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非類似度

2016年6月17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。この場合の解析として、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行う方法が考えられるが、このようなデータを我々は**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個の連続変数およびq個のカテゴリー変数（カテゴリー変数 jにおけるカテゴリー値の数はm_j 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。

【非類似度の分解】

2つのカテゴリー変数k_1およびk_2がそれぞれm_{k_1}個、m_{k_2} 個のカテゴリーをもち、$s_{j_{l_1} j_{l_2}}^{(g_i, k_{l_1} k_{l_2})}$ $(i = 1,2;\ l_1, l_2 = 1,2)$が$k_1$および k_2により作られる分割表において各変数のカテゴリー値がそれぞれj_{l_1} 、j_{l_2} である個数とするとき、全ての組み合わせのLRTSの総和である非類似度 $d_{(cc)}^{(g_1, g_2)}$は以下の通り分解することができる。

$$\begin{aligned} d_{(cc)}^{(g_1,g_2)} = & \sum_{k_1=1}^{q-1} \sum_{k_2=k_1+1}^{q} \left[\sum_{j_1=1}^{m_{k_1}} \sum_{j_2=1}^{m_{k_2}} \left\{ s_{j_1 j_2}^{(g_1,k_1k_2)} \log \frac{\left(s_{j_1 j_2}^{(g_1,k_1k_2)}\right)^2}{s_{j_1 j_1}^{(g_1,k_1k_1)} s_{j_2 j_2}^{(g_1,k_2k_2)}} \right. \right. \\ & + s_{j_1 j_2}^{(g_2,k_1k_2)} \log \frac{\left(s_{j_1 j_2}^{(g_2,k_1k_2)}\right)^2}{s_{j_1 j_1}^{(g_2,k_1k_1)} s_{j_2 j_2}^{(g_2,k_2k_2)}} \\ & \left. \left. - (s_{j_1 j_2}^{(g_1,k_1k_2)} + s_{j_1 j_1}^{(g_2,k_1k_1)}) \log \frac{\left(s_{j_1 j_2}^{(g_1,k_1k_2)} + s_{j_1 j_2}^{(g_2,k_1k_2)}\right)^2}{\left(s_{j_1 j_1}^{(g_1,k_1k_1)} + s_{j_1 j_1}^{(g_2,k_1k_1)}\right)\left(s_{j_2 j_2}^{(g_1,k_2k_2)} + s_{j_2 j_2}^{(g_2,k_2k_2)}\right)} \right\} \right] \\ & + \sum_{k_1=1}^{q} (q-1) \left[\sum_{j_1=1}^{m_{k_1}} \left\{ s_{j_1 j_1}^{(g_1,k_1k_1)} \log \frac{s_{j_1 j_1}^{(g_1,k_1k_1)}}{n^{(g_1)}} + s_{j_1 j_1}^{(g_2,k_1k_1)} \log \frac{s_{j_1 j_1}^{(g_2,k_1k_1)}}{n^{(g_2)}} \right. \right. \\ & \left. \left. - (s_{j_1 j_1}^{(g_1,k_1k_1)} + s_{j_1 j_1}^{(g_2,k_1k_1)}) \log \frac{s_{j_1 j_1}^{(g_1,k_1k_1)} + s_{j_1 j_1}^{(g_2,k_1k_1)}}{n^{(g_1)} + n^{(g_2)}} \right\} \right] \end{aligned}$$

この分解において、等号の右側の下の2行は「それぞれの変数に関する周辺分布に基づく項」の総和、残りは「2変数の分布部分に基づく項」の総和である。すなわち、前者において k_1に関する項の値が大きければその変数に関して2集団間の差異が大きいことを示し、後者において k_1と k_2の組み合わせに関する項の値が大きければ各変数の周辺分布の影響を除いた変数の組み合わせに関して2集団間の差異が大きいことを示す。

【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り、10種類の連続変数および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、連続変数を離散化した上で各々の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）

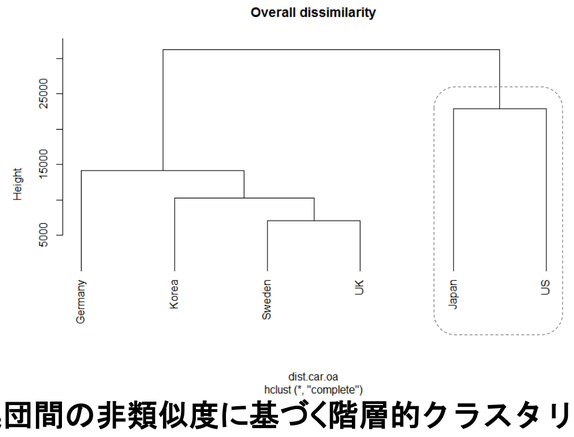

図1: 6つの集団間の非類似度に基づく階層的クラスタリング結果

ここで日本車と米国産車との間の非類似度に注目し、2変数ごとの組み合わせのLRTSをヒートマップとしてプロットしたのが図2である。濃い色であるほど大きな値となっている。左下から右上にかけての対角成分が各変数の周辺分布に基づく値であり、左図より変数8(車輪直径)などに関して差異が大きいことがわかる。また、各変数の周辺分布に基づく値を除去し2変数の分布に基づく値だけに注目すると、右図より変数9(車長)と変数10(車幅)の組み合わせなどにおいて他の組み合わせよりも差異が大きいことがわかる。

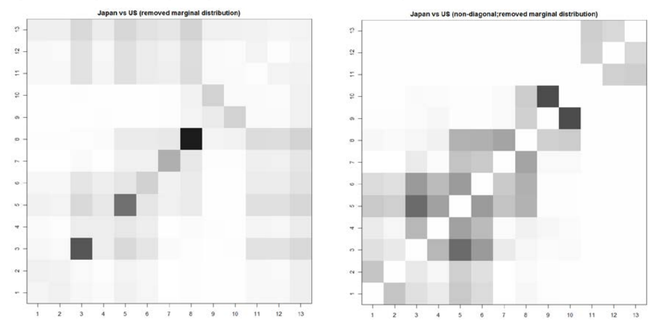

図2: 日本車と米国産車との間の非類似度を分解した結果
（左：分解後の値をそのまま表示、右：各変数の周辺分布に基づく値を除去）