第1章

緒論

1.1 研究の背景

クラスタリング[1, 2, 3, 4]とは図1.1のようにいろいろ異なった性質のものがまざり合っている対象の中で，互いに似た者同士を集めてクラスタを作り，それらを分類しようとする方法である。これはクラスタ分析 (cluster analysis) とも呼ばれ，1950年代から本格的に研究が行なわれるようになり，様々な手法が提案されている。クラスタリングについてもう少し数学的に述べると，教師なし分類，すなわち外部的分類基準なしに，対象に関するデータの相互的類似度あるいは非類似度にもとづいて互いに近い対象は同じグループに入り，互いに遠い対象は違うグループに入るようにグループを生成する手法であるといえる。しかし分割に対する評価はクラスタリング手法によって異なるので，それぞれの手法によってその結果は異なったものになる。また外的基準が存在しないので，類似度・非類似度の定め方によりクラスタリング結果は大きく異なる。一般に類似度・非類似度はクラスタリングに用いる変数空間上のユークリッド距離によって定められることが多く，クラスタ分析にどの変数を用いるか，各変数を標準化するのかどうかという吟味もクラスタリング手法の選択とともに非常に大切である。一般的にクラスタリングでは，ユーザーが望むような（期待したよう

な) クラスタリング結果を得ることを目的とする。しかし，それは扱うデータや何のためにクラスタリングを行うのかといったことにより異なってくる。ユーザーは，それを考慮してクラスタリングを行う必要がある。

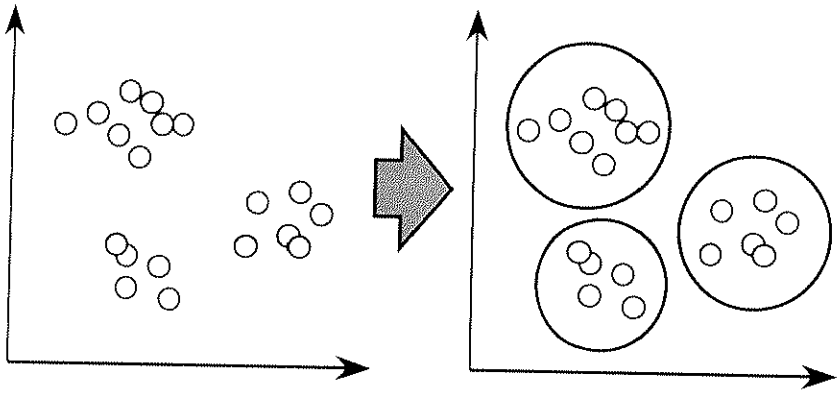

図 1.1: クラスタリングの概念図

大きく分けてクラスタリングには，階層的な方法と非階層的な方法とがある。階層的な方法は図 1.2 のような樹形図 (dendrogram) を得る方法で，とくにクラスタ数は決めず，対象の階層的構造を求めるものである。この方法は，目的に応じて，大分類から小分類までいろいろ利用できるところに特徴がある。一方，非階層的な方法は，あらかじめクラスタ数を定めておき，対象が属しているクラスタの重心との距離が最小となるという意味で，最良の分類を得ようとする手法である。

crisp k-means 法はクリスプ k-平均法とも呼ばれ，McQueen[5] や Duda and Hart[6], Forgy[7] らによって 1960 年代から 1970 年代に示された非階層的なクラスタリング手法である [1, 8]。本論文ではこの crisp k-means 法に，L.A.Zadeh によって 1965 年に提案されたしたファジィ理論 [9] を適用したファジィクラスタリング手法を扱う。クリスプなクラスタリングでは，クラスタリングによって各対象は必ず 1 つのクラスタに属するが，ファジィクラスタリングでは対象はそれぞれのクラスタに属する度合いを定めることによってクラスタを表現する。(図 1.3 参照) これによって，各個体のより自然な分割を考えることができる。さらに第 3 章で導入する分類関数によって，各クラスタをファジィ集合とした，自然な表現を考えることができる。

