変数間の相関係数に基づく 集約的シンボリックデータの変数選択

2019年6月5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多数存在し、かつ混在するデータ集合では、値の多様性が極めて小さい変数や、類似性が極めて高い変数の組み合わせが出現することが往々にして起こる。これらの変数は解析上冗長であり、除去することでデータ集合の特徴をより適切に捉えつつ高速に解析を行えるようになる。本報告では、順序変数同士の相関として定義されているポリコリック相関係数を、名義変数や限られた記述統計量の情報のみが解っている連続変数を含む相関にも拡張する。これにより冗長な変数の適切な発見および削除のための材料を提供し、ASDを用いたデータ集合のクラスタリングの例を示す。

【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。

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

$n^{(g)}$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。

【ASD間の非類似度】

$X^{(g)}$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。2つのASDg_1 およびg_2 が同じ性質をもつ場合の分割表の各セルの出現個数の期待値を考え、それに基づいたカイ2乗統計量 $\chi^{2(g_1g_2,k_1k_2)}$ を求める。これを全ての組み合わせについて総和をとったカイ2乗統計量 $d_{(cc)}^{(g_1g_2)} = \sum_{k_1=1}^{q-1} \sum_{k_2=k_1+1}^{q} \chi^{2(g_1g_2,k_1k_2)}$ 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ついては、適当な個数に分割した各区間をカテゴリー値とみなし、各区間もしくは各領域ごとの確率にカテゴリー値もしくは集団全体の個数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

$$d_{(rr)}^{(g_1g_2)} = \sum_{l_1=1}^{p-1} \sum_{l_2=l_1+1}^{p} \chi^{2(g_1g_2,l_1l_2)} \quad \text{および} \quad d_{(rc)}^{(g_1g_2)} = \sum_{l=1}^{p} \sum_{k=1}^{q} \chi^{2(g_1g_2,lk)}$$

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。

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。すなわち、冗長な変数が多いと非類似度が必要以上に大きくなり計算時間も長くなる。

【ASDにおける各変数間の相関の表現】

カテゴリー変数を含む異なる2変数間の相関については、カテゴリー値が順序尺度であるカテゴリー変数同士に関してポリコリック相関、順序尺度をもつカテゴリー変数と連続変数の生データの間でポリシリアル相関がそれぞれ既に定義されている。ただしASDにおいては、カテゴリー変数が名義尺度をもつ場合が存在し、連続変数についても限られた記述統計量の情報のみで定義しているため、異なる変数間の相関を考える上でこれらの相関はそのままでは使えない。そこで、名義変数や連続変数を含む組み合わせの相関を以下で定義する。

【名義変数を含む組み合わせに関する2乗相関係数】

名義変数の場合、変数内のカテゴリー値の順番は定まらない。そこで、カテゴリー値の全ての順列ごとにポリコリック相関を考え、正規分布の当てはまりが最も良くなる場合の値を採用する。相関については正負の区別をせず2乗値を考える。

【連続変数とカテゴリー変数の組み合わせに関する2乗相関係数】

ASDでは連続変数lを平均 μ_lおよび分散σ_{ll} で表しているため生データの利用によるポリシリアル相関は使えないが、連続変数lの領域をカテゴリー変数kのカテゴリー値の総数m_k個の領域に分割し、各カテゴリー値ごとのi番目の領域確率 $p_{ji}^{(l)}$と各カテゴリー値の総数n_jの積を各セルの生起数とする分割表を考える。これにポリコリック相関を適用し、他の変数の組み合わせの場合と同様に考えるため2乗値を使用する。

【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2種類の連続型変数および62種類のカテゴリー変数を含むが、相関が非常に高く冗長と考えられる変数の組を調べ、変数の意味を考慮した上で適切に除去することを考える。そこでデータ全体に対し、カテゴリー変数を含む組み合わせについて2乗相関係数を計算した。表2はその中における、カテゴリー変数同士の相関が高い組み合わせの例を示したものであり、この結果を元にカテゴリー変数の意味を考慮して「新築」「管理費有無」など5つの変数を除去した。

