カイ2乗統計量に基づく 集約的シンボリックデータの非類似度

2018年6月15日　統計数理研究所　オープンハウス

清水　信夫　　データ科学研究系　助教

【研究の背景】

近年の計算機科学の発展により、大規模かつ複雑な多変量データ集合が多数出現している。それらを記述、解析する上でデータ構造を柔軟に定義した枠組みとしてDidayにより提案されたシンボリックデータ (SD)があり、それらを解析する枠組みとしてシンボリックデータ解析 (SDA)が提唱されている。

最近の大規模多変量データ集合では、連続（実数）変数とカテゴリー変数が混在する場合が多く、また特徴的な属性に関し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が少なからず存在する。この場合、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する方法が考えられるが、これらのデータを我々は**集約的シンボリックデータ**（Aggregated Symbolic Data, ASD）と呼ぶ。

連続変数とカテゴリー変数が混在するデータ集合において2つのASD間の非類似度を考える場合、連続変数を離散化してあたかもカテゴリー変数であるかのように考えることで、全体をカテゴリー変数のみからなるデータ集合とみなして各集団間の非類似度を一貫した基準で考えられる。本報告では、連続変数を適当な有限個の区間に分割してそれらの区間をカテゴリー値とするカテゴリー変数とみなし、2つずつの変数の分割表に関するASD間のカイ2乗統計量を非類似度と考え、その基準を実データに対して適用した例を示す。

【変数型が混在する大規模データにおける集団の表現】

p個の連続型変数およびq個のカテゴリー変数（カテゴリー変数kにおける カテゴリー値の数は m_k個）のデータ集合Xのうち、集団 g におけるデータ行列 $X^{(g)}$を下記のように表す。

$$X^{(g)} = \begin{bmatrix} \boldsymbol{x}_{\boldsymbol{11}}^{(g)} & \cdots & \boldsymbol{x}_{\boldsymbol{1p}}^{(g)} & \boldsymbol{x}_{\boldsymbol{11}}^{(g,1)} & \cdots & \boldsymbol{x}_{\boldsymbol{1m_1}}^{(g,1)} & \cdots & \boldsymbol{x}_{\boldsymbol{11}}^{(g,q)} & \cdots & \boldsymbol{x}_{\boldsymbol{1m_q}}^{(g,q)} \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ \boldsymbol{x}_{\boldsymbol{n^{(g)}1}}^{(g)} & \cdots & \boldsymbol{x}_{\boldsymbol{n^{(g)}p}}^{(g)} & \boldsymbol{x}_{\boldsymbol{n^{(g)}1}}^{(g,1)} & \cdots & \boldsymbol{x}_{\boldsymbol{n^{(g)}m_1}}^{(g,1)} & \cdots & \boldsymbol{x}_{\boldsymbol{n^{(g)}1}}^{(g,q)} & \cdots & \boldsymbol{x}_{\boldsymbol{n^{(g)}m_q}}^{(g,q)} \end{bmatrix}$$

$n^{(g)}$個のデータをもつ$X^{(g)}$ において、左のp列がp個の連続変数値、それ以外がq個のカテゴリー変数ごとのダミー変数値である。連続変数およびカテゴリー変数に対しては、異なる2変数間の関係の確率モデルを2次モーメントまでの範囲で定義する。

【2つのカテゴリー変数の組み合わせに関する非類似度】

$X^{(g)}$から生成された各ASD g における異なる2つのカテゴリー変数の組み合わせは分割表として表され、全ての組み合わせに関する分割表をまとめたものがBurt行列として表される。ここでの各セルにおける値 $s_{j_1 j_2}^{(g,k_1k_2)}$はカテゴリー変数の組 (k_1, k_2) における各カテゴリー値の組が (j_1, j_2) となる場合の生起数である。ASD g_1およびg_2 が同じ性質をもつ場合、分割表のセル(j_1, j_2)の出現個数の期待値の推定量は

$$\widehat{E(s_{j_1 j_2}^{(g_a,k_1k_2)})} = \frac{s_{j_1 j_2}^{(g_1,k_1k_2)} + s_{j_1 j_2}^{(g_2,k_1k_2)}}{n^{(g_1)} + n^{(g_2)}} n^{(g_a)} \quad (a = 1, 2)$$