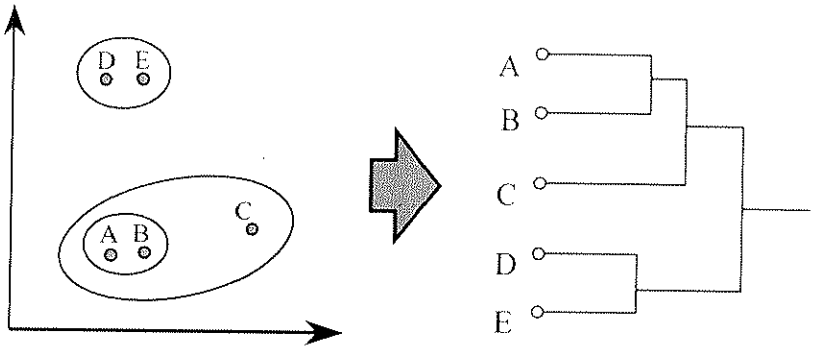

図 1.2: 樹形図 (dendrogram) の例

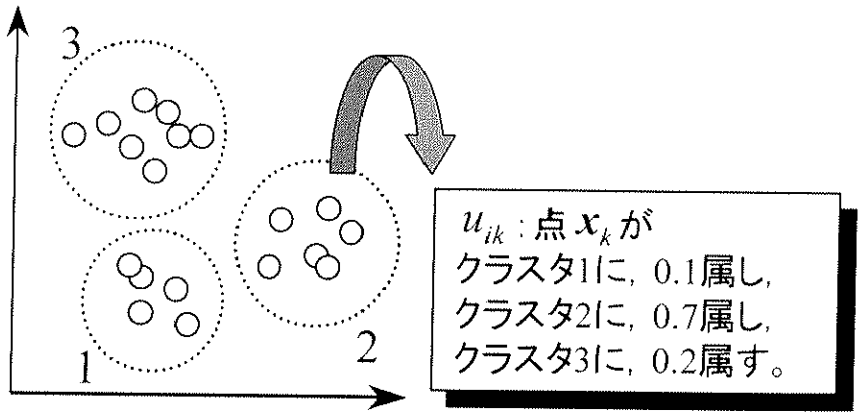

図 1.3: ファジィクラスタリングの概念図

現在，crisp k-means 法をファジィ化したファジィクラスタリングの手法として，Bezdek らによる fuzzy c-means 法 [10, 11, 12] が広く用いられている。本論文はこの fuzzy c-means 法について，第 I 部・第 II 部に以下の 2 つの観点から，研究成果をまとめたものである。

1.2　正則化によるファジィクラスタリング

第 I 部では，fuzzy c-means 法を crisp k-means 法の正則化であるという視点に立って fuzzy

c-means 法を考察する。正則化の概念は関数方程式の「適切でない問題 (ill-posed problem)」を解くための手法 [13] として，従来より多くの現実的問題に用いられてきた。一般に正則化とは，ある意味で特異な問題から正則なものへの修正を意味する。特異な問題は解くことが難しいが，正則なものは扱いやすい。正則化された問題の解が元の問題の解の近似となっているとき，その問題を元の問題の正則化という。

クラスタリングの場合 crisp k-means 法は通常の意味での「適切でない問題」とはいえないが，crisp 解をあたかも特異であるかのようにみなし fuzzy 解を正則であるように考えることによって，fuzzy c-means 法は crisp k-means 法の正則化であると考えることができる。これによって，crisp k-means 法の別の正則化によるファジィクラスタリング手法を考えることができる。本論文では，新たな2種類の異なる正則化によるクラスタリング手法を提案し，さらにそれらの手法についてクラスタリング結果を示すとともに，分類関数を導入してそれぞれのクラスタリング手法を比較検討する。

