カテゴリー変数が含まれる場合の集約的シンボリックデータのモデリング

2017年6月16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ASDの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、さらに2つずつの変数の組み合わせが従う2変量確率モデルから導出される集団間の尤度比検定統計量(LRTS)を非類似度と考えることで、カテゴリー変数のみからなるデータ集合として各集団間の非類似度を一貫した規準で考えることができる。

本報告では、データ集合において全ての変数がカテゴリー変数化された場合のASDが従う確率モデルを考え、ASD間の非類似度に関する性質について考察し、得られた結果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。

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

$n^{(g)}$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。

【集団間の非類似度の考え方】

異なる集団 g_1 およびg_2 の間の非類似度の定義を以下の手順で定める。

1. 各集団ごとに2変数間の確率モデルについて最尤推定量を考える
2. 連続変数の定義域を極めて微小な幅となる多数の区間に分割し、各区間における1つの個体の生起数が1もしくは0となるように考え、取り得るカテゴリー値(＝微小区間)が極めて多くスパースなカテゴリー変数と考える
3. g_1 およびg_2 に関し共通の2変数間の確率モデルの2種類の最大対数尤度を全ての組み合わせについて以下の通り考える
   - 同一パラメータモデル（g_1および g_2のパラメータが同じ値）の最大対数尤度 $\hat{l}_0$
   - 個別パラメータモデル（g_1および g_2のパラメータが違う値も可）の最大対数尤度 $\hat{l}_1$
4. 各々の組み合わせごとにLRTS= $-2(\hat{l}_0 - \hat{l}_1)$ を計算してそれらの総和を非類似度とする

この手順により、2つの集団間の非類似度はカテゴリー変数のみからなるデータ集合における異なる2つずつの変数の組み合わせのLRTSの総和として求めることができる。

【異なる2つのカテゴリー変数の組み合わせ】

各集団 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$s_{i_{k_1} i_{k_2}}^{(g,k_1 k_2)}$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$(i_{k_1}, i_{k_2})$となる場合の生起数である。Burt表は以下のように表現される。

$$\begin{bmatrix} s_{1}^{(g,1)} & \cdots & s_{1}^{(g,12)} & \cdots & s_{1m_2}^{(g,12)} & \cdots & s_{11}^{(g,1q)} & \cdots & s_{1m_q}^{(g,1q)} \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ s_{m_1}^{(g,1)} & & s_{m_1 1}^{(g,12)} & \cdots & s_{m_1 m_2}^{(g,12)} & \cdots & s_{m_1 1}^{(g,1q)} & \cdots & s_{m_1 m_q}^{(g,1q)} \\ & & s_{1}^{(g,2)} & & & \cdots & s_{11}^{(g,2q)} & \cdots & s_{1m_q}^{(g,2q)} \\ & & \vdots & & & & \vdots & & \vdots \\ & & & & s_{m_2}^{(g,2)} & \cdots & s_{m_2 1}^{(g,2q)} & \cdots & s_{m_2 m_q}^{(g,2q)} \\ & & & & & & & & \\ & & & & & & s_{1}^{(g,q)} & & \\ & & & & & & & & \\ & & & & & & & & s_{m_q}^{(g,q)} \end{bmatrix}$$

ここで各分割表ごとに各セルの出現確率 $p_{i_{k_1} i_{k_2}}^{(g,k_1 k_2)}$が多項分布に従うと仮定する。Burt表内の各セルの値は、各分割表内および共通変数をもつ分割表間でそれぞれ制約があり、出現確率に関する各分割表ごとの尤度関数を全て独立に考えるのは厳密には適当ではないが、出現確率の疑似最尤推定量 $\hat{p}_{i_{k_1} i_{k_2}}^{(g,k_1 k_2)} = s_{i_{k_1} i_{k_2}}^{(g,k_1 k_2)} / n^{(g)}$は出現確率が満たすべき条件を満足している。これを用い、異なる2つのカテゴリー変数の組み合わせにおける疑似LRTSの、全ての組み合わせに関する総和を集団間の(全体の)非類似度と考える。

