

題 目 隣接関係に基づくクラスター分析 -クラスター数決定に関する考察-

氏 名 高橋 直也

指導教員 大沼 進

クラスター分析は、データをグループ間では類似せず、グループ内では類似するグループ(クラスター)に分類するのに用いられる有用な多変量解析手法の 1 つであり、多変量データに潜在する個体のクラスター構造とその特徴を抽出するための探索的な方法として利用される。近年では、調査や実験等から得られたデータだけでなく機械学習や画像処理など広範な分野に、クラスター分析の応用が広がっている。クラスター分析は、階層的手法・非階層的手法に大別され、さらにそれぞれに多くの手法が含まれるが、一部の例外を除きクラスターと個体の距離について最適化することにより分類を行う。このため、分類の結果得られるクラスターの形状は、(超)楕円体またはそれに近い形状に制限される。また、既存の(とくに非階層的)クラスター分析手法の多くに、データを何個のクラスターに分けるかを解析者が明示的に示さなければいけないという共通する特徴がある。したがって、分析者は手法の適用に先立ち、クラスター数を指定しなければならない。クラスター数を選択する方法はこれまでもいくつか提案されている(e.g., Thorndike, 1953; von Rousseeuw, 1987)が決定的なものは無く、主観的にクラスター数を決定することが多い。本論文は、クラスターの形状を限定しない隣接関係に基づく分析方法に注目し、隣接関係を記述する隣接行列の改善を提案するとともに、適切なクラスター数の決定の問題を解決するために、グラフラプラシアン固有ベクトルのギャップを用いたクラスター数決定方法である Eigenvector Gap Method (EGM) に関する検討を行う。提案手法は、ドロネー図により類似度グラフを構築したグラフの隣接関係をもとにしてデータの並び替えを行い、クラスター構造を明らかにする。そして、グラフラプラシアンの固有ベクトルのギャップを用いて、最終的なクラスター数の決定とクラスターへの割り当てを行う。シミュレーションの結果、苦手とするデータは存在するものの、提案手法により解析者が明示的にクラスター数を示さなくても、多様なデータに対して内的に自然なクラスター数の決定ができることを示された。