クラスタリングは「教師なし分類」であり，その結果については一般的に正解・不正解は存在していない。しいていえば，正解が存在しているとすれば，それはユーザーが望むような（期待したような）クラスタリング結果のことことである。その意味で crisp k-means 法のファジィ化にはこれまで，標準的な fuzzy c-means 法しか存在しなかったが，いくつかの方法が存在するということは，ユーザーが期待することのできるクラスタリングに対する選択の幅を広げることになり，クラスタリングの目的に応じて手法の使い分けをすることが可能となる。しかしこれはまた，クラスタリングを行うとき，どの手法を用いるのがよいのかということが新たな問題となる。本研究では，これに対する回答として，各クラスタリング手法の異なる性質について新たに分類関数を導入することによって各種法の性質を調べた比較研究を行っている。

1.3 線形ファジィクラスタリング

第II部では，多次元空間上に分布するデータのモデリングを行うために，多次元空間上のデータの部分線形構造をとらえたファジィクラスタリングを考える。(図 1.4 参照) 部分

構造モデルは，人間が把握しやすい部分的線形構造とし，大規模なデータからのいくつかの線形部分空間の発見を目的とする。

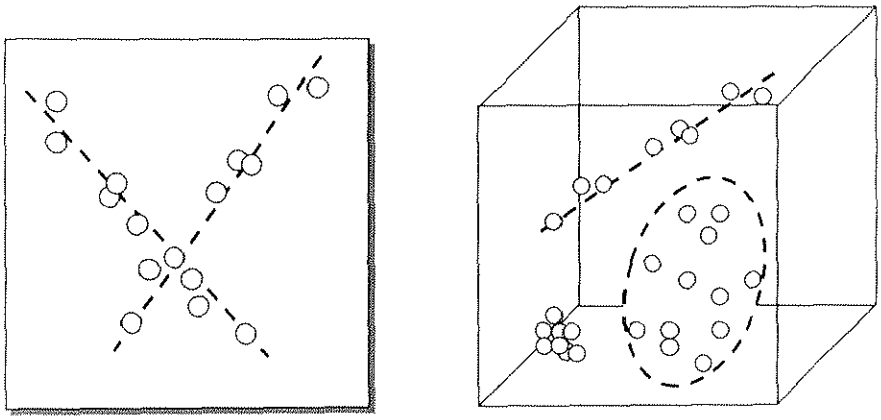

図 1.4: 2 次元及び，3 次元上の線形構造によるクラスタリングの図

多次元上に分布するデータ解析の手法には，データ全体を 1 つの集合ととらえるものと複数のグループととらえるものにより，分布モデル，回帰分析，主成分分析等の手法や，判別分析，各種クラスタリング手法等がある。そして扱っているデータや問題に応じ，アナリストは各種手法を使い分けてデータの性質を探ってきた。しかし扱っているデータに十分な知識がないときや，データが複雑で各変数間の関係を予測できない大規模な問題に，これらの手法を適用するのは困難である。これはデータ全体を 1 つの集合ととらえられないが，判別には距離だけでなく線形関係も考慮し，さらにそれぞれの分布により変数選択も行わなければならないという，複雑に関係したこれらの解析手法にある。本研究ではこの問題をふまえ，データを複数の線形関係を考慮したクラスタリングを行い，そのデータ構造を提示することで問題構造の解明し，アナリストの主観に大きく依存しないモデル作りを目指す。

これに関連するクラスタリングに関する研究は，海外では盛んに行われている。しかし海外での研究目的は，画像認識のための技術として研究されていることが多い。そのため二次元上のデータが主な対象となっている。代表的なファジィクラスタリング法である fuzzy c-means 法の評価規範にクラスタのファジィ散布行列の固有値・固有ベクトルを用いる fuzzy c-varieties 法が，1981 年に Bezdek らにより提案された。これはクラスタの構造に線形関係があると仮定し，複数の線形構造を発見することのできるクラスタリング手法である。その

