次元縮約とクラスタリングの同時達成手法の研究開発

松永　和磨

多変量データ解析には解析対象となるデータと目的に合わせて様々な手法が存在する．多次元のデータを低次元空間に近似する主成分分析と，個体を特定数のクラスター（群）に分類するクラスター分析はその中でも代表的な 2 手法である．行列分解型の分析法は行列近似問題に帰結するため，分布の仮定を必要とせず標本そのものの情報を得たいときに有効となる．例えば，データ行列の行が「国・県であるなどの有限母集団の全て」である場合や，データ行列そのものが特定のブランドの印象評定データのように「個体と変数のどちらにも興味がある」場合などには非常に有用である．

しかし，そうした利点に対し，応用領域においてはこういった分析がそぐわない場合がある，そうした場合，しばしば Tandem Clustering といった手順が取られる．これは[1]データに対し目的（例えば次元縮約など）に即した解析法を適用し，その分析結果を得る，[2]得られた分析結果に対し，K-means 法(MacQueen, 1967)を始めとするクラスター分析を適用し個体を分類する，という 2 段階の手続きである．例えば，変数の数が大きく，そのうちのいくつかはクラスター構造に寄与していないと考えられるデータに対して，主成分分析を行った後に，その主成分得点を利用してクラスタリングを行う，などが挙げられる．しかしタンデムクラスタリングは簡便である反面，[1]の段階で得られた結果がクラスター構造を反映するとは限らず，寄与する次元をうまく抽出できないとして批判されている（De Soete & Charrol, 1994）．実際に Vichi & Kiers (2001)は数値実験によって，主成分分析におけるタンデムな分析は失敗に終わることを例証している．

タンデムクラスタリングにおいて，こうした問題が生まれる原因は，異なる目的の分析手法のために，それぞれ別個の目的関数を取り扱うことにあるとされている．その対処として特定の分析手法とクラスター分析法を同一の目的関数に組み込み，同時に行う手法が提唱されている． 代表として，主成分分析とクラスター分析を同一の目的関数によって表現し，かつ，同時に最適化する手法である Reduced K-means 法(De Soete ¥& Carroll,　1994)と Factorial K-means 法(Vichi & Kiers, 2001)が挙げられる．この 2 つの手法は，目的関数を統一することで低次元空間でのクラスタリングを行っており，先述の問題点を解決している．しかし，Reduced K-means 法の目的関数には得点行列が存在せず，クラスター中心は推定されるが個体得点が推定されないという問題が存在する．またこれらのアルゴリズムによって導かれる主成分負荷量行列は，必ずしも解釈が容易になるとは限らない．また，Factorial K-means 法は特定のデータに対して有用なクラスタリングを行うが，その他のデータに対するクラスタリング精度は非常に低くなってしまっている，という問題が存在する．

以上のように，次元縮約とクラスター分析を組み合わせ，解釈に容易な結果を目標とした手法が考えられてきたが，本研究では，それらの先行研究の問題点を改善するモデルを提案した．提案手法では 2 つのモデルを同一の目的関数に組み込み，それらをチューニングパラメータを用いて重みづけすることで，より柔軟なデータへのフィッティングを目指した．

提案手法によって 2 種の実データを解析し，既存手法との比較実験を行った結果，Iris データ(Fisher,1936)を用いた正分類率の比較では既存手法を大きく上回る正分類率を示し，OECD の経済指標データ(Vichi & Kiers, 2001)を用いた散布図比較では，視覚的に提案手法の優位性が明らかに見て取れた．以上のように先行研究の問題点の解決と，クラスタリングの精度の良さが示され，提案手法の有用性が例証された．（行動統計科学）