第3章　区間データに対するユークリッド距離に基づくファジィクラスタリング

3.1　はじめに

クラスタリングにおいて，区間データを扱う研究は，シンボリックデータ解析の分野で行われてきた [23, 24, 25, 15]．シンボリックデータ解析では，区間データの他に，通常の数値データ，名義データ等を統合的に扱う．シンボリックデータ解析においてファジィc-平均法を行った研究 [15] もあるが，厳密な交互最適化は行われていない．

また，最近，各個体に不確定性を含むデータに対するアルゴリズムの研究が盛んになってきている [28, 39, 49]．しかしながら，それらには厳密さや一般性について制約がある．Hathaway ら [28]，Pedrycz ら [49]，および Lee ら [39] のアプローチでは，厳密に交互最適化が行われていない．

本章では，従来の定式化より一般性かつ厳密性をもち，多数かつ多次元のデータに適用できる方法を考察する．ユークリッド空間を扱っている従来の手法では，厳密な交互最適化が行われていないことが問題であったが，本手法により厳密な交互最適化を実現する．

ここで考えるべき点として，個体における不確定性のクラスタリングにおける意義をどうとらえるかという問題がある．個体における不確定性は，個体どうしの間の距離に不確定性が生じることを意味している．この不確定性はやはり区間の形に表されるが，クラスタリングにおいて区間全体を一度に処理することは大変困難である．このようなとき，各個体における不確定性から生じた個体間の距離の不確定性を特徴づける値を用いてクラスタリングを行うのが良いと考えられる．また，これらの特徴づけのための値としては，距離の不確定性における上限と下限を選んだ．

ところで，距離の上限と下限は，後の節で述べる最長距離法と最短距離法にそれぞれ対応している．これら2つの'極端な'値をとり，それらを比較することによって，不確定性がどのような影響を及ぼすかをみることができる．

本論文では区間の直積で表されるデータを取り扱う．これを区間データと呼ぶことにする．他にファジィ数で表されるような不確定性を扱っている研究 [28, 39, 49] もあるが，区間は様々な種類のファジィ数で表される不確定性の最大値を扱うことになる．

ファジィc-平均法を区間データを扱うことができるように拡張する．最も極端な最長距離法，最短距離法の2種類の集合間の距離を用いて目的関数を定義する．本来ファジィc-平均法は，目的関数が定義され，その目的関数を最適化することによりクラスタリングを行う手法である．目的関数には個体のクラスターへの帰属度と，クラスター中心の2つの変数があり，一度に最適化することができないので交互最適化を用いる．目的関数を厳密に最適化するためには，交互最適化のそれぞれのステップで，個体のクラスターへの帰属度と，クラスター中心がそれぞれ厳密に最適化されなければならない．しかしながら区間データを扱う場合には，クラスター中心に関する最適化は，目的関数がクラスター中心に関して単純な2次関数の形をしていないため，従来のファジィc-平均法同様に簡単に最適解を求めることができない．そこで，クラスター中心に関する最適化のアルゴリズムを新たに開発し，それを用いて目的関数の最適化を行う．このアルゴリズムを用いることで，目的関数の厳密な交互最適化が実現できる．

2種類の区間データを準備し，不確定性を含んだまま本手法でクラスタリングした結果と，同じデータについて代表点をとって不確定性を含まないとした場合に，通常のファジィc-平均法を用いてクラスタリングした結果を比較する．

3.2　区間データ

個体の各々の成分が幅を持った値で与えられる場合，その幅は区間であるとみなすことができる．幅 (不確定性) をもたない値についても，区間の両端の値が一致したものとみなすことにより，区間として扱うことができる．

各々の成分が区間で表現される場合に，個体はそれらの区間のデカルト積である

とする．分類すべき区間データは

$$M = \{M_1, M_2, \cdots, M_n\} \tag{3.1}$$

と表され，個体 M_k の各々の成分 M_k^l は区間で表されるとする．

$$M_k^l = [f_{k1}^l, f_{k2}^l], (l = 1, \cdots, p). \tag{3.2}$$

個体 M_k は，区間のデカルト積で表される．

$$M_k = [f_{k1}^1, f_{k2}^1] \times \cdots \times [f_{k1}^p, f_{k2}^p]. \tag{3.3}$$

このような区間データをファジィc-平均法を用いてクラスタリングすることを試みる．

3.3　区間データに対する目的関数

既存の標準的なファジィc-平均法とエントロピー正則化を用いたファジィc-平均法の目的関数は，

$$J_m(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m \parallel x_k - v_i \parallel^2$$

$$J^\lambda(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik} \parallel x_k - v_i \parallel^2 + \lambda^{-1}\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}\log u_{ik}$$

であり，クラスタリングすべき個体 x_k は不確定性を含まないと仮定している．それぞれの目的関数では，ユークリッドノルムの二乗 $\parallel x_k - v_i \parallel^2$ が含まれている．不確定性を含まない個体 x_k のかわりに区間の直積で表される個体 M_k を扱うには，ユークリッドノルムを新たに定義する必要がある．

そこで，区間データに対する標準的なファジィc-平均法の目的関数とエントロピー法の目的関数を

$$J_m(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m D_{ik} \tag{3.4}$$

$$J^\lambda(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}D_{ik} + \lambda^{-1}\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}\log u_{ik} \tag{3.5}$$

のように書き直す．ただし，その中で用いるユークリッド距離の二乗は

$$D_{ik} = D_2(M_k, v_i)^2 = \parallel M_k - v_i \parallel^2 \tag{3.6}$$

である．制約条件は (2.3) と同じである．

$$\mathcal{M} = \{(u_{ik}) \mid u_{ik} \in [0,1], \sum_{i=1}^{c} u_{ik} = 1, k = 1, 2, \cdots, n\}.$$