と考えられる。一方で 2つのASDが異なる場合は$s_{j_1 j_2}^{(g_a,k_1k_2)}$ および $\widehat{E(s_{j_1 j_2}^{(g_a,k_1k_2)})}$ を用いてカイ2乗統計量を

$$\chi^{2(g_1g_2,k_1k_2)} = \sum_{\substack{a=1 \\ s_{j_1 j_2}^{(g_1,k_1k_2)} + s_{j_1 j_2}^{(g_2,k_1k_2)} \geq 1}}^{2} \sum_{j_1=1}^{m_{k_1}} \sum_{j_2=1}^{m_{k_2}} \frac{\left\{ s_{j_1 j_2}^{(g_a,k_1,k_2)} - \widehat{E(s_{j_1 j_2}^{(g_a,k_1k_2)})} \right\}^2}{\widehat{E(s_{j_1 j_2}^{(g_a,k_1k_2)})}}$$

と考えることができ、これを $k_1 < k_2$なる全ての (k_1, k_2) に関し考え総和をとった

$$d_{(cc)}^{(g_1g_2)} = \sum_{k_1=1}^{q-1} \sum_{k_2=k_1+1}^{q} \chi^{2(g_1g_2,k_1k_2)}$$

がBurt行列におけるASD間の非類似度と考えられる。

【連続変数を含む組み合わせに関する非類似度】

連続変数lについては、$-\infty = h_0^{(l)} < h_1^{(l)} < \cdots < h_{N-1}^{(l)} < h_N^{(l)} = \infty$ なる $h_j^{(l)}$ $(j = 0,1,\cdots,N)$ を境界値とするN個の区間に分割し、各区間をカテゴリー値として考えると、他の変数との組み合わせについてカテゴリー変数同士の分割表と同様に考えることができる。すなわち、連続変数同士の組み合わせにおいては集団全体の個数に各領域ごとの出現確率を掛けた値、連続変数とカテゴリー変数の組み合わせにおいてはカテゴリー変数における各カテゴリー値ごとの全体の個数に各区間ごとの出現確率を掛けた値を、分割表のセルの個数の近似値として考えることにより、カテゴリー変数同士の組み合わせの場合と同じくカイ2乗統計量の近似値を計算することができる。

連続変数同士の組み合わせにおけるカイ2乗統計量を$\chi^{2(g_1g_2,l_1l_2)}$ 、連続変数とカテゴリー変数の組み合わせにおけるカイ2乗統計量を $\chi^{2(g_1g_2,lk)}$とすると

および

$$d_{(rr)}^{(g_1g_2)} = \sum_{l_1=1}^{p-1} \sum_{l_2=l_1+1}^{p} \chi^{2(g_1g_2,l_1l_2)}$$

$$d_{(rc)}^{(g_1g_2)} = \sum_{l=1}^{p} \sum_{k=1}^{q} \chi^{2(g_1g_2,lk)}$$

がそれぞれの組み合わせの全体の非類似度と考えられる。
なお、連続変数の分割数Nの値の定め方については、適当な範囲内でいくつかの場合に対する結果を求め、その中で適当なものを選べばよい。

連続変数をカテゴリー化して考えることにより、$d_{(cc)}^{(g_1g_2)}, d_{(rr)}^{(g_1g_2)}, d_{(rc)}^{(g_1g_2)}$ は全てカテゴリー変数同士の組み合わせにおける非類似度と考えられるため、この総和

$$d^{(g_1g_2)} = d_{(cc)}^{(g_1g_2)} + d_{(rr)}^{(g_1g_2)} + d_{(rc)}^{(g_1g_2)}$$

