混合効果モデルに基づく隣接情報を考慮した関数データクラスタリング

岡　響希

本論文では、時間経過とともに繰り返し測定して得られる経時測定データを対象としたクラスタリングにおける新たな手法の開発を行い、その有効性を検討した。

経時測定データを解析する統計手法として関数データ解析がある。経時測定データを、時間に対して連続的に変化する関数としてとらえたデータが関数データである。従来の多変量データを対象とした統計手法を、関数データを対象に拡張した手法が関数データ解析である。中でも関数データクラスタリングは、分析対象となる関数データをデータ間の距離や類似性などの基準に基づいて分類する手法である。

関数データクラスタリングの解析対象となる経時測定データは、測定の際に位置情報を得られることがある。クラスタリングを行う上で空間関係を考慮することはデータの解釈に新たな示唆を与える可能性がある。これに着目して、データ同士が隣接関係にあるかどうかを考慮した非階層型の関数データクラスタリングの新手法の開発を試みた。

本論文では先行研究を参考に、関数データクラスタリングを適用する上で、関数データをB-スプライン基底関数による基底関数展開によって表現した。そして、各対象の関数データが、所属するクラスタごとの固定効果と、個体差を表す変量効果を持つとする混合効果モデルによって表現されると仮定した。加えて、各対象がクラスタの個数と等しい数の成分を持つ混合分布に従って生成されるとする混合モデルをもとに、関数データクラスタリングのモデル化を行った。

そして、各対象から最も近い距離にある異なる対象を「隣接している」として定義した。提案手法では、隣接関係にあるデータの推定クラスタが一致している場合にのみ作用するように正則化項を付与してパラメータ推定を行うようにした。また、正則化項を付与する際に、その影響力を決定する変数を与えるようにした。

提案手法を用いて、隣接関係を反映したクラスタリングが可能であるかを検討するために、人工データを生成してシミュレーションを行った。また、クラスタリングの精度を比較するため、従来の関数データクラスタリング手法でも同様のデータに適用した。クラスタリングの評価指標として、真のクラスタと推定クラスタとの一致率と類似度を用いた。

その結果、いくつかの状況において、従来の関数データクラスタリング手法より精度良くクラスタの推定を行うことができた。ただし、推定精度の向上度合いは限定的であり、その原因としては、隣接関係の定義、正則化項が作用する条件、正則化項の影響力の決定方法、シミュレーションにおけるサンプルサイズと生成データの特徴などが関連していると考えられる。（行動統計科学）