$D_2(M_k, v_i)^2$ を求めるには，各々の軸方向にそれぞれ独立にユークリッド距離の二乗 $D_2^l(M_k, v_i)^2$ を求め，それらを加え合わせればよい．

$$D_2(M_k, v_i)^2 = \sum_{l=1}^{p} D_2^l(M_k, v_i)^2.$$

さて，本論文では，クラスター中心は不確定性を含まないと仮定する．すると，クラスター中心 v_i は 1 つの要素からなる集合であり個体 M_k の各成分は区間つまり集合である．区間の直積で表される個体 M_k と不確定性を含まないクラスター中心 v_i のユークリッド距離の二乗 $D_2^l(M_k, v_i)^2$ を最長距離と最短距離を用いて定義する．集合 K と集合 L に対し距離 $d(K, L)$ は，

最長距離法 (Farthest Neighbor)

$$d(K, L) = \max\{d(x, y) : \ for\ all\ x \in K, y \in L\}$$

最短距離法 (Nearest Neighbor)

$$d(K, L) = \min\{d(x, y) : \ for\ all\ x \in K, y \in L\}$$

である．この他に最長距離法と最短距離法の間の性質を持つハウスドルフ距離が考えられるが，本手法においては最長距離法と一致することが前章において示されているので，上の 2 種類の距離を用いる．

これらの距離を用いて $D_2^l(M_k, v_i)^2$ を定義する．最長距離法を用いた場合は，

$$D_2^l(M_k, v_i)^2 = \max\{(v_i^l - f_{k1}^l)^2, (v_i^l - f_{k2}^l)^2\}$$

最短距離法を用いた場合は，

$$D_2^l(M_k, v_i)^2 = \begin{cases} 0, & (f_{k1}^l \le v_i^l \le f_{k2}^l) \\ \min\{(v_i^l - f_{k1}^l)^2, (v_i^l - f_{k2}^l)^2\}, & (otherwise) \end{cases}$$

のように定まる．このようにして集合間の距離を用いて (3.6) の D_{ik} が定まり，区間データに対するファジィ c-平均法の目的関数 (3.4)，(3.5) が定まる．

3.4　クラスター中心に対する最適化

目的関数が定義されたので，それを最適化することを考える．既存のファジィc-平均法同様アルゴリズム**FCM**を用いて，2つの変数(U,V)を交互最適化することにより最適解を求める．

$$J(U,V)=J_m(U,V)\quad \text{もしくは}\quad J(U,V)=J^{\lambda}(U,V).$$

と目的関数を置き換え，アルゴリズム**FCM**を適用する．

ファジィc-平均法アルゴリズム

FCM1. $\bar{U}$と$\bar{V}$の初期値を適当に決める．

FCM2. $\min_{U\in\mathcal{M}} J(U,\bar{V})$の最適解$\hat{U}$を求め，$\bar{U}=\hat{U}$と置き換える．

FCM3. $\min_{V} J(\bar{U},V)$の最適解$\hat{V}$を求め，$\bar{V}=\hat{V}$と置き換える．

FCM4. 最適解$(\bar{U},\bar{V})$が収束していれば終了，
そうでなければステップ**FCM2**に戻る．

FCM2における解は，既存のファジィc-平均法の解と同様に標準的なファジィc-平均法では，

$$\bar{u}_{ik}=\left[\sum_{j=1}^{c}\left(\frac{D_{ik}}{D_{jk}}\right)^{\frac{1}{m-1}}\right]^{-1} \tag{3.7}$$

エントロピー法では，

$$\bar{u}_{ik}=\frac{e^{-\lambda D_{ik}}}{\sum_{j=1}^{c} e^{-\lambda D_{jk}}} \tag{3.8}$$

で与えられる．D_{ik}は(3.6)で与えられる．

ところが**FCM3**の解は通常のファジィc-平均法と同様に求めることができない．不確定性を含まないデータを扱う既存のファジィc-平均法では目的関数が最適化すべき変数v_iに対して2次関数であるのでv_iに関する偏導関数を0とするv_iが最適解となり(2.5)，(2.11)で与えられる．区間データに対する標準的なファジィc-平均法の目的関数は(3.4)であり，エントロピー法の目的関数は(3.5)である．D_{ik}は集合間の距離を用いて定義され，これらの目的関数はクラスター中心v_iに関して単純な

2次関数ではない．そこで，目的関数 (3.4)，(3.5) に対する FCM3 の最適解を求めるアルゴリズムを新たに開発する．

3.5　最適化アルゴリズム

目的関数のクラスター中心に関する偏導関数を求めると，それぞれの成分に対し，区分線形で単調増加であることがわかる．この区分線形関数を探索するアルゴリズムを用いることにより，目的関数のクラスター中心に関する最適解が得られる．

まず，クラスター中心 v_i に関する目的関数の形状を知るために v_i に関する一階偏導関数をそれぞれの目的関数に対して求める．

$$\frac{\partial J_m(U,V)}{\partial v_i^l} = \sum_{k=1}^{n}(u_{ik})^m \frac{\partial}{\partial v_i^l} D_2^l(M_k, v_i)^2. \tag{3.9}$$

$$\frac{\partial J^\lambda(U,V)}{\partial v_i^l} = \sum_{k=1}^{n}(u_{ik}) \frac{\partial}{\partial v_i^l} D_2^l(M_k, v_i)^2. \tag{3.10}$$

最長距離法の場合

偏導関数の後半部分について，最長距離法を用いた場合には，

$$\begin{aligned}\frac{\partial}{\partial v_i^l} D_2^l(M_k, v_i)^2 &= \frac{\partial}{\partial v_i^l}\left[\max\{(v_i^l - f_{k1}^l)^2, (v_i^l - f_{k2}^l)^2\}\right] \\ &= \begin{cases} 2(v_i^l - f_{k1}^l), & \text{for } v_i^l \geq \frac{f_{k1}^l + f_{k2}^l}{2} \\ 2(v_i^l - f_{k2}^l), & \text{for } v_i^l \leq \frac{f_{k1}^l + f_{k2}^l}{2} \end{cases}\end{aligned} \tag{3.11}$$

と表せる．図 3.1 は (3.9) 式におけるある k についての $(u_{ik})^m \frac{\partial}{\partial v_i^l} D_2^l(M_k, v_i)^2$ を示している．黒点は $v_i^l = \frac{f_{k1}^l + f_{k2}^l}{2}$ における値であり，(3.11) 式にみられるように不連続性が生じる．

