類似度に基づいて，2つのクラスターを逐次的に併合する手法である．予め設定したクラスター数になるまで併合が行われたときに分類処理が終了し，データの階層構造が得られる．よく知られている代表的な例としては，群平均法やウォード法などがある．一方，非階層的手法 (分割最適化手法) とは，クラスター数を予め指定して適当な基準により一度に複数のクラスターを形成する手法であり，k-means 法や混合分布モデルなどが代表的な例としてよく知られている．

また，近年その非階層的手法に対し，人間の柔軟な判断を反映し得るソフトクラスタリング手法の開発が進められている．具体的には，通常のクラスタリング (ハードクラスタリングとも呼ばれることがある) において，各データはいずれか1つのクラスターに必ず属することを強いられる．ところが，糖尿病の疾患原因を調べる場合において，様々な合併症を伴うと各疾患データは複数の疾患群 (クラスター) に跨って所属することもあり得る．そこで，ソフトクラスタリングでは，データが複数のクラスターに属することを認めて，前述の短所を補っている．

ソフトクラスタリング手法の先駆けとして，Bezdek[1] により 1970 年代にファジィ概念を導入した Fuzzy c-means(FCM) モデルが提案されている．それは k-means 法にファジィ概念を導入したものであり，各クラスター内でのデータの散らばり具合 (変動) を最小化することによって，クラスター内の同質性を保証している [2, 7]．それゆえに，同質性の基準に焦点を当てたソフトクラスタリング手法が多く提案されていた．ところが，その基準だけでは以下に述べるように不十分であるということが後に認識された．複数のクラスターに各々のデータが跨ることを許したため，幾つかのクラスター中心が非常に近接した領域に存在するようなクラスタリング結果は，それぞれのクラスターの特徴づけの差別化が困難となり，ソフトクラスタリングの意味が失われる．従って，各々のクラスターがより適切に分離されるための基準「整分離性 (well-separation)[7]」は Wang et.al において考慮されるようになり，Wang et.al[7] は前述の同質性および整分離性基準を考慮した 2 目的 FCM(Bi-objective FCM : BOFCM) モデルとして定式化した．さらに，彼らは BOFCM モデルによる分類処理の後，それぞれのクラスタリング結果の妥当性 (validity) は分割エントロピー (partition entropy) を用いて評価している．

ここで注意しなければならないのは，Wang et.al の BOFCM モデルから得られたクラスタリング結果について妥当性の評価を行っても，それが必ずしも良いことが保証されるわけではない．そこで，

本論文において我々はBOFCMモデルの計算過程にその結果の妥当性評価基準を追加して分類を行う3目的FCM(Tri-objective FCM:TOFCM)モデルを提案する.

2 1目的FCMモデルおよび2目的FCMモデルについて

2.1 帰属の概念

本節では,「各データが複数のクラスターに所属する度合い」について述べる.

例えば,クラスタリング結果として2つのクラスターC_1およびC_2があるとする.ここで,データを$\boldsymbol{x}$と表す.k-means法では,あるデータ$\boldsymbol{x}$がC_1に属し,C_2には属さないとすれば,それは別の表現を用いるとデータ$\boldsymbol{x}$がC_1に帰属する度合いは1と表現される.また,このときデータ$\boldsymbol{x}$がC_2に帰属する度合いは0と表せる.

この表現を拡張して,帰属の度合いのとりうる値を0から1までの実数とする.このとき,例えば「$\boldsymbol{x}$がC_1に帰属する度合いを0.8」,「$\boldsymbol{x}$がC_2に帰属する度合いを0.2」などの表現が可能となる.

従って,データ$\boldsymbol{x}$が2つのクラスターC_1およびC_2へ同時に帰属する(“所属する”)ことを表現できる.この考えがFCMモデルの根底となっている.以下では,帰属する度合いを「メンバーシップ値」と呼ぶことにする.

