第1章　序論

1.1　研究の背景

一般に分類と呼ばれる手法は，外的な基準を持つものと持たないものに分けられる．外的な基準を持つものとして判別分析などがあるが，分類の基準はデータの外にあり，その分類基準によって分類を行う．外的な分類基準を持つことから，教師あり分類と呼ばれる．もう一方の外的な基準を持たない分類手法は，教師なし分類と呼ばれ，本研究で扱うクラスタリングはこれにあたる．クラスタリングでは，外的基準を用いることなくデータ相互の距離の近さをもとに分類を行う．

クラスタリング (clustering) はクラスター分析 (cluster analysis) とも呼ばれ，データ解析の諸技法のなかで，外的基準なしに分類を行う方法のことである．外的基準がないため，分類すべき個体 (データ) の間に定義された類似性や距離に基づいて，グループ分けを行う．クラスタリングは，与えられたデータの集まりをいくつかのグループに分ける手法のことで [46]，多くの応用がなされている．

たとえば，数量分類学は生物学に応用されている分野であり，その中でクラスター分析が用いられている [55]．医学の分野では，患者を症候群のグループに分類したり [18, 29, 64]，バクテリアを分類 [21] する研究が行われている．考古学の分野では，発掘された石器具等を，クラスタリングを用いて類似のものからなるグループに分類する [30, 31, 32]．また，マーケティングの分野では，都市の商業的，人口学的な変数からなるデータを用いて，都市のグループ化が行われたり [20, 26]，商品の銘柄間の類似度をもとに銘柄に対するクラスタリングが行われている [36]．また，データマイニングにおいて，顧客データベースから類似部分集団の検出 (データベース・セグメンテーション) を行う際にクラスタリングが用いられている [7, 9]．

さて，クラスタリング技法には，階層的クラスタリングと非階層的クラスタリングの2種類がある．階層的クラスタリング [17] とは，グループが入れ子を構成するようにクラスを生成していく手法であり，非階層的クラスタリングでは，単にグルー

プが形成される．

非階層的クラスタリングには，グループの中心を考える手法と考えない手法がある．グループの中心を考えない手法の中では，グループ内の個体間の総当りの距離を考え，全てのグループに対する重み付き総和を評価関数とし，この評価関数を最小化する手法が代表的なものの一つである．

グループの中心を考える代表的なクラスタリング手法にc-平均法がある．c-平均法では，グループの中心と個体の分割を変数とする目的関数を最小化する．目的関数は，グループの中心とそれに属する個体の間の距離を求め，この距離を全ての個体に対して加え合わせたものである．本論文では，c-平均法の分割にあいまいさ(ファジィ)を導入したファジィc-平均法を取り扱う．

本研究は，クラスタリングにおいて不確定性を含むデータを扱っていることが特徴である．不確定性を含むデータの例として次の三つを挙げる．

1. 不確定性を含むデータ
   データユニット自身がもともと不確定性を含む場合である．データの入力項目がもともと幅を持った場合が考えられる．20-29のように年齢の20代を表す区間は，25歳で代表されるよりも20-29という区間のままで扱うほうがより自然である．また，人が何らかの質問に対して答える場合，曖昧さを含んだ回答を行う場合があり[61]，このようなデータに対しても曖昧さを含んだままデータを解析することは十分意義がある．
2. 誤差を含むデータ
   ある値がなんらかの測定機器で計測される場合，測定された各々の値は誤差を含んだ測定値となる．測定値は誤差を含まないものと仮定され，解析される場合も多い．しかしながら，測定値が誤差という不確定性を含んでいるのは事実であり，誤差を含んだ測定値をそのまま解析できれば，誤差をふまえた解析結果を得ることができる．ここでは誤差をファジィ数として扱うが，このような立場は他の研究者にもみられ，DuboisとPradeは次のように言っている[5]．"The calculus of fuzzy quantities derives from the introduction of the notion of a fuzzy set in the older area of interval analysis. Interval analysis defines rules for the propagation of errors (described by intervals) in calculation processes."
3. 個体における一部の成分が欠損しているデータ
   個体の一部の成分が欠損しているデータは欠損値を含むデータと呼ばれ，多く

の場合，データ解析において欠損値を含む個体は無視され，他の欠損値がない個体のみを用いて解析が行われてきた．データ解析の一分野であるクラスタリングにおいても通常，欠損値を含む個体は無視され，欠損値を含まない個体のみを用いてクラスタリングが行われる．しかしながら，欠損値を含む個体にも欠損値以外の情報が含まれており，この情報はクラスタリング結果に反映されるべきである．欠損値を含むデータは不確定性を含むデータとみなすことができる．個体のある成分の不確定性が無限大になったのが欠損値であるとみなすことにより，欠損値を含むデータを，不確定性を含むデータの一部であると考えることができる．