(注) 右導関数と左導関数が一致しない場合は両方の値を表示し，二値関数としている．

図 3.1 ではわかりやすくするために折れ線を 2 つだけ示しているが，これらをすべての k に対して加え合わせた折れ線は図 3.2 のようになり，(3.9),(3.10) 式の例示はこのような形になる．

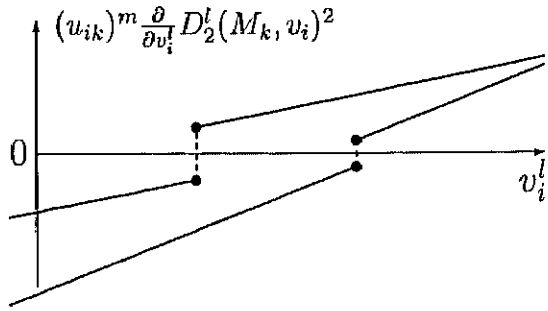

図 3.1: 2 つの異なる k についての $(u_{ik})^m \frac{\partial}{\partial v_i^l} D_2^l(M_k, v_i)^2$ の例示 (最長距離法)

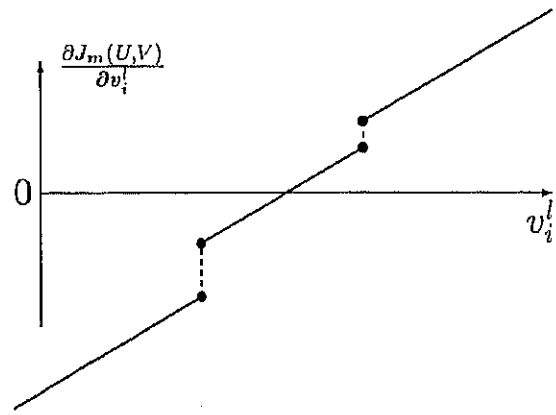

図 3.2: (3.9) 式の例示 (最長距離法)

このように一階偏導関数は，区分線形な関数を加え合わせた形状をしている．また，(3.11) 式は単調増加関数なので (3.9)，(3.10) は単調増加することがわかる．このことから目的関数 (3.4)，(3.5) は変数 v_i に関して凸関数であることがわかる．この目的関数を最小にする v_i を求めるには，一階偏導関数が負から正に転ずる v_i を探索すればよい．

探索のためのアルゴリズムを以下に示す．以下のアルゴリズムでは (3.9) 式に対して探索を行っているが，(3.10) 式に対しても $m=1$ と置くことにより同様に探索を行うことができる．(3.9) 式は線形な区間で v_i^l に関して共通の傾き $\sum_{k=1}^{n} 2(u_{ik})^m$ を持っていることに注意する．(アルゴリズムの中ではこれを U とおく) また不連続点を

$$X_k^l = \frac{f_{k1}^l + f_{k2}^l}{2}$$

と置き，

$$X_1^l \leq \cdots \leq X_k^l \leq \cdots \leq X_n^l$$

のように並べかえる．

Algorithm (Searching v_i^l in Farthest Neighbor)

begin
  $U := \sum_{k=1}^{n} 2(u_{ik})^m$;
  $S := \sum_{k=1}^{n} (u_{ik})^m (-2)(f_{k2}^l - X_1^l)$;
  $k := 0$;
  while (1) **do begin**
    $k := k + 1$;
    $J_k := (u_{ik})^m (-2)(f_{k1}^l - f_{k2}^l)$;
    $S := S + J_k$;
    if $S > 0$ **then** $\bar{v}_i^l := X_k^l$; **break**;
    $S := S + U(X_{k+1}^l - X_k^l)$;
    if $S > 0$ **then** $\bar{v}_i^l := X_{k+1}^l - \frac{S}{U}$; **break**;
  end;
  output $\bar{v}_i^l$ **as the** l-th coordinate of the
  cluster center v_i
end.

J_k は (3.9) 式の不連続点におけるジャンプの大きさを示している．このアルゴリズムでは，目的関数 (3.4) 式の最適解を求めるために，図 3.2 のような形状をした不連続な一階偏導関数 (3.9) 式が 0 を横切る位置を探索する．そこで，図 3.2 における不連続点 (•) を下から順序探索する．S は探索中の不連続点における (3.9) 式の値を示している．探索中に不連続点で縦軸つまり (3.9) 式の値が 0 を超えた場合は上記のア

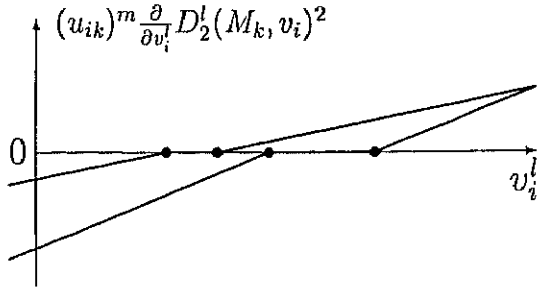

図 3.3: 2つの異なる k についての $(u_{ik})^m \frac{\partial}{\partial v_i^l} D_2^l(M_k, v_i)^2$ の例示 (最短距離法)

ルゴリズム中の上側の **break** で，その不連続点の位置をクラスター中心として出力する．連続な区間で 0 を超えた場合は下側の **break** で (3.9) 式が 0 となる位置をクラスター中心として出力する．

最短距離法の場合

最短距離法を用いた場合の目的関数のクラスター中心に関する偏導関数の後半部分は，

$$
\begin{aligned}
D_2^l(M_k, v_i)^2 &= \begin{cases} 0, & (f_{k1}^l \le v_i^l \le f_{k2}^l) \\ \min\{(v_i^l - f_{k1}^l)^2, (v_i^l - f_{k2}^l)^2\}, & (otherwise) \end{cases} \\
&= \begin{cases} (v_i^l - f_{k1}^l)^2, & (v_i^l \le f_{k1}^l) \\ 0, & (f_{k1}^l \le v_i^l \le f_{k2}^l) \\ (v_i^l - f_{k2}^l)^2, & (f_{k2}^l \le v_i^l) \end{cases}
\end{aligned}
$$