2.2 標準的なFCMモデル(評価基準が単一の場合)

最初に,FCMモデルを説明するのに用いる記号を列挙する.

- C_t:t番目のクラスター$(t=1,\cdots,c)$
- $\boldsymbol{x}_i \in \mathbb{R}^k$:$i$番目のデータ$(i=1,\cdots,n)$
- $\boldsymbol{v}_t \in \mathbb{R}^k$:クラスター$t$の中心
- u_{it}:データ$\boldsymbol{x}_i$のクラスターC_tに対するメンバーシップ値

  ただし,以下の条件を満たす.

$$0 \leqslant u_{it} \leqslant 1, \quad \sum_{t=1}^{c} u_{it} = 1, i = 1, \cdots, n \tag{1}$$

- $U=[u_{it}]$:メンバーシップ値u_{it}を要素にもつ$n \times c$行列
- $V=[\boldsymbol{v}_t]$:クラスター中心$\boldsymbol{v}_t$を要素にもつ$n \times c$行列

次に,各データとそれぞれのクラスター中心とのユークリッド2乗距離の重み付け和で表現される同質性基準(目的関数)を述べ,それを制約条件(1)式のもとで最小化するFCMモデルを数理計画問題として,Bezdekは提案している.

$$\begin{aligned} \min \quad & J(U,V) = \sum_{t=1}^{c}\sum_{i=1}^{n} (u_{it})^m \|\boldsymbol{x}_i - \boldsymbol{v}_t\|^2 \\ \text{s.t.} \quad & \sum_{t=1}^{c} u_{it} = 1,\ i = 1, \cdots, n, \\ & 0 \leqslant u_{it} \leqslant 1,\ i = 1, \cdots, n; t = 1, \cdots, c \end{aligned} \tag{2}$$

ここで,mはあいまいさの強弱を調整するパラメータ$(m>1)$であり,予め決めておく必要がある.$m=1$の場合,FCMモデルの目的関数がk-meansモデルの場合に一致し,最適解は$0 \leqslant u_{it} \leqslant 1$の範囲内に存在しないと述べられている[3].このとき,メンバーシップ値をファジィ化するだけではソフトクラスタリングが行われないことが示される.そのため,Bezdekは目的関数$J(U,V)$をu_{it}について非線形化し,改善したのである.

以下に，Bezdek による FCM の解法手順を示しておく．

[FCM モデルの解法手順]
1) 初期設定
データ集合 $\{x_1, \cdots, x_n\}$ を与えて，クラスター数 $t(2 \leqslant t \leqslant c)$ および $m \in (1, \infty)$ を固定する．メンバーシップ u_{it} の初期値 $U^{(0)} = \{u_{it}^0\}$，十分に小さな正の数 ε を与える．

2) クラスター中心を次式により計算

$$v_t^p = \sum_{i=1}^{n} (u_{it}^p)^m x_i \Big/ \sum_{i=1}^{n} (u_{it}^p)^m \tag{3}$$

3) メンバーシップ値 u_{it}^p から u_{it}^{p+1} への更新は次式を用いて行う．

$$u_{it}^{p+1} = \left\{ \sum_{s=1}^{c} \left(\frac{\|x_i - v_t^p\|^2}{\|x_i - v_s^p\|^2} \right) \right\}^{-\frac{1}{m-1}}, \forall i, t \tag{4}$$

4) もし $\|u_{it}^{p+1} - u_{it}^p\| < \varepsilon$ が成立すれば終了し，そうでなければ p を $p+1$ に変更して 2) へ戻る．
なお，v_t^p および u_{it}^{p+1} は，ラグランジュ未定乗数法により得られる．