このようなデータは，世の中に存在し解析の対象となっており，不確定性を排除してから解析を行うのではなく，不確定性を含んだままデータ解析 (ここではクラスタリング) を行うべきである，というのがこの研究の根本的な動機である．

1.2　関連研究

不確定性を含むデータの解析については，ここで扱うファジィc-平均法以外にファジィ回帰分析がある．ファジィデータに対するファジィ回帰分析は以前から盛んに行われている．線形回帰モデルのパラメータがファジィ数で表されるようなファジィ線形回帰モデルにおいては，入力データに対応する出力データが与えられたとき，ファジィ線形回帰式を仮定して，回帰式のファジィ係数を推定する．ファジィデータに対するファジィ回帰分析は，坂和 [52]，田中 [62, 63] により研究されている．このため，ファジィデータに対するファジィクラスタリング手法も同様に考察されるべきである．

ファジィクラスタリングの分野では，不確定性を含まないデータに対するファジィc-平均法は，Bezdek[4] により提案されたものが最も標準的なものである．また，その変形についても考察がなされており，宮本ら [48] はエントロピー正則化を用いたファジィc-平均法を提案している．ファジィc-平均法では，クラスターの中心と，個体のクラスターへの所属の度合いを示すメンバシップの，2つの変数からなる目的関数を最小化することによりクラスタリングが行われる．一般的には，2つの変数を含む評価関数の最適化 (最小化) には，クラスター中心とメンバシップのそれぞれの変数に関して交互に最適化を収束するまで行う，2段階交互最適化 (ファジィc-平

均法のアルゴリズムと呼ばれる) が行われる.

ファジィクラスタリングでは, ユークリッド距離の二乗が多く用いられるが, L_1 距離を用いた場合の考察も行われている. 不確定性を含まないデータに対する L_1 距離に基づくファジィ c-平均法は, Bobrowski ら [8] や Jajuga [34] によって研究されている. また, 目的関数のクラスター中心に関する効率的な最適化アルゴリズムは, 宮本ら [43] によって提案されている.

さて, 不確定性を含むデータに対する研究に目を向けると, ユークリッド距離に基づくファジィ c-平均法において, 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. Hathaway ら [28] や Pedrycz ら [49] および Lee ら [39] の研究では, 不確定性を含むデータが扱われ, ユークリッド距離の二乗が含まれるファジィ c-平均法の目的関数を用いているが, 目的関数に対する厳密な交互最適化を行っていない.

L_1 距離に基づくファジィ c-平均法において, 不確定性を含むデータを厳密な交互最適化を用いてクラスタリングする研究も行われていない.

ここでは, 最も単純な性質を持つ不確定性を含むデータとして, 区間やファジィ数のデカルト積を考えている. 佐藤ら [53, 54] は, 各々の個体に超楕円体の不確定性を仮定したファジィクラスタリング手法を提案している. しかしながら, 一般に超楕円体を用いた解析は複雑で多くのデータを厳密に扱うには適さない.

また, 遠藤ら [16] は, 2つのファジィ集合間の距離を定義することにより, データが一般的なファジィ集合で与えられた場合の, ファジィ c-平均法に基づく新しいクラスタリングアルゴリズムを提案しているが, ここでも厳密な交互最適化アルゴリズムは示されていない.

1.3　研究の意義

いままでの研究では不確定性を含むデータを, いかに不確定性のないデータとして扱うかが考えられていた. 1つは, 不確定性をなくしてある代表値を決め, 不確定性のないデータとして扱うことによりクラスタリングを行う手法である. つまり, 不確定性がある場合には無いデータに変換して, クラスタリングを行うという視点に立っている. もう1つは, 不確定性を含むデータの数自体が少ないのであれば, 解析の対象から外してしまい, 残りの不確定性を含まないデータのみでクラスタリン

グを行うという視点である．しかしながら上の2つの考え方は，不確定性をあるがままにとらえるのではなく，不確定性を含むデータに何らかの方法を適用することにより，不確定性がないデータとして解析するという視点に立っている．実際，不確定性を含むデータの不確定性を解消，あるいは情報を持っているデータをそれ自身無視している．