と書き直すと

$$
\frac{\partial}{\partial v_i^l} D_2^l(M_k, v_i)^2 = \begin{cases} 2(v_i^l - f_{k1}^l), & (v_i^l \le f_{k1}^l) \\ 0, & (f_{k1}^l \le v_i^l \le f_{k2}^l) \\ 2(v_i^l - f_{k2}^l), & (f_{k2}^l \le v_i^l) \end{cases} \tag{3.12}
$$

とかける．

図 3.3 のような関数を k に関して加え合わせることにより図 3.4 のような v_i に関する一階偏導関数 (3.9) が求められる．

(3.12) 式は単調増加関数なので一階偏導関数 (3.9)，(3.10) は単調増加することがわかり，目的関数は変数 v_i に関して凸関数であることがわかる．先ほどと同様に一

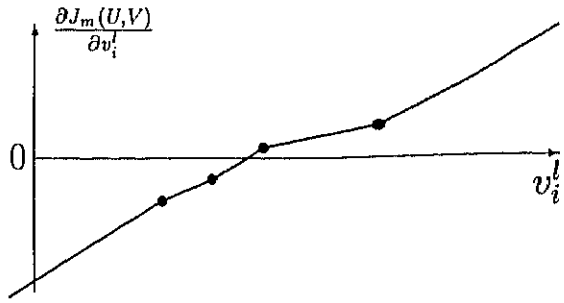

図 3.4: (3.9) 式の例示 (最短距離法)

階偏導関数が0の値をとる v_i を探索すればよく，そのアルゴリズムを以下に示す．まず $f_{k1}^l(k=1,\cdots,n)$ と $f_{k2}^l(k=1,\cdots,n)$ を小さい順にソーティングし，

$$X_1^l \leq \cdots \leq X_j^l \leq \cdots \leq X_{2n}^l$$

のように並べかえる．

Algorithm (Searching v_i^l in Nearest Neighbor)

begin
  $U := \sum_{k=1}^{n} 2(u_{ik})^m$;
  $S := \sum_{k=1}^{n} (u_{ik})^m 2(X_1^l - f_{k1}^l)$;
  $j := 1$;
  while $(S < 0)$ **do begin**
    if $X_j^l = f_{k1}^l$ **then** $U := U - 2(u_{ik})^m$;
    if $X_j^l = f_{k2}^l$ **then** $U := U + 2(u_{ik})^m$;
    $S := S + U(X_{j+1}^l - X_j^l)$;
    $j := j + 1$;
  end;
  $\bar{v}_i^l := X_j^l - \frac{S}{U}$
  output $\bar{v}_i^l$ as the l-th coordinate of the
  cluster center v_i
end.

このアルゴリズムでは，目的関数 (3.4) 式の最適解を求めるために，図 3.4 のような形状をした連続な一階偏導関数 (3.9) 式が 0 を横切る位置を探索する．そこで，図 3.4 における $X_j^l(\bullet)$ を下から順序探索する．探索中に縦軸つまり (3.9) 式の値が 0 を超えた場合はループを抜け，(3.9) 式が 0 となる位置をクラスター中心として出力する．

このようにして，区間データに対して **FCM3** の最適解をアルゴリズムによって解くことができる．これらのアルゴリズムは，アルゴリズムの形からもわかるように

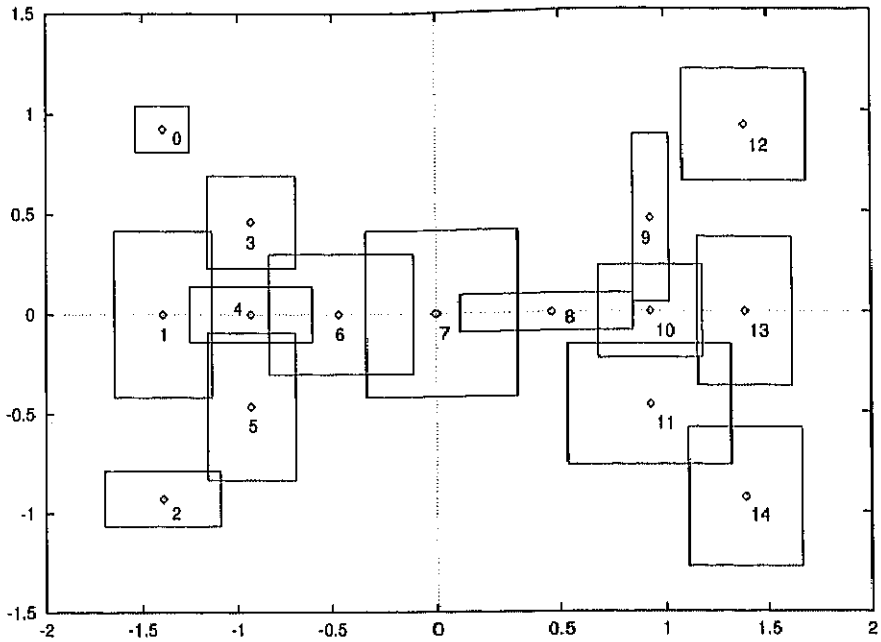

図 3.5: バタフライデータ (不確定性あり)

アルゴリズムを行う前のソーティングを除けば，計算量は $O(n)$ で充分である．既存のファジィ c-平均法でクラスター中心を求めるステップでの計算量も $O(n)$ である．

3.6　数値例とその結果

まず，人工的に作成されたバタフライデータと呼ばれるデータに，区間の不確定性を仮定したもの，それから，人物の印象に対する幅をもった回答からなるデータに対し数値実験を行った．それぞれのデータが不確定性を含まないと仮定した場合の従来のファジィ c-平均法の結果と，不確定性を含んだまま，ここで提案した手法でクラスタリングした結果を比較する．

バタフライデータに対する数値実験

簡単なバタフライデータと呼ばれる 15 個体からなる不確定性を含まないデータ (図 3.5 におけるそれぞれの長方形の中心点からなるデータ) を，既存のファジィ c-平均法でクラスタリングした結果と，同じデータに中心から両側に対称に不確定性

