# 凸計画に基づくスペクトラル・クラスタリング手法

水谷 友彦
東京工業大学 工学院 経営工学系

クラスタリングは，与えられたデータを似ているものは同じグループに，似ていないものは異なるグループに分類するというタスクである．スペクトラル法はこのタスクに対して有効であると経験的に知られており，実際に多くの実問題に対して利用されている．スペクトラル法ではデータの分類をグラフの分割問題として定式化する．具体的には，データ集合をグラフとして表現する．データはグラフの頂点として表し，2 つのデータ間の類似度が高いとき対応する頂点間に辺を置く．クラスタリングのタスクは，このようなグラフを辺で密に繋がっている部分（頂点集合の部分集合がクラスターのように見える部分）ごとに分割する問題とみなすことができる．

このような分割を求めるために，スペクトラル法では 2 つのステップを実行する．最初のステップではグラフに付随する行列（具体的にはラプラシアン）の固有ベクトルを計算する．次のステップでは得られた固有ベクトルを利用して，グラフの頂点をユークリッド空間上の点に写す．そして，得られた点集合に対して $k$ 平均法などの既存のクラスタリング手法を適用し，頂点集合を幾つかの部分集合に分類するというものである．現在では，ステップ 2 において $k$ 平均法を用いるのが，スペクトラル法の標準的な実装となっている．

2014 年度の COLT において Peng らは $k$ 平均法に基づくスペクトラル法の性能を理論的に評価した．グラフの頂点集合の部分集合に対して定まるコンダクタンスという量は，そのクラスターらしさを定量化するために用いられる指標である．Peng らはクラスタリングのタスクをコンダクタンスの最大値の最小化問題（コンダクタンス問題）として定式化した．そして，その問題の最適解とスペクトラル法の出力がどのくらいの近さにあるのかを解析した．その解析において，彼らはコンダクタンス問題の最適値とラプラシアンの $k+1$ 番目に小さい固有値の比によって定まる値 $\Upsilon_G$ を導入した．$\Upsilon_G$ はグラフが $k$ 個のクラスターらしい部分を持つことを表す指標となっている．特に，$\Upsilon_G$ の値が大きいグラフは，$k$ 個のクラスター構造が明確に認識できるようなグラフである．彼らは，$\Upsilon_G$ が大きいとき，スペクトラル法の出力はコンダクタンス問題の最適解に近くなり，$\Upsilon_G$ が大きくなるに従ってそれらは接近することを示した．

本研究では，スペクトラル法のステップ 2 において楕円を利用する手法を考案し，その性能を考察した．理論的な成果として，$\Upsilon_G$ がある閾値を超えれば，考案手法の出力とコンダクタンス問題の最適解は一致するという結果を得た．加えて，考案手法と既存手法である $k$ 平均法に基づくスペクトラル法の性能を実験的に評価した．$k$ 平均法の出力は初期値の与え方に影響を受ける．したがって，$k$ 平均法を複数回実行して，得られた結果の平均を取った．実験結果から，考案手法の性能は少なくとも既存手法の性能の平均値と同程度であるという示唆を得た．