2.3 BOFCM：2 目的 FCM モデル (評価基準が 2 種類の場合)

前節で述べたクラスタリングでは，各クラスター内の同質性を高めることを目的としていた．これに対して，それぞれのクラスターの離れ具合を広げることを考える．つまり，それぞれ 2 つのクラスター中心間の距離の最大化

$$\max \quad L(V) = \sum_{t=1}^{c} \sum_{s<t} \|v_t - v_s\|^2 \tag{5}$$

を図る．それを以下では well-separation 基準と呼ぶことにする．2 目的 FCM(BOFCM) モデルは，目的関数 $J(U,V)$ および $L(V)$ を同時に最適化する 2 目的非線形計画問題として，次のように Wang et.al によって提案された．
(BOFCM)

$$\begin{aligned} \min \quad & J(U,V) = \sum_{t=1}^{c} \sum_{i=1}^{n} (u_{it})^m \|x_i - v_t\|^2 \\ \max \quad & L(V) = \sum_{t=1}^{c} \sum_{s<t} \|v_t - v_s\|^2 \\ \text{s.t.} \quad & \sum_{t=1}^{c} u_{it} = 1, \ i = 1, \cdots, n, \\ & 0 \leqslant u_{it} \leqslant 1, \ i = 1, \cdots, n; t = 1, \cdots, c \end{aligned} \tag{6}$$

BOFCM モデル (6) は，2 目的非線形計画問題であるため，このままでは求解は実質的に困難であり，完全最適解 [5] が存在するとは限らない．従って，BOFCM モデルはスカラー化手法によって，次の代替的な問題に変換される．

$$\begin{aligned} \min \quad & \widehat{JL}(U,V) = \beta \sum_{t=1}^{c} \sum_{i=1}^{n} (u_{it})^m \|x_i - v_t\|^2 - \alpha \sum_{t=1}^{c} \sum_{s<t} \|v_t - v_s\|^2 \\ \text{s.t.} \quad & \sum_{t=1}^{c} u_{it} = 1, \ i = 1, \cdots, n, \\ & 0 \leqslant u_{it} \leqslant 1, \ i = 1, \cdots, n; t = 1, \cdots, c \end{aligned} \tag{7}$$

ここで，α はクラスター内での散らばり具合 J に対する正のウエイトであり，β はクラスター間の離れ具合 L に対する非負のウエイトである．これらのウエイトは，$\alpha+\beta=1$ を満たすものとする．$\alpha=0$ の場合は，各々のクラスターがより適切に分離されるための基準「整分離性 (well-separation)」を無視した Bezdek による従来の FCM モデルと一致する．一般的に，(7) 式に対する解法が知られていないため，Wang et.al は次の定理を導出して BOFCM モデル (7) に最適解が存在することを示している．

[定理] BOFCM モデルの性質 (文献 [2] による)
$x_i \neq v_t$ $(\forall i,t)$ と仮定するとき，モデル (7) 式がステップワイズな最適解をもつための必要十分条件は

$$\begin{aligned}\alpha &< \min_t\left\{\frac{\sum_{i=1}^n (u_{it})^m}{\sum_{i=1}^n (u_{it})^m + c - 1}\right\},\\ \beta &> \max_t\left\{\frac{c-1}{\sum_{i=1}^n (u_{it})^m + c - 1}\right\}\end{aligned} \tag{8}$$

である．この BOFCM モデルの性質によって，次の解法手順が Wang et.al によって示されている．

[BOFCM モデルの解法手順]
1) 初期設定
データ集合 $\{x_1,\cdots,x_n\}$ を与えて，クラスター数 $t(2 \leqslant t \leqslant c)$ および FCM モデルで用いられる $m \in (1,\infty)$ を固定する．メンバーシップ u_{it} の初期値 $U^{(0)}=\{u_{it}^0\}$，非常に小さな正の数 δ および ε を与える．

2) 各々の目的関数に対するウエイトの計算
(8) 式を満たす α および β を次式によって計算する．

$$\alpha = \min_t\left\{\frac{\sum_{i=1}^n (u_{it}^p)^m}{\sum_{i=1}^n (u_{it}^p)^m + c - 1}\right\} - \delta \tag{9}$$

$$\beta = \max_t\left\{\frac{c-1}{\sum_{i=1}^n (u_{it}^p)^m + c - 1}\right\} + \delta \tag{10}$$

