題目：スペクトラルクラスタリングの応用 -誤差のあるデータへの適用-

氏名：尾本　倫太郎

指導教員：大沼進（中島晃）

クラスター分析は、多変量データを構成する各個体をその類似度に基づいて複数のグループに分類する多変量解析の手法である。その分析手法は天文学、医学、考古学、市場調査、社会学、心理学などの幅広い領域においても利用されており、多変量データに潜在するクラスターの構造を抽出する事が可能となる。

クラスター分析は大きく階層的クラスター分析法と非階層的クラスター分析法に分けられ、階層的クラスター分析法では最短距離法や Ward 法、非階層的クラスター分析法では k-means 法や Fuzzy c-means 法、混合分布モデルがよく知られている。

近年、機械学習や画像処理などの分野で使用される機会が増えているスペクトラルクラスタリングは、2000 年代に提案されたクラスター分析手法の一つであるが、グラフ理論に基づいて分類を行うという点で従来の分析法とは異なる。またスペクトラルクラスタリングは従来の k-means などの手法では分類することが難しかった複雑な形状のクラスターを分類することができるという特徴がある。

しかしながら、実験や研究のデータを解析する際に誤差の影響により正しく分類できないというケースも存在する。特に人文科学系の実験では比較的サンプルサイズが小さく誤差が大きいため、その影響を強く受ける傾向にある。

本論文ではスペクトラルクラスタリングを誤差の大きなデータに適用するための方法を提案する。スペクトラルクラスタリングの処理過程において通常用いられる k-means ではなく Fuzzy c-means を適用する。Fuzzy c-means は k-means 法とは異なり、各個体が複数のクラスターに属する。その際、k-means では各個体はクラスターへの所属度を 0 か 1 かで表すのに対し、Fuzzy c-means では各個体は各クラスターへの所属度を 0~1 の間の数値で確率的に表される。誤差の影響等により複数のクラスターに重なりが生じた場合などに、k-means 法は推定が不安定になりがちになるが、Fuzzy c-means はクラスターが重なった部分の個体が複数のクラスターに所属するため、k-means と比較して正確に推定できる。本研究では、まず人工データを従来のスペクトラルクラスタリングに適用して誤差の影響を検証し、次に誤差の大きなデータに提案する手法を適用して Fuzzy c-means を利用することの有用性を検証した。