不確定性を含むデータは，それ自身一つのデータとしての情報を含むのであり，またさらに不確定性というもう一つの情報を含んでいるとは考えられないだろうか．不確定性を含むデータを不確定性を保持したままクラスタリングを行うことにより，あるがままに解析の対象であるデータをとらえ，より正確な解析を行うことができるのではないか．

次に，不確定性を含むデータのクラスタリングを行おうとすると，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生じる．個体における不確定性は，個体の間の距離に不確定性が生じることを意味する．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が適切である．ところで，距離の上限と下限は，後に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，不確定性がクラスタリング結果に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，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．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．まず，目的関数に含まれるクラスター中心とファジィデータ間の距離を，最長距離法，最短距離法を用いて定義する．目的関数を交互最適化で最適化する際に，クラスター中心に関する最適解は，従来のファジィc-平均法と同様に求めることができない．そこで，新たにクラスター中心計算アルゴリズムを開発し，目的関数の厳密な交互最適化を実現する．

1.4　本論文の構成

第2章では，ファジィc-平均クラスタリングと距離空間について述べる．まず，ユークリッド距離に基づくc-平均法を紹介する．c-平均法 (c-means) は，クラスタリングの代表的な手法の一つであり，ユークリッド距離に基づく手法がよく用いられる．c-平均法では分類すべき個体が与えられたときに，c個のクラスター中心と，個体がクラスターに所属するか否かを表す2値変数 (以後メンバシップと呼ぶ) を含む目的関数を最小化することによりクラスタリングを行う．最適化はクラスター中心とメンバシップについて交互に最適化を行うアルゴリズムを用いて行われる．最適化の結果，クラスター中心と，メンバシップ (c個の分割) が得られる．

c-平均法のメンバシップに0から1のあいまいさを導入したのが，ファジィc-平均法である．c-平均法でただ単にメンバシップが0から1の値を連続的にとることができるように許容したとしても，交互最適化におけるメンバシップの解は0か1の値しか得られない．

Bezdek[4] は目的関数にパラメータmを導入し，標準的なファジィc-平均法の目的関数を定義した．この目的関数を最適化すると0から1の間の値をもつファジィなメンバシップ解が得られる．また，宮本ら [48] はファジィc-平均法はc-平均法の正則化ととらえ，c-平均法の目的関数にエントロピー正則化項を加えることにより，エントロピー正則化を用いたファジィc-平均法の目的関数を定義した．これらの手法を紹介する．

ファジィc-平均法の目的関数には個体とクラスター中心との間の距離が含まれており，ユークリッド距離の二乗が多く用いられる．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では，標準的手法においてもエントロピー正則化を用いた手法においても，目的関数の交互最適化におけるそれぞれの変数に関する最適解は単純な式で与えられる．この導出過程について述べる．

ユークリッド距離のかわりにL_1距離を用いた場合についても交互最適化のそれぞれの変数に対する最適解の導出が必要になる．メンバシップに関する最適解の導出は，ユークリッド距離の二乗を用いた場合と同様に行うことができ，最適解についても，同様の形の式で与えられる．しかしながら，クラスター中心の最適解はユークリッド距離に基づくファジィc-平均法のように簡単には求めることはできない．不確定性を含まないデータに対し，L_1距離を用いた場合の解に関して研究が行われており，この解について記述する．

不確定性を含むデータに対してファジィc-平均法の目的関数を定義する際に集合間の距離を用いる．不確定性を含むデータをファジィc-平均法で扱うためには，不確定性を含む個体とクラスター中心の間の距離を定義しなければならない．この距離を3種類の集合間の距離を用いて定義する．最長距離，最短距離，ハウスドルフ距離について述べる．ハウスドルフ距離は，最長距離と最短距離の間の性質を持つ．ところが，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している．この場合，片方の集合がただ一つの要素からなっている場合に相当し，ハウスドルフ距離と最長距離は一致することが示される．

第3章では，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を行うための新たなクラスター中心計算アルゴリズムについて述べる．

ここで扱うユークリッド空間は，ファジィクラスタリングにおいて多く用いられ，区間データは，不確定性を含むデータを考える場合に最も単純なものである．区間データは各々の成分が区間で表現され，その直積である区間の直積で表されると仮定する．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．

既存のファジィc-平均法を，区間データを扱えるように拡張することを試みる．その際，個体とクラスター中心との間の距離を定義する必要があり，最も極端な最長距離，最短距離の2種類の集合間の距離を用いて，この距離を定義する．他の集合間の距離を用いることも可能であるが，最長距離，最短距離は，他の距離の上限と下限をとることになる．定義された距離を用いて，ファジィc-平均法の目的関数を定義する．