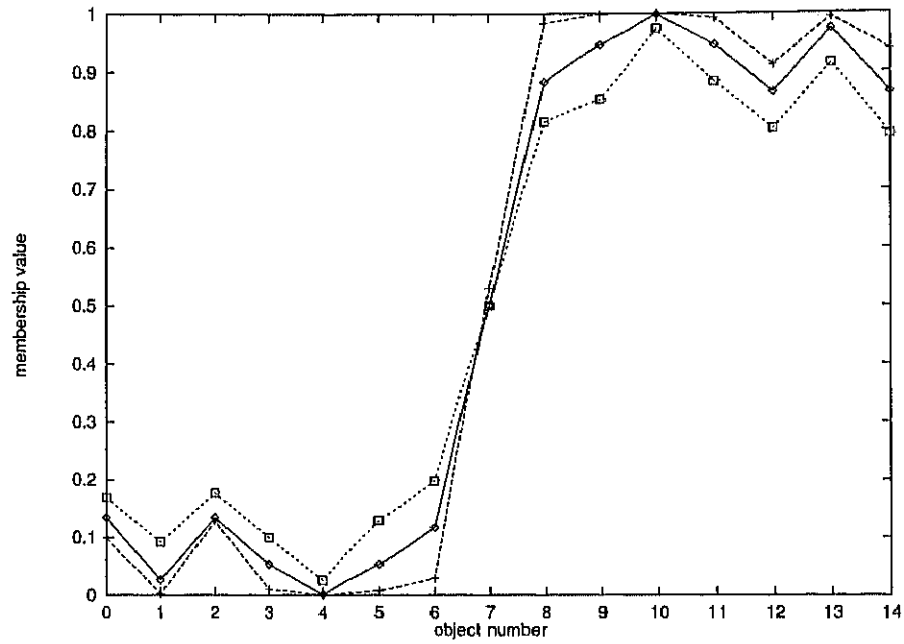

図 3.6: バタフライデータに対するクラスタリング結果 (個体番号 k に対するメンバシップ u_{1k}，標準的なファジィ c-平均法)

を仮定した図 3.5 のような区間データを本手法で 2 つにクラスタリングした結果を示す.

クラスタリング手法は，標準的なファジィ c-平均法 (パラメータ $m = 2.0$) と，エントロピー法 (パラメータ $\lambda = 0.7$) を用いた. アルゴリズム **FCM** における収束判定基準は，各ループでのメンバシップの変化の最大値がある値より小さいことを使った.

$$\max_{i,k} \mid u_{ik} - \bar{u}_{ik} \mid < \varepsilon, \quad \varepsilon = 1.0 \times 10^{-6}. \tag{3.13}$$

図 3.6，図 3.7 は標準的なファジィ c-平均法とエントロピー法を用いて，クラスタリングした結果である. 横軸は個体番号で，縦軸は右側のクラスターに対するメンバシップ値を表している. 区間データに対する最長距離法，最短距離法を用いた結果はそれぞれ□，＋で表される. また，見やすくするために点線で結んでいる. 不確定性を含まないデータに対する結果は◇で表され，実線で結んでいる.

図にみられるように，同じデータでも不確定性を含まない場合と不確定性を仮定した場合はかなり違った結果を示す. また最長距離法を用いたメンバシップと最短距離法を用いたメンバシップは異なっており，不確定性を含まない場合のメンバシッ

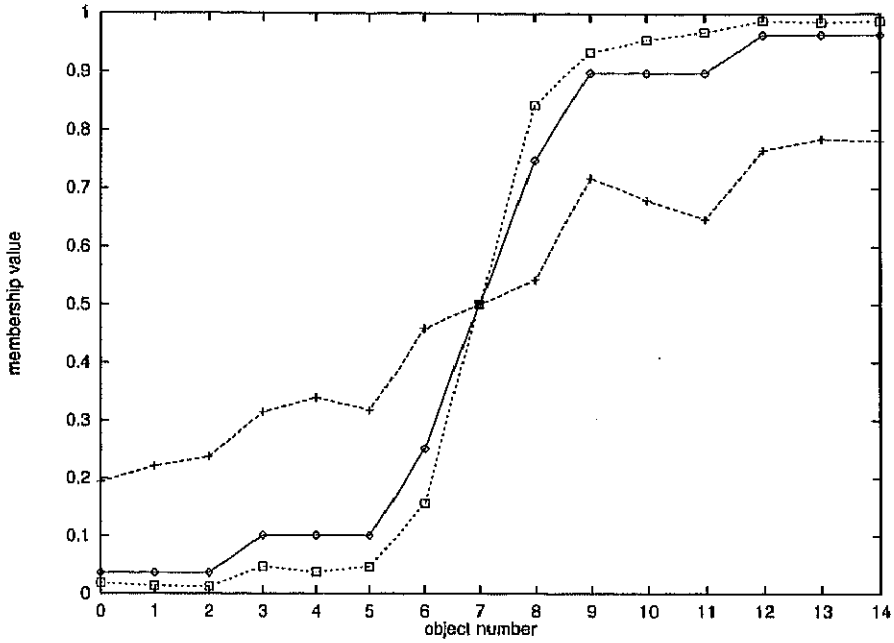

図 3.7: バタフライデータに対するクラスタリング結果 (個体番号 k に対するメンバシップ u_{1k}，エントロピー法)

プを両側からはさんでいる．ただし，この不確定性を含まない場合のメンバシップを不確定性を含む場合のメンバシップがはさんでいることについては，このデータに限った場合である．実際，次に述べる実データにおけるクラスタリング結果では，このメンバシップの関係が成り立たない場合がある．

標準的なファジィc-平均法とエントロピー法を用いた場合でも，不確定性を含まないと仮定した場合，不確定性を含むデータに対し最長距離，最短距離を用いた場合で，それぞれメンバシップの値が異なる．このように同じデータに対して異なる結果が得られた場合には結果が安定しているとはいえず，それらの結果を比較検討する必要がある．また，同じ結果が得られた場合には，結果は安定しているといえる．クラスター中心の位置に関しては，それぞれの場合で差異がほとんどみられなかった．

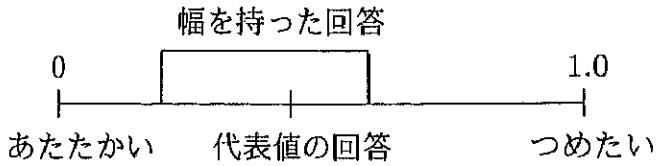

図 3.8: 1つの成分の回答例 ([61] による)

人物の印象に関する実データに対する数値実験

次に，実際に調査されたデータに対する数値実験の結果を示す．被験者は私立K女子大学およびK女子短期大学学生400名であった．質問紙は，京都市内のK女子大学およびK女子短期大学で行われた俳優M氏の講演会へ出席する意図と，M氏についてのイメージに関するものであった．なお，M氏の講演会は1990年10月16日であり，調査時点は1990年7月中旬であった [61].