3) クラスター中心の計算
α および β，u_{it}^p を用いて，次式によりクラスター中心を計算する．

$$v_t^p = \frac{r_t}{q_t} - \frac{\alpha\sum_{s=1}^c \frac{r_s}{q_s}}{1+\alpha\sum_{s=1}^c \frac{1}{q_s}}, \quad \forall t \tag{11}$$

ここで，

$$q_t = \beta\sum_{i=1}^n (u_{it}^p)^m - \alpha c, \quad r_t = \beta\sum_{i=1}^n (u_{it}^p)^m x_i$$

$\alpha=0$ の場合，Bezdek の FCM モデルにおけるクラスター中心の計算式と一致する．

4) メンバーシップ値 u_{it}^{p} から u_{it}^{p+1} への更新
v_t^p および次式を用いて行う.

$$u_{it}^{p+1} = \left\{ \sum_{s=1}^{c} \left(\frac{\|\boldsymbol{x}_i - \boldsymbol{v}_t^p\|^2}{\|\boldsymbol{x}_i - \boldsymbol{v}_s^p\|^2} \right) \right\}^{-\frac{1}{m-1}}, \forall i, t \tag{12}$$

このメンバーシップ値の更新式は，Bezdek の FCM モデルにおいても用いられている.
5) もし u_{it}^{p+1} が (8) 式を満たさなければ，p を $p+1$ に変更して 2) へ戻る.

6) もし $\|u_{it}^{p+1} - u_{it}^{p}\| < \varepsilon$ が成立すれば終了し，そうでなければ p を $p+1$ に変更して 3) へ戻る.
BOFCM モデルにおいても，Bezdek の FCM モデルと同様に v_t^p および u_{it}^{p+1} は，ラグランジュ未定乗数法により得られる.

3 BOFCM モデルの妥当性評価の改善

3.1 分割エントロピー (クラスターの妥当性評価基準)

最初に本節では，BOFCM モデルによって得られたクラスタリング結果の妥当性を評価するための基準について述べる.

クラスタリングの分類処理が完了した後，各クラスターは様々な初期値をもつ最終的なクラスタリング結果を得る．もし，正確な結果に対して様々な初期値を決定する十分な情報がなければ，クラスタリング結果の妥当性を評価する主観的尺度が必要とされる．その妥当性を評価する幾つかのアプローチは文献 [1] で議論されているが，しかしどのアプローチもそのまま適用することはできない．その中で，よく知られた測度が Bezdek によって提案され，分割のあいまいさが $H(u_{it}, c)$ によって定義される分割のエントロピーで判断される.

$$H(u_{it}, c) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{c} u_{it} \log u_{it}, \quad 0 \leqslant H(u_{it}, c) \leqslant \log c$$

$H(u_{it}, c)$ が小さければ小さいほど，良い分類が行われる.

しかし，BOFCM モデルによるクラスタリング結果に対して妥当性の評価を行っても，その良し悪しまでは保証されていない．それゆえに，我々は BOFCM モデルの計算過程においてクラスタリング結果の妥当性評価を行うほうが良い結果を得ると考え，次節で BOFCM モデルの拡張を図る.