【自動車データへの適用例】

表1は2004年に米国で販売された世界各国の自動車のうち約400台についてのデータの一部である。このデータには10種類の連続型変数および4種類のカテゴリー変数が含まれる。このデータをカテゴリー変数”Country”に関して製造元の本社が所属する国別に6つの集団に分け、各々のASD間の非類似度を計算して階層的クラスタリングを行った結果を図1に示す。

Vehicle Name	Country	Price	・・・	Length	Type	・・・	Drive
Chevrolet Aveo 4dr	US	11690	・・・	167	Sedan	・・・	front
Hyundai Santa Fe GLS	Korea	21589	・・・	177	Sedan	・・・	front
Saab 9-5 Aero	Sweden	40845	・・・	190	Wagon	・・・	AWD
Honda Odyssey LX	Japan	24950	・・・	201	Mini Van	・・・	front
Nissan Murano SL	Japan	28739	・・・	188	Wagon	・・・	rear
Jaguar XKR coupe 2dr	UK	81995	・・・	187	Sports Car	・・・	rear
BMW X3 3.0i	Germany	37000	・・・	180	SUV	・・・	AWD
・	・	・	・・・	・	・	・	・

表1: 2004年に米国で販売された世界各国の自動車データ（一部）

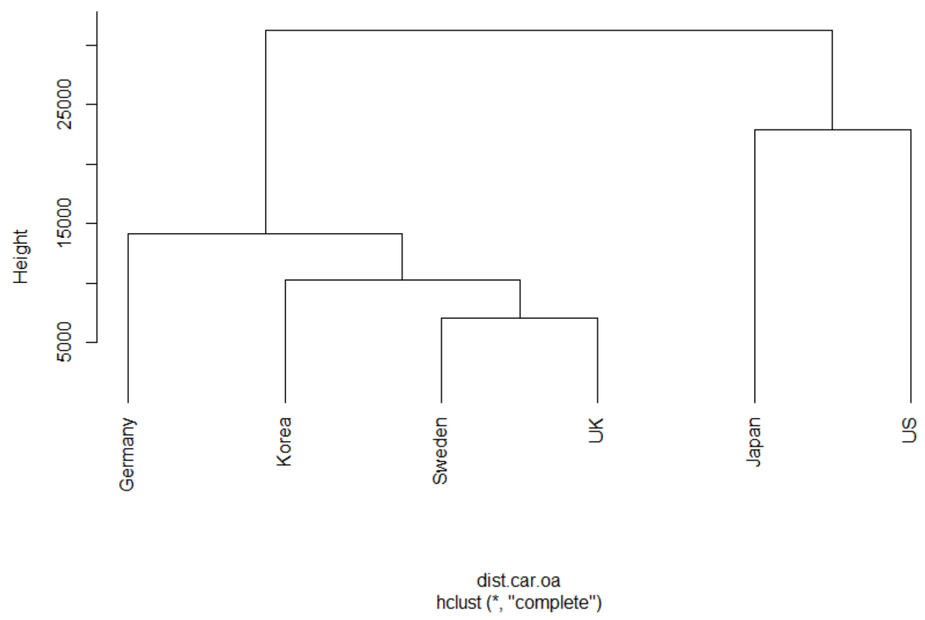

図1: 6つの集団間の非類似度に基づく階層的クラスタリング結果

図1より、米国産車（US）を除く5つの集団のうち、最も早い段階で米国産車と同一のクラスターとしてまとめられているのは日本車（Japan）であり、他の4つの集団についてのクラスターは米国産車を含むクラスターと大きな差異がみられる。この結果より、2004年時点ではこのデータにおける日本車の集団が他の非米国車の集団よりも米国産車の集団に相対的に近いことを示しており、米国市場により適応的であったと考えられる。