カテゴリー変数を含む集約的シンボリックデータのカイ2乗統計量

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　はじめに

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。変数が連続変数値データのみの場合は、各グループをその平均および分散共分散行列を用いて表すことが考えられる。実際のデータにおいては連続変数の他にカテゴリー変数も含まれている場合が多数ある。このような状況において、連続変数・カテゴリー変数いずれに対しても同じように非類似度を考えるために、それぞれの 2 次のモーメントまでを考え、連続変数をカテゴリー変数に変換して全ての変数をカテゴリー変数とみなす。そしてその Burt 表を ASD と考える。本報告では 2 つの Burt 表の同一性を検定するためのカイ 2 乗統計量を非類似度とみなし、これを用いて ASD 間のクラスタリングを行う。

2　集約的シンボリックデータ間のカイ2乗統計量

連続変数をカテゴリー変数に変換するために、連続変数値が 1 個または 0 個含まれるような極めて小さな等間隔の区間を考えたものをカテゴリー値とし、その場合の出現確率を連続変数の分布を正規分布と考えたときの確率密度関数で近似する。カテゴリー変数の出現確率はカテゴリー値ごとにパラメータとして設定する。このようにするとグループ g について

$$
\left[\begin{array}{ccc|c|ccc}
s_1^{(g,1)} & & & & s_{11}^{(g,1q)} & \cdots & s_{1m_q}^{(g,1q)} \\
& \ddots & & \cdots & \vdots & \ddots & \vdots \\
& & s_{m_1}^{(g,1)} & & s_{m_1 1}^{(g,1q)} & \cdots & s_{m_1 m_q}^{(g,1q)} \\
\hline
& \vdots & & \ddots & & \vdots & \\
\hline
s_{11}^{(g,q1)} & \cdots & s_{1m_1}^{(g,q1)} & & s_1^{(g,q)} & & \\
\vdots & \ddots & \vdots & \cdots & & \ddots & \\
s_{m_q 1}^{(g,q1)} & \cdots & s_{m_q m_1}^{(g,q1)} & & & & s_{m_q}^{(g,q)}
\end{array}\right]
$$

という、異なる 2 変数ごとの分割表の組み合わせである Burt 表を考えることができる。$s_{m_{k_1} m_{k_2}}^{(g,k_1 k_2)}$ は変数の組 (k_1, k_2) におけるカテゴリー値がそれぞれ i_{k_1} および i_{k_2} である場合のセルの値である。

各 ASD における Burt 表の中の各分割表においては、各分割表内および共通変数をもつ分割表間で周辺分布に関する制約があり、それらの制約を全て考慮した確率分布モデルに基づく尤度比検定統計量を考えるのは難しい。ただ、2 つの分割表において各セルごと値が示されているとき、これに対して同一性の仮説の下で 2 つの ASD を合わせた場合の各セルの予測値も考えられる。そこでこれらからカイ 2 乗統計量を求めることができる。本報告ではそれらの総和を非類似度とみなし、以前に考察した疑似的な尤度比検定統計量の総和を用いた場合との比較を行う。詳細および適用例は当日に示す。