後 fuzzy c-means 法と fuzzy c-varieties 法をうまく融合させることにより，さらに実用的な fuzzy c-elliptotypes 法や adaptive fuzzy c-elliptotypes 法が Bezdek や Dave らにより開発されている。また，クラスタのファジィ散布行列の代わりに，データの線形構造を発見するために，クラスタに線形回帰モデルを用いる fuzzy c-regression models 法等の手法も報告されている。

本論文では，まず新しい線形ファジィクラスタリング手法，及びそれを用いたモデリングに関する提案を行う。また，第I部で述べた正則化の概念を線形ファジィクラスタリングの次元係数に用いた手法についての提案を最後に行う。これによって，これまでの手法にあった理論的な問題点を解消することができる。

1.4 本論文の構成

各章の概要は，以下の通りである。

第2章では，この論文を通して基礎となるファジィクラスタリング手法である fuzzy c-means 法について簡単に述べる。これはクリスプな crisp k-means 法がファジィ化された手法であり，ファジィクラスタリング手法の代表的な手法で，現在広く用いられている手法である。この手法ではクラスタリングのあいまいさを表すスムージングパラメータを導入し，この値を変化させることにより，crisp k-means 法に近い結果から，あいまいさの度合いの大きな結果まで，様々な結果を得ることができる。

本章では，この crisp k-means 法と，これがどのようにファジィ化されているかについて簡単に述べる。これを crisp k-means 法の正則化であると考え，続く第3章，第4章では，この fuzzy c-means 法とは異なった2種類のファジィクラスタリング手法を考える。

第3章では crisp k-means 法に正則化項としてエントロピー関数を持つエントロピー正則化法を考える。エントロピーにもとづく方法として，李と向殿よるエントロピー最大化法 [14, 15] が提案されているが，正則化の概念を導入することによって，エントロピーを用いる方法を，一般的な fuzzy c-means 法の交互最適化アルゴリズムの枠において議論できるようになる。このことは，数多く研究されている fuzzy c-means 法の変形をエントロピー正則

化を用いて行うことができることを意味している。

また，従来の fuzzy c-means 法から生成されるファジィプロトタイプ分類関数に対応するプロトタイプ分類関数がエントロピー正則化から得られる。これら 2 種類の分類関数の理論的性質を調べ，計算幾何学における Voronoi 図との関連を明らかにする。2 つのクラスタリング手法によって得られる分類関数は，クラスタの中心や無限遠方において大きな違いがあることが示される。しかし，そこから得られるクラスタの領域は等しいものとなり，クラスタの中心を含む Voronoi 図と一致することが示される。さらに数値例によっても，これら 2 つの方法によるクラスタリング結果を示し，ファジィ分類関数のこれらの性質を確認する。

第 4 章では，この正則化のアイデアを更に進め，crisp k-means 法に正則化項としてメンバシップの 2 次関数を用いる 2 次正則化法を提案する。この手法では，クラスタの中心の計算は通常の fuzzy c-means 法と似ているが，メンバシップ値の計算は異なった形となる。ここでは，メンバシップ計算のための効率のよいアルゴリズムを提案する。

さらにエントロピー正則化法同様に，ファジィ分類関数を 2 次正則化手法によって求めることができ，fuzzy c-means 法やエントロピー正則化法のファジィ分類関数の性質を比較する。この 2 次正則化法から得られる分類関数は区分線形型となるという著しい性質があることが示される。その分類関数は，標準的な fuzzy c-means 法のものよりはエントロピー正則化法のそれに近いが，ある有限の範囲にあるクラスタの中心にしか影響されないという意味で，より crisp k-means 法に近いということができる。数値例によって，これらの方法によるクラスタリング結果を示し，ファジィ分類関数のこの性質を確認する。