3.2 TOFCM モデル：3 目的 FCM モデル (評価基準が 3 種類の場合) の提案

我々は，Wang et.al が提案した BOFCM モデルにクラスタリング結果の妥当性評価基準を追加したモデルを次のように提案する.

$$\begin{aligned}
\min \quad & J(U, V) = \sum_{t=1}^{c} \sum_{i=1}^{n} (u_{it})^m \|\boldsymbol{x}_i - \boldsymbol{v}_t\|^2 \\
\max \quad & L(V) = \sum_{t=1}^{c} \sum_{s<t} \|\boldsymbol{v}_t - \boldsymbol{v}_s\|^2 \\
\min \quad & H(U) = -\sum_{i=1}^{n} \sum_{t=1}^{c} u_{it} \log u_{it} \\
\text{s.t.} \quad & \sum_{t=1}^{c} u_{it} = 1, \ i = 1, \cdots, n, \\
& 0 \leqslant u_{it} \leqslant 1, \ \forall i = 1, \cdots, n; t = 1, \cdots, c
\end{aligned} \tag{13}$$

我々の提案モデル (13) においても (6) 式と同様に求解が難しいので，次の代替的な問題に変換する．

$$\begin{aligned} \min \quad & K(U,V) = \beta \sum_{t=1}^{c}\sum_{i=1}^{n}(u_{it})^m \|\boldsymbol{x}_i - \boldsymbol{v}_t\|^2 - \alpha \sum_{t=1}^{c}\sum_{s<t} \|\boldsymbol{v}_t - \boldsymbol{v}_s\|^2 - \gamma \sum_{t=1}^{c}\sum_{i=1}^{n} u_{it} \log u_{it} \\ \text{s.t.} \quad & \sum_{t=1}^{c} u_{it} = 1,\ i = 1, \cdots, n, \\ & 0 \leqslant u_{it} \leqslant 1,\ i = 1, \cdots, n; t = 1, \cdots, c \end{aligned} \tag{14}$$

ここで，α, β は BOFCM モデル (7) 式で用いているウエイトと同様であり，γ はクラスタリング結果の妥当性を評価する基準 H に対する正のウエイトである．3つのウエイトは，$\alpha + \beta + \gamma = 1$ を満たすものとする．$\alpha = \gamma = 0$ の場合は，Bezdek による FCM モデルに一致し，$\gamma = 0$ の場合は Wang et. al による BOFCM モデルに一致する．

4 おわりに

本稿では，Bezdek の FCM モデルおよび Wang et.al の BOFCM モデルの議論から展開して，データを複数のクラスターへ分類を行う際にエントロピーによるクラスタリング結果の妥当性評価を同時に考慮する 3 目的 FCM(TOFCM) モデルを提案した．さらに，エントロピーを用いたクラスタリング結果の妥当性基準を BOFCM モデルによる分類を行う際に併用することで，提案モデルはクラスタリングの正則化 [4] も同時に行っている可能性があるとも考えられる．

今後の課題として，提案モデルによるクラスタリング結果の妥当性評価が BOFCM モデルよる評価よりも有効であるかを検証することである．また，提案モデルに対してスカラー化手法により代替的な問題に変換しているが，目的関数が凸関数の差であるので D.C. 計画問題に基づく解法について考察したい．

参考文献

[1] J.C.Bezdek : " Pattern Recongniton with Fuzzy Objective Function Algorithms", (Plenum Press, New York and London, 1981).

[2] Hung,T.-W. : " The bi-objective fuzzy c-means cluster analysis for TSK fuzzy system identification", *Fuzzy Optimization Decision Making*, 6, pp.51-61 (2007)

[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).

[4] S.Miyamoto and M.Mukaidono: " Fuzzy c-means as a regularization and maximum entropy approach", *Proc. of the 7th International Fuzzy Systems Association World Congress(IFSA '97)*, Vol.II, pp.86-92, (1997)

[5] 中山弘隆，谷野哲三 :" 多目的計画法の理論と応用", コロナ社（計測自動制御学会編・発行），(1994)

[6] 齋藤堯幸，宿久　洋 :" 関連性データの解析法 (多次元尺度構成法とクラスター分析)", 共立出版 (2006)

[7] Wang,H.-F., Wang,C., and Wu, G.-Y. : " Bi-criteria fuzzy c-means analysis", *Fuzzy Sets and Systems*, 64, pp.311-319 (1994)