この調査の中のM氏のイメージに関する回答を本数値実験で用いた．データはあたたかい－つめたい，おもしろい－つまらない，ばかだ－かしこい，やさしい－いじわるだ，魅力がない－魅力がある，嫌いだ－好きだ，の6つの成分からなる．被験者はそれぞれの成分に関し幅を持って回答し，またその代表値も回答した (図3.8). それぞれの回答は $[0,1]$ に入るように標準化されている．

データは6つの成分からなるが，そのうちの2つの成分 (あたたかい－つめたい，おもしろい－つまらない) の代表値は図3.9のように分布している．図3.10，3.11，3.12，3.13は代表値を用いて既存のファジィc-平均法を行った結果 (+) と，幅を持った回答を本手法でクラスタリングした結果 (最長距離法＊，最短距離法×) である．それぞれクラスター中心の位置を示している．標準的なファジィc-平均法ではパラメータ $m = 1.8$，エントロピー法ではパラメータ $\lambda = 20.0$ とし，収束判定基準は (3.13) を用いた．横軸は (あたたかい－つめたい)[0,1]，縦軸は (おもしろい－つまらない)[0,1] である．ここでは2分類，3分類の結果を示している．4分類，5分類に関しても数値実験を行ったが，クラスター中心の配置の傾向が3分類と似ていたため，結果は省略する．特に3分類を行った場合にクラスター中心の位置が手法によって異なっている．また，図3.11，図3.13の最長距離法を用いた結果＊によく表れているように，よりあたたかく感じる人はよりおもしろく感じる傾向と，よりつめたくつまらなく感じる傾向の他に，比較的あたたかく感じているけれど，あまりおもし

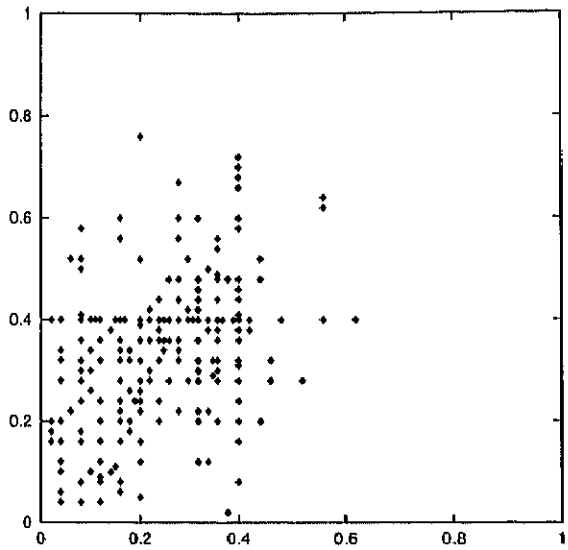

図 3.9: M 氏の印象 6 次元データの 2 成分 (横軸:あたたかい－つめたい，縦軸:おもしろい－つまらない)

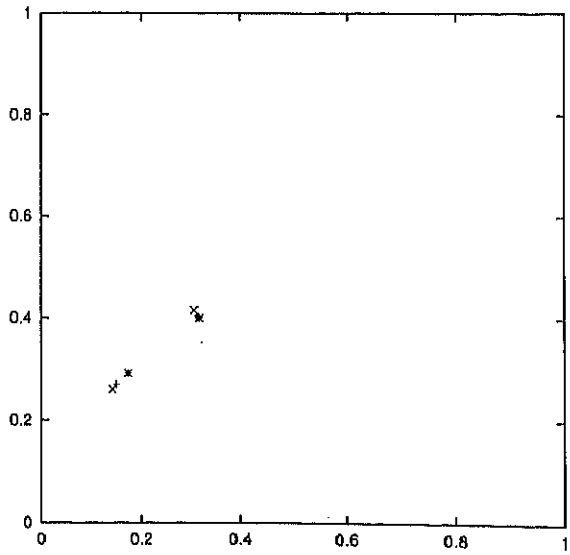

図 3.10: M 氏の印象 6 次元データのクラスタリング (標準的なファジィ c-平均法，2 分類) 結果の 2 成分 (横軸:あたたかい－つめたい，縦軸:おもしろい－つまらない)

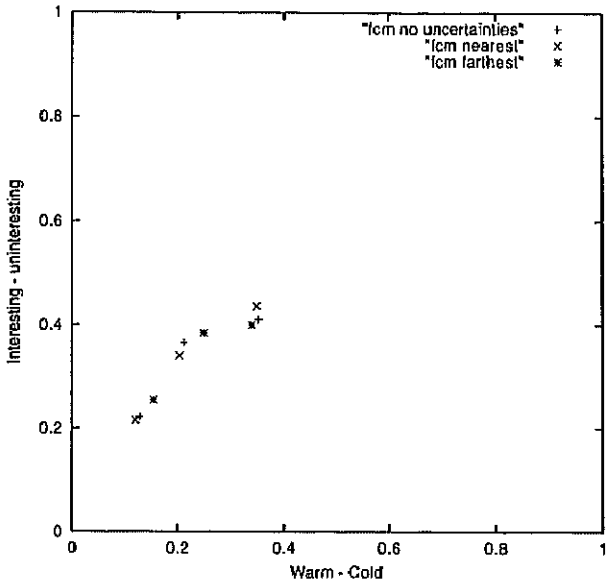

図 3.11: M 氏の印象 6 次元データのクラスタリング (標準的なファジィ c-平均法，3 分類) 結果の 2 成分 (横軸:あたたかい－つめたい，縦軸:おもしろい－つまらない)