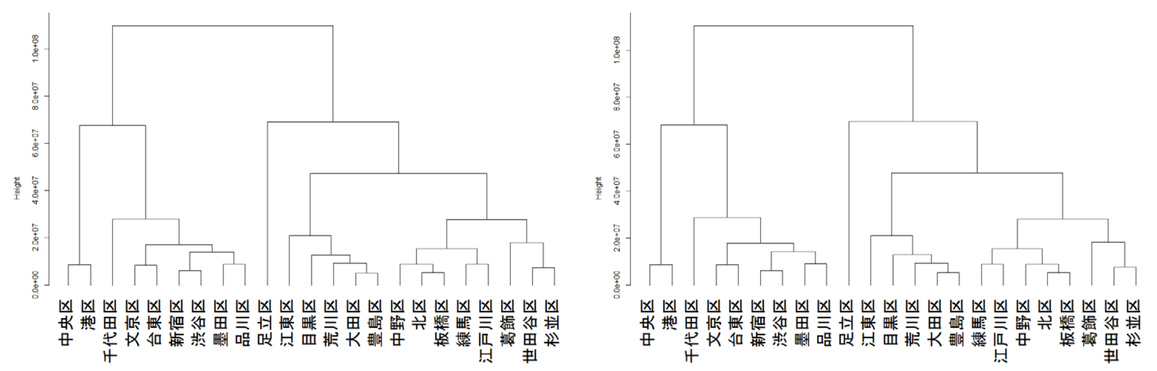
がASD間の全体のカイ2乗統計量に基づく非類似度と考えることができる。

【不動産情報データへの適用例】

表1はある不動産検索サイトにおける2013年時点の東京23区の賃貸住宅データ（有効総件数が約79万件）の一部である。このデータは5種類の連続型変数および76種類のカテゴリー変数を含む。このデータをカテゴリー変数“区”に関して物件が所在する区別に23の集団に分け、各々のASD間の非類似度を計算して最長距離法による階層的クラスタリングを行った。連続変数の分割数Nが4および5の場合の結果を図1に示す。

No.	区	賃料	面積	・・・	敷金月数	部屋数	所在階	・・・	即入居
1	荒川区	8.25	26.83	・・・	0.00	1	3	・・・	可
2	目黒区	5.10	9.96	・・・	1.00	0	2	・・・	可
⋮	⋮	⋮	⋮	・・・	⋮	⋮	⋮	・・・	⋮
4594	港区	37.20	114.93	・・・	2.00	3	5	・・・	不可
⋮	⋮	⋮	⋮	・・・	⋮	⋮	⋮	・・・	⋮
17277	中央区	27.00	89.72	・・・	2.00	3	14	・・・	可
⋮	⋮	⋮	⋮	・・・	⋮	⋮	⋮	・・・	⋮
494893	板橋区	5.50	20.16	・・・	0.50	1	5	・・・	可
⋮	⋮	⋮	⋮	・・・	⋮	⋮	⋮	・・・	⋮

表1: 不動産検索サイトにおける東京23区の賃貸住宅データ（一部）



(a) N=4　　(b) N=5

図1: 23区の相互間の非類似度に基づく階層的クラスタリング結果

図1より、いずれのNでも(1)中央区と港区の組(2)千代田区や新宿区など7区(3)足立区(4)世田谷区や板橋区など8区(5)大田区や荒川区など5区　の5つのクラスターが形成されているとみることができる。

この方法では連続変数の領域をN個のカテゴリー値として表すためNの値により結果に差異が起こり得るが、不動産情報データに関してはカテゴリー変数の数が連続変数の数よりも圧倒的に多いため、カテゴリー変数同士のみからの非類似度から導出される構造が大きく影響し、連続変数を含む部分の構造があまり影響していないと考えられる。