集約的シンボリックデータの変数選択

統計数理研究所　清水 信夫
統計数理研究所　中野 純司
徳島文理大学　　山本 由和

1　集約的シンボリックデータ

大量の個体をもつ多変量データが自然なグループに分かれている場合、そのような各々のグループに関しての推論に興味がある場合が考えられる。このとき、グループを表すためのいくつかの記述統計量の集合をデータと考えたものを集約的シンボリックデータ (Aggregated Symbolic Data, ASD) と呼ぶ。実際のデータにおいては多数の連続変数と多数のカテゴリー変数が両方含まれている場合がよく見られる。われわれはこのような状況において、連続変数およびカテゴリー変数のいずれの場合に対しても同じ基準で集約的シンボリックデータ間の非類似度を考え、そのクラスタリングを行う方法を提案してきた。

個体データは連続変数とカテゴリー変数の両方を含むものとし、それがいくつかのグループに分けられているとする。各グループにおいて、各々の変数、および異なる 2 つずつの変数の組に関して、2 次までのモーメントを用いて求められる記述統計量を考え、それらの集合を ASD と考える。ASD を形成する記述統計量は、連続変数同士に関しては平均や分散共分散行列、カテゴリー変数同士であればBurt 行列などである。

異なる ASD 間の非類似度は、異なる 2 変数より求められる非類似度統計量の値を、全ての場合の組み合わせについて考えた時の総和として表す。

2　集約的シンボリックデータにおける各変数間の従属性

カテゴリー変数において、ほとんどすべてが 1 つのカテゴリー値を取るようなものにはあまり情報が含まれない。また、連続変数間に相関の高いものがあるように、カテゴリー変数間にも従属性の高いものがある。そのような変数は情報としては冗長であり、それらをそのまま用いれば非類似度がその変数の存在の分だけ強調されてしまう。そこで、データ全体でばらつきが極めて小さな変数、従属性が極めて高い変数の組み合わせのうちの片方は冗長な変数として削除する必要がある。

カテゴリー変数についてはカテゴリー値が順序尺度の場合 (順序変数) と名義尺度の場合 (名義変数) が存在する。順序変数同士の従属性を表す指標としてはポリコリック相関係数、順序変数と連続変数との相関を表す指標としてはポリシリアル相関係数がそれぞれ知られている。

ここでは、名義変数を含む組み合わせの従属性および相関についてはカテゴリー値の順序を並べ替えた上で最も当てはまりの良い並びにおけるポリコリック相関係数およびポリシリアル相関係数を考える。その上で、それらの係数の 2 乗が 1 に近い値となる組み合わせが見つかった場合、両変数の意味を吟味した上で変数を 1 つ除去することで、冗長な組み合わせを減らすことにする。そのようにして選択された変数を用いて異なる ASD 間の非類似度を求め、それらを用いたクラスタリングを行う。詳細および適用例は当日に示す。