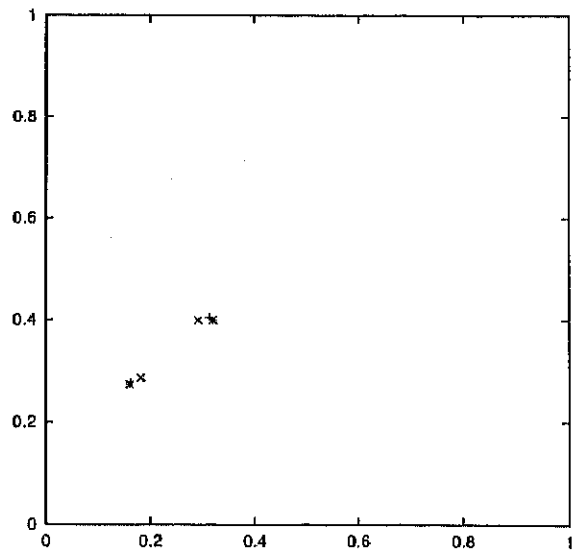

図 3.12: M 氏の印象 6 次元データのクラスタリング (エントロピー法，2 分類) 結果の 2 成分 (横軸:あたたかい－つめたい，縦軸:おもしろい－つまらない)

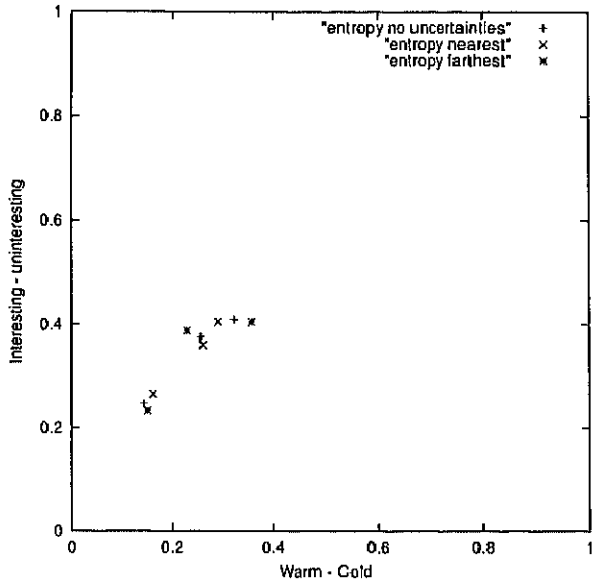

図 3.13: M 氏の印象 6 次元データのクラスタリング (エントロピー法，3 分類) 結果の 2 成分 (横軸:あたたかい－つめたい，縦軸:おもしろい－つまらない)

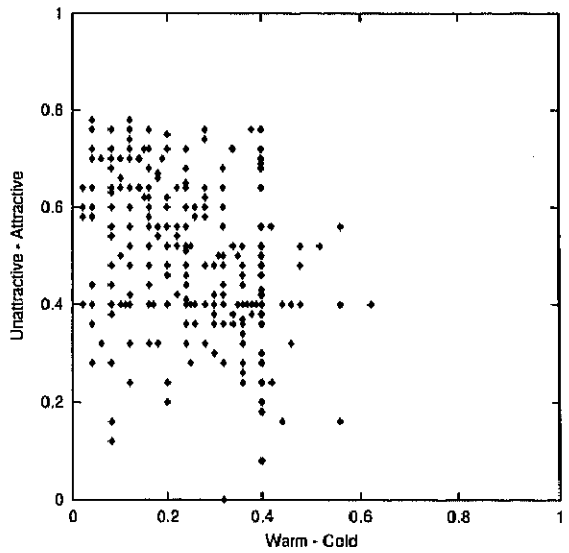

図 3.14: M 氏の印象 6 次元データの 2 成分 (横軸:あたたかい－つめたい，縦軸:魅力がない－魅力がある)

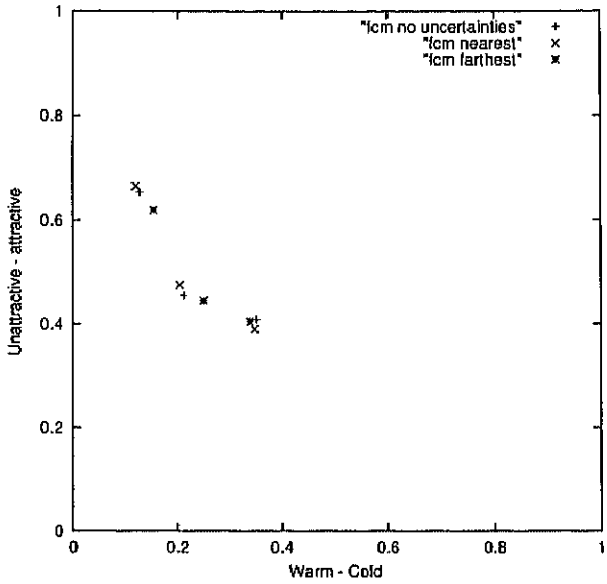

図 3.15: M 氏の印象 6 次元データのクラスタリング (標準的なファジィ c-平均法，3 分類) 結果の 2 成分 (横軸:あたたかい－つめたい，縦軸:魅力がない－魅力がある)

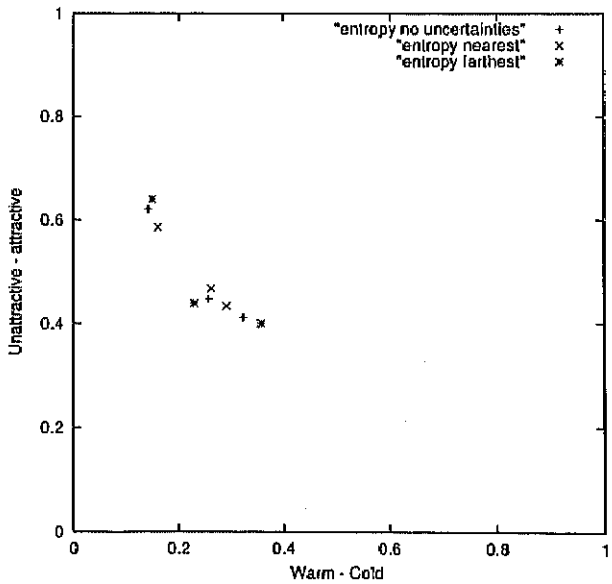

図 3.16: M 氏の印象 6 次元データのクラスタリング (エントロピー法，3 分類) 結果の 2 成分 (横軸:あたたかい－つめたい，縦軸:魅力がない－魅力がある)