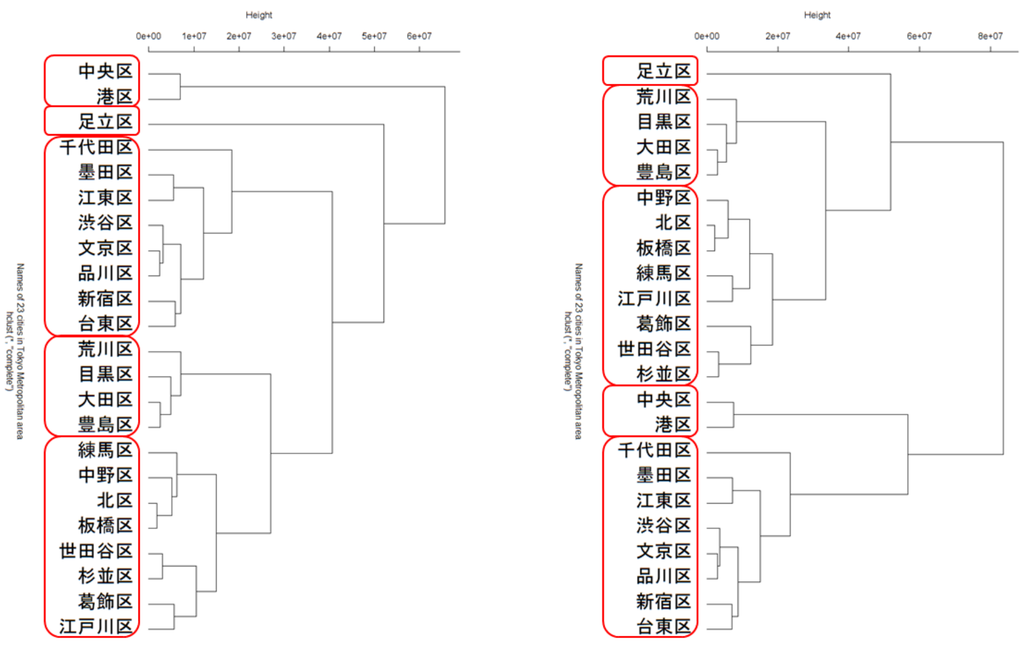
表1: 不動産検索サイトにおける東京23区の賃貸住宅データ（一部）

No.	区	賃料	面積	物件種別	構造種別	・・・	管理形態
1	荒川区	8.25	26.83	マンション	鉄筋コン	・・・	記載なし
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4588	港区	22.30	71.28	マンション	鉄筋コン	・・・	巡回管理
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
498088	足立区	6.40	33.34	アパート	軽量鉄骨	・・・	記載なし
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
714202	新宿区	16.40	55.64	マンション	鉄骨鉄筋	・・・	常駐管理
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

表2: カテゴリー変数同士の相関が高い組み合わせの例（一部）

変数 1	変数 2	2乗相関係数
新築	築年数	0.9515
管理費有無	管理費区分	0.9047
エレベーター	地上階建て数	0.8794
エレベーター	物件種別	0.8670
オートロック	宅配ボックス	0.8214
エレベーター	構造種別	0.8033
⋮	⋮	⋮

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。連続変数の分割数を5とし、カテゴリー変数が57個(5個削除)の場合および62個(元データのまま)の場合の結果を図1に示す。



(a) 57個　　(b) 62個

図1: 元データから冗長なカテゴリー変数を除去した場合およびそのまま利用した場合それぞれの階層的クラスタリング結果

図1より、いずれの場合でもデンドログラムを構成する主要なクラスターの構造には大きな変化がないと考えることができる。また(a)においては(b)より計算時間も短くなっており、冗長な変数の適切な除去による効果が現れていると考えられる。