2つの名義変数間の相関係数

清水　信夫　　データ科学研究系　助教

• 研究の背景および動機

連続(実数)変数とカテゴリー変数が混在する大規模多変量データにおいて、自然に分けられた集団が存在し、それらに関する情報に興味がある場合を考えたい

- 各集団ごとに変数のいくつかの記述統計量(平均、分散、etc.)の集合をデータと考えて解析⇒**集約的シンボリックデータ（Aggregated Symbolic Data, ASD）**と呼ぶ
- 各変数ごとの性質だけでなく、2つの変数間の関係を表す記述統計量についても考えたい
  - 2つの連続変数同士であれば**Pearson相関係数**が定義されている
  - 2つのカテゴリー変数同士の各カテゴリー値ごとの組み合わせは分割表で表され、これより2つのカテゴリー変数間の相関に相当する統計量が考えられる
  - 順序変数同士の場合における分割表から求められる関係として**ポリコリック相関**、順序変数と連続変数の組み合わせにおける関係として**ポリシリアル相関**が存在
    ⇒Pearson相関係数などとの整合性がない
    ⇒名義変数の場合が考慮されておらず、カテゴリー値の全ての並べ替えの中での最大相関を考えると計算時間が長くなる
  - 2つの名義変数同士の値について、既存の相関係数と対応が取れる形の指標を定義し、それとほぼ同等な値を現実的な時間で導出することを考えたい

• カテゴリー変数同士の相関の定義

2つの異なるカテゴリー変数についてそれぞれのスコアを考え、それらの間の相関が最大となる場合をカテゴリー変数間の相関とする
⇒連続変数における、標準化された数値のPearson相関に対応

集団gにおける2つの名義変数k_a, k_bのダミー変数行列をそれぞれ$X^{(g,k_a)}, X^{(g,k_b)}$とすると$X^{(g,k_a)\prime}X^{(g,k_b)} = N^{(g,k_ak_b)}$は2変数の分割表となる。
ここで$a = \left[a_1 \cdots a_{m_{k_a}}\right]', b = \left[b_1 \cdots b_{m_{k_b}}\right]'$としてスコア$X^{(g,k_a)}a$とスコア$X^{(g,k_b)}b$の相関が最大となる場合を考える。
スコアに条件をつけない場合、名義変数同士の相関は確率行列$P^{(g,k_ak_b)} = N^{(g,k_ak_b)}/n^{(g)}$を標準化した行列を特異値分解した場合の最大特異値$\lambda_1$となる。
それを与える(a, b)はλ_1に対応する最大特異ベクトルの組(a_1, b_1)を用いて$(a, b) = (D_{k_a}^{-1/2}a_1, D_{k_b}^{-1/2}b_1)$と求まる。
（D_{k_a}, D_{k_b}は、それぞれ$P^{(g,k_ak_b)}$のk_a, k_b方向の周辺分布ベクトル$\boldsymbol{v}_{k_a}, \boldsymbol{v}_{k_b}$の各成分を対角成分とする正方行列とする）

分割表に順序変数が含まれる場合は、順序変数について順番を固定したまま、相関に対応する値および各変数のスコアを名義変数同士の相関と同様に考える。
2つのカテゴリー変数のうち順序変数が1つだけでそれをk_aとするとき、対応するスコアを線形増加数列を成分とするベクトル$a_{ns} = \left[1,2,\cdots,m_{k_a}\right]'$に関して

$$\tilde{a} = \frac{(I_{m_{k_a}} - \mathbf{1}_{m_{k_a}}\boldsymbol{v}_{k_a}')a_{ns}}{\sqrt{a_{ns}'(D_{k_a} - \boldsymbol{v}_{k_a}\boldsymbol{v}_{k_a}')a_{ns}}}$$

と変換した形で表し、2変数間の相関が最大となるようにもう1つの変数のスコアbを一意に計算できる。
2つとも順序変数のときは、それぞれのスコアを線形増加数列の変換の形で表し、相関を求めることができる。

• 名義変数が含まれる場合の相関についての検証と修正、および発見的なアルゴリズム

名義変数においてはカテゴリー値の並び順は任意に決定可能だが、相関を考えるには全ての並びの中での特定の並び順の場合における順序変数同士の場合と一致させないとカテゴリー変数を含む全体の様々な場合を考える上で整合性が取れない⇒スコアに何らかの条件をつける必要がある

