変数型が混在する場合の 集約的シンボリックデータのクラスタリング

2015年6月19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。

SDAにおいては、主に多変量の連続型変数に区間データを想定し、それらに既存の各種統計手法を拡張する研究が多く行われてきたが、最近多数出現している連続型変数とカテゴリカル変数が混在する大規模データ場合に対して一貫した規準で対応できている研究は少ない。

一方、最近の大規模多変量データ集合においては、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。

ASD を考えることにより、従来のSDAの枠組みにおいてあまり考慮されてこなかった異なる2つの連続型変数間の相関係数を解析に織り込むことが可能になる。また、連続型変数とカテゴリカル変数が混在するデータ集合に対しても、それぞれの変数型に対応した記述統計量が従う確率モデルを用いて解析可能になる。特に、クラスタリングにおいては変数が従う確率モデルから導出される尤度比検定統計量を非類似度とすることで、変数型が混在する場合でも一貫した規準を作ることが可能になる。

本報告では、連続型変数とカテゴリカル変数が混在するデータ集合におけるASDのクラスタリングを行うにあたり、連続型変数を離散化してカテゴリカル変数として扱うことで、異なるASD間の非類似度を異なる2つずつのカテゴリカル変数の組み合わせに関する尤度比検定統計量の総和として考える。また、そのような非類似度を用いたクラスタリング手法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個のカテゴリカル変数（カテゴリカル変数 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個の連続型変数値、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。ここで連続型変数およびカテゴリカル変数に対し、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。すなわち(a) 異なる2つのカテゴリカル変数 (b) 異なる2つの連続型変数 (c) 1つの連続型変数と1つのカテゴリカル変数 の3種類の組み合わせにおける確率モデルを考える。

【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. 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上記(a)～(c)の組み合わせ全て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$\hat{l}_1$
3. 上記(a)～(c)の組み合わせごとに尤度比検定統計量$-2(\hat{l}_0 - \hat{l}_1)$を計算し、それらの総和を非類似度とする

ただし、この段階では(a)～(c)それぞれの組み合わせにおける尤度比検定統計量を同等には扱えない。そこで、連続型変数を離散化してカテゴリカル変数とみなし、(b)および(c)についても(a)と同様の組み合わせとして尤度比検定統計量を考えることで、全ての尤度比検定統計量を同等に扱った上で総和を非類似度と考えることができる。連続型変数の離散化方法は後述。

【異なる2つのカテゴリカル変数の組み合わせ】

各集団における異なる2つのカテゴリカル変数の組み合わせ（(a)）は分割表として表されるが、ここでの各セルにおける値はそれぞれの各カテゴリー値の組み合わせごとの生起数であり、これらが多項分布に従うと仮定する。この場合の集団間の非類似度 $d_{(cc)}^{(g_1,g_2)}$ は異なる2つのカテゴリカル変数の組み合わせにおける尤度比検定統計量の総和として求められる。

【連続型変数の離散化】

各集団において連続型変数を含む場合の組み合わせ（(b)および(c)）では連続型変数の平均ベクトルおよび分散共分散行列が正規分布に従うと仮定する。ここで、連続型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となるようにすると考えると、各連続型変数はと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カル変数と考えることができる。その上で(b)および(c)においても 集団間の非類似度 $d_{(rr)}^{(g_1,g_2)}$および$d_{(rc)}^{(g_1,g_2)}$を(a)と同様の考え方で求めることができ、連続型変数が従うと仮定した正規分布のパラメータの最尤推定量を用いて表される。

このように3種類の組み合わせ全てで尤度比検定統計量の求め方を揃えることにより、全体の非類似度 $d^{(g_1,g_2)}$ は全ての非類似度の総和で表せる。

$$d^{(g_1,g_2)} = d_{(cc)}^{(g_1,g_2)} + d_{(rr)}^{(g_1,g_2)} + d_{(rc)}^{(g_1,g_2)}$$

【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種類のカテゴリカル変数が含まれる。このデータをカテゴリカル変数”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）

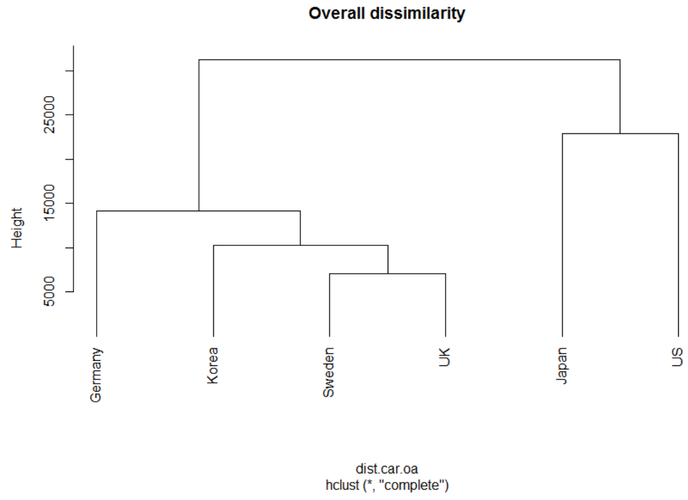

図1: 6つの集団間の非類似度に基づく階層的クラスタリング結果

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。