定義された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が用いられる．目的関数を厳密に交互最適化するためには，それぞれの最適化ステップで，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．目的関数のクラスター中心に関する偏導関数を考えると，単調増加な区分線形関数であることが示される．クラスター中心に関する最適解を求めるためには，この区分線形関数が負から正に転じる点を求めればよく，この解を求めるために新たにアルゴリズムを開発

する．本アルゴリズムでは，偏導関数の区分点を順に探索し，偏導関数が0となる解を出力する．このアルゴリズムを用いることで，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，目的関数に対する厳密な交互最適化が実現される．また，アルゴリズムの計算量は$O(n)$であり，従来のファジィc-平均法におけるクラスター中心に関する最適解を求めるステップでの計算量も$O(n)$であることが示される．

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．

第4章では，ファジィデータに対するL_1距離に基づくファジィクラスタリングについて述べる．

まず，L_1距離を用いる手法を考察する動機について述べる．ファジィc-平均法では通常，ユークリッド距離の二乗が用いられる．個体をユークリッド空間の1点としてとらえることは，座標軸の回転を許容することであり，その便利さのために最も頻繁に用いられている．しかしながら，多変量の統計ではつねにユークリッド空間が用いられるとは限らず，マンハッタン空間と呼ばれるL_1空間もしばしば用いられている．

ところで，個体1つ1つに不確定性があると考える場合，ここで用いる仮定のようにファジィ数の直積で表現されるファジィデータが最も自然である．ところが，このようなファジィデータでは，座標軸を回転すると，不確定性を含むデータはもはやファジィ数の直積ではなくなる．したがって，回転に対して個体データの性質が不変ではないので，回転を許容しない空間を考えるほうがむしろ適切である．すなわち，L_1空間は不確定性を含むデータを扱う場合に，その仮定の自然さ，適切さにおいてユークリッド空間に劣ることはない．

次に，ここで扱うファジィデータについて記す．各々の個体がファジィ数の直積で表現されるようなファジィデータを考える．ファジィ数の中で，本論文では三角ファジィ数に関して考察を進める．任意のファジィ数に関しても同様に議論することは可能であるが，実際の計算が複雑になるので，ここでは三角ファジィ数を扱っている．

このファジィデータに対する目的関数が定義される．既存のファジィc-平均法の目的関数にファジィデータを適用しようとするとクラスター中心とファジィデータとの間のL_1距離を新たに定義する必要がある．ところで，ファジィ数の直積で表現

されるファジィデータのα-カットを考えると区間の直積となる．ファジィデータのαカットとクラスター中心の距離を最長距離と最短距離を用いて求める．この距離をαに関して0から1まで積分したものをファジィデータとクラスター中心との間の距離と定義する．この距離を用いて，ファジィデータに対する目的関数が定義される．

ファジィc-平均法の解は目的関数を最小化することにより求められる．これらの目的関数はクラスター中心とメンバシップの2つの変数を含んでおり，2変数の交互最適化アルゴリズムを用いて最小化を行う．メンバシップに関する最適化は，ファジィデータを扱う場合も不確定性を含まない通常のデータに対する既存のファジィc-平均法同様に最適解が導出される．しかしながらクラスター中心に関する最適解を同様に求めることはできない．そこで，目的関数のクラスター中心に関する偏導関数を考えると，区分線形な単調増加関数であることが示される．最適解を求めるには，この偏導関数が0となる解を求めればよい．

この最適解を求めるために三角ファジィ数の直積で表されるファジィデータに対するクラスター中心に対する厳密な最適解を求めるアルゴリズムを新たに開発する．このアルゴリズムでは，区分線形な偏導関数の区分点を順に探索し，偏導関数が0の値をとる場所を出力する．最長距離法と最短距離法を用いた場合の，クラスター中心に関する最適解を求めるアルゴリズムがそれぞれ示される．これらのアルゴリズムを用いることによりクラスター中心に関する厳密な最適解が得られ，ファジィデータに対するL_1距離に基づく目的関数を，厳密な交互最適化によって最適化できることが示される．なお，このアルゴリズムの最初の段階でソーティングを行うが，それを除いた計算量は$O(n)$である．既存のファジィc-平均法における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．

ある人物のイメージに関する不確定性を含む回答があり，この回答が三角ファジィ数の直積で表現されると仮定し，本章の手法を用いて数値実験を行った結果を示す．このデータが不確定性を含まないと仮定した場合に既存のファジィc-平均法を行った結果と，本数値実験の結果を比較する．