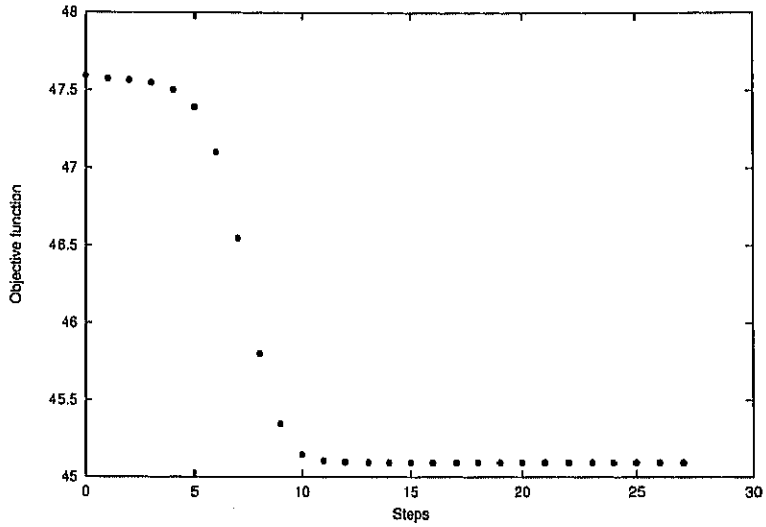

図 3.17: 目的関数が単調減少している例

ろく感じていない第3の傾向が存在することを示している．また，同様に (あたたかい－つめたい，魅力がない－魅力がある) の2成分に関して，代表値は図 3.14 のように分布している．図 3.15，3.16 は代表値を用いて既存のファジィ c-平均法を行った結果 (+) と，幅を持った回答を本手法でクラスタリングした結果 (最長距離法＊，最短距離法×) である．それぞれクラスター中心の位置を示している．この結果でも，それぞれの場合で異なったクラスター中心の配置が得られた．

本章で述べたクラスター中心計算アルゴリズムでは，目的関数はクラスター中心に関して厳密に最適化 (最小化) が行われている．これは，ファジィ c-平均法のアルゴリズムのクラスター中心の最適化のステップ **FCM3** において，目的関数が変化しないかあるいは減少していることを意味する．また，ステップ **FCM2** の最適化も厳密に行われているので，目的関数は同様に変化しないかもしくは減少する．このステップ **FCM2**，ステップ **FCM3** を繰り返すのがファジィ c-平均法のアルゴリズムなので，目的関数はファジィ c-平均法によって明らかに単調減少する．

このことを例示するために，本数値実験において，実際に調査されたデータに対する実験を行ったとき，アルゴリズム **FCM** の各ループにおける目的関数の値を図 3.17 に示す．ここでは，標準的なファジィ c-平均法 (パラメータ $m = 1.8$) において

最長距離法を用いて，2分類を行った場合を示している．クラスター中心に関する最適化 (最小化) を厳密に行っているので，目的関数の値は単調に減少している．

3.7　まとめ

ファジィクラスタリングにおける区間データの扱い方を提案した．各個体がクリスプな区間の直積で表されるデータを，ファジィc-平均法でクラスタリングする手法を考えた．

最長距離法，最短距離法の2種類の集合間の距離を用いて，区間データとクラスター中心との間の距離を定義し，目的関数を定義した．

最長，最短距離法以外の集合間の距離を用いることで，評価関数を定義し直すことは可能であるが，他の集合間の距離は最長および最短距離法の間の性質を持つので，本手法は任意の集合間の距離を用いた場合の両端の性質を持つと考えられる．

区間データの区間の大きさを無限大に大きくしたものは，欠損値とみなすことができる．本章の手法において最長距離法を用いた場合には距離が無限大となり目的関数が定義できないが，最短距離を用いた場合は，目的関数が定まり，クラスタリングが実現できる．これは，欠損値を含むデータのクラスタリング手法 [45] で扱われている手法と等価になる．

目的関数を FCM のアルゴリズムで厳密に最小化しようとすると，クラスター中心に関する最適化を既存のファジィc-平均法と同様に単純に行うことはできない．目的関数のクラスター中心に関する偏導関数をみると，各々の成分に対し区分線形で単調増加な関数である．このことから目的関数はクラスター中心に関して凸であることがわかる．凸関数の最小値を求めるためには，単調増加である偏導関数が負から正に転ずる場所を見つければよい．

そこで目的関数のクラスター中心に関する偏導関数をそれぞれの成分に関して，探索するアルゴリズムを新たに開発した．このアルゴリズムを用いることにより，偏導関数が負から正に転じる場所が求められる．これは目的関数のクラスター中心に関する厳密な最適解が得られることを意味し，目的関数をファジィc-平均法のアルゴリズムで厳密に交互最適化することができ，区間データに対するファジィc-平均法が確立した．

数値実験を行うことにより，不確定性を含まない場合に既存のファジィc-平均法

を適用した結果と，区間データに対して本手法において最長距離法と最短距離法を用いた結果を比較した．バタフライデータに対する実験結果ではメンバシップ値がそれぞれの場合で異なり，実際に調査されたデータにおいては，クラスター中心の位置が異なった．

本章では，ユークリッド距離の二乗を用いたファジィクラスタリング手法を開発したが，他の距離を用いた場合の解析も考えられる．L_1 距離や，より一般的なミンコフスキー距離などがそれにあたる．L_1 距離を用いた場合には，目的関数のクラスター中心に関する偏導関数は，階段状の関数となり，本章で述べたアルゴリズムを単純化することにより，クラスター中心に関する最適解を得ることができる．このアルゴリズムは，L_1 距離を用いた不確定性を含まないデータに対するアルゴリズム [43] と類似した形になる．