第 II 部，第 5 章では，ファジィクラスタリング手法をデータ解析の手法としてとらえ，fuzzy c-means 法を応用した線形ファジィクラスタリング手法について議論する。クラスタリングの規範としてはユークリッド距離を関連性の尺度として用いられることが多い。しかしクラスタの構造に線形や非線型の何らかのモデルを仮定することによって，データの構造を考慮したクラスタリング手法も多く報告されている。本章ではデータが持つ部分線形関係に注目し，データの線形構造をとらえたファジィクラスタリング手法を考える。こういった線形クラスタリングに関するこれまでの研究は，画像認識へ応用するためのものが多く，2 次元上のデータを対象とした研究が中心に行われてきた。本研究ではデータ解析の見地からクラスタリング手法をとらえ，多次元上のデータに対して異なる次元の線形構造をとらえた線形ファジィクラスタリング手法の提案を行う。

さらに観測データには誤差やはずれ値が含まれていることが多いが，これらの問題についても考える。これに対応するための手法として，ノイズクラスタ [16, 17] が Dave らによって提案されているが，これを用いた線形ファジィクラスタリング手法について考察を行う。ノイズクラスタは，一般のクラスタのいずれにも分類されないデータのためのクラスタでゴミ箱クラスタとも呼ばれている。この章で新たに提案したクラスタリング手法にノイズクラスタを適用することにより，簡単な数値例を用いて，クラスタリング結果が誤差やはずれ値の影響を低く抑えることができたことを示す。

第6章ではさらに，ファジィクラスタリングによるモデリングを行うために，楕円型メンバシップ関数で表される新しいタイプのファジィモデルを提案する。著者は，大規模システムのファジィモデルによるモデリングとシミュレーションに関する研究 [18, 19, 20] を行ってきた。このときの大きな問題の1つは，モデルを構築するためにはその対象をよく知る必要があるということである。しかしシステムを理解することはモデリングの目的でもあり，「卵が先か，鶏が先か」という問題である。これに対して中森はその著書 [21] の中で次のように述べている。大規模システムのモデリング目的の1つは，そのシステムをより理解することにあり，これは互いに相反する要求である。これに対して中森はその著書 [21] の中で次のように述べている。「モデリングは対象の認識を表現する行為であり，そして表現を用いて思考することにより，さらに認識を深めることができる。科学の発達はこのフィードバックに支えられてきた。（中略）大事なことは，そのプロセスの中でいかに思考支援がなされるかである。」

ここではファジィモデリング [22] における課題として，システムの部分的な構造をとらえたサブモデル群をいかに発見するかを考え，ファジィクラスタリング手法を用いた新たなファジィモデリングおよびシミュレーション手法を提案する。従来のファジィモデリングでは，事前に変数間の構造モデルを仮定する。しかし，複雑で大規模なシステムに対しては，構造を探求するためにデータからモデルを構築してみるという側面がある。また多くの場合，数値データは一様に存在せず，変数の選択や，サブモデルの担当領域の決定が困難である。こういった状況でモデリングを実行するためには，対象をよく理解する必要がある。この章で提案する楕円型ファジィモデルは，変数間の関係をデータに忠実に表現しようとするもので，最終モデルというよりは，対象をよりよく理解するための道具である。実際のモデリングの例として，日本の河川に関する適用例を紹介する。

第7章では最後に，第5章で提案した線形ファジィクラスタリング法の次元係数の決定に，第3章，第4章で用いた正則化の概念を導入した新たな定式化による手法を提案する。1990年にDaveによって提案された適応型 (adaptive) 手法では，次元係数の決定を各クラスタのファジィ散布行列の固有値を用いて決定する。これによってクラスタの構造をとらえたよい結果を得ることができる。第5章で提案した手法もこの考えを応用したものである。しかしこれはファジィクラスタリングの2段階アルゴリズムの考え方に合致しない面があり，目的関数の単調現象性が失われていることを指摘する。また，その具体的な例を示す。そして次元係数の決定に正則化の概念を導入することにより，その問題を解消する。また，本手法によるいくつかの数値例によるクラスタリング結果を紹介する。

なお，第3章は文献[23, 24]，第4章は文献[25, 26]，第7章は文献[27, 28]，第6章は文献[29, 30]，第5章は文献[31]の成果に，それぞれもとづいている。