- どんな条件をつけるか？
  ⇒条件をつけない場合において求められた最適なスコアにおける成分$(D_{k_a}^{-1/2}a_1, D_{k_b}^{-1/2}b_1)$がそれぞれ昇順となるようにカテゴリー値の順番を並べ替える
  ⇒成分の中のいくつかの要素が同じ値をもつ場合、成分$(D_{k_a}^{-1/2}a_2, D_{k_b}^{-1/2}b_2)$、$(D_{k_a}^{-1/2}a_3, D_{k_b}^{-1/2}b_3)$、… を調べてそれらが昇順となるように並べ替える
  ⇒条件をつけない場合の並べ替え済みの最適なスコアを、単調増加数列を成分とするベクトルの変換により求められたスコアで置き換える
- これにより名義変数のカテゴリー値の最適な並べ替えをした場合の相関を、順序変数同士の相関と同様に考える
- 最適と考えられる並べ替えを見つける時間を大幅に短縮できる
  ⇒これによる並べ替えが真の値（＝カテゴリー値の全ての並べ替えの中での最大値）を与えることは保証されていないが、多くの場合で妥当な近似値が得られる

• 名義変数同士の相関係数の導出例

以下は名義変数同士の相関係数の例である。それぞれの名義変数のカテゴリー値の数はそれぞれ6および7である。この2つの名義変数における元々の分割表をN_1およびN_2、それらのカテゴリー値の並びをそれぞれランダムに並べ替えたものをN_{1r}およびN_{2r}とする。それぞれの分割表においてカテゴリー値の全ての並べ替えのうち最大の値をρ_{all}、発見的アルゴリズムにより求めた値をρ_{heu}とする。ここでN_1はブロック対角行列だが、N_2はブロック対角行列ではない。
この時にN_1とN_{1r}、N_2とN_{2r}はそれぞれ適切な並べ替えにより同じ相関係数を持つはずである。これらの分割表に関してρ_{all}を計算すると、どの場合でも数分程度かかる（カテゴリー値の数が多いほど時間がかかる）。一方、ρ_{heu}に関しては長くとも数秒程度で計算できる。値についてはブロック対角行列の場合（もしくはそれに近い場合）には適切な並べ替えの場合との間で差異が出やすい傾向があるが、適切な並べ替えを行ってもブロック対角行列に近い形とならないような場合は、発見的なアルゴリズムによる導出でも適切な並べ替えの場合とほぼ同じ値となる。
全ての並べ替えにおいて相関係数を計算する場合に各名義変数におけるカテゴリー値の数の階乗の積に比例した計算時間がかかることを考えると、提案した発見的アルゴリズムは妥当な時間で相関係数の近似値を求めるには有効と考えられる。

$$N_1 = \begin{bmatrix} 30 & 10 & 0 & 0 & 0 & 0 & 0 \\ 10 & 30 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 20 & 10 & 0 & 0 & 0 \\ 0 & 0 & 10 & 20 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 20 & 30 & 10 \\ 0 & 0 & 0 & 0 & 10 & 30 & 20 \end{bmatrix} \quad \rho_{all} = 0.94,\ \rho_{heu} = 0.93$$

$$N_2 = \begin{bmatrix} 30 & 10 & 0 & 5 & 10 & 0 & 5 \\ 10 & 30 & 5 & 0 & 5 & 10 & 0 \\ 5 & 0 & 20 & 10 & 6 & 0 & 8 \\ 0 & 2 & 10 & 20 & 0 & 4 & 0 \\ 8 & 4 & 6 & 2 & 20 & 30 & 10 \\ 0 & 6 & 2 & 4 & 10 & 30 & 20 \end{bmatrix} \quad \rho_{all} = 0.62,\ \rho_{heu} = 0.62$$

$$N_{1r} = \begin{bmatrix} 0 & 10 & 0 & 20 & 0 & 0 & 0 \\ 30 & 0 & 0 & 0 & 0 & 10 & 0 \\ 0 & 20 & 0 & 10 & 0 & 0 & 0 \\ 0 & 0 & 30 & 0 & 10 & 0 & 20 \\ 10 & 0 & 0 & 0 & 0 & 30 & 0 \\ 0 & 0 & 30 & 0 & 20 & 0 & 10 \end{bmatrix} \quad \rho_{all} = 0.94,\ \rho_{heu} = 0.93$$

$$N_{2r} = \begin{bmatrix} 0 & 10 & 0 & 20 & 8 & 5 & 6 \\ 10 & 5 & 0 & 0 & 5 & 30 & 10 \\ 4 & 2 & 30 & 6 & 10 & 8 & 20 \\ 2 & 20 & 4 & 10 & 0 & 0 & 0 \\ 30 & 0 & 10 & 5 & 0 & 10 & 5 \\ 6 & 40 & 30 & 2 & 20 & 0 & 10 \end{bmatrix} \quad \rho_{all} = 0.62,\ \rho_{heu} = 0.62$$

図1. 名義変数同士の分割表における全ての並べ替えによる相関係数(ρ_{all})および発見的なアルゴリズムによる相関係数(ρ_{heu})の例