2013年6月14日　統計数理研究所　オープンハウス

混合正規分布モデルの利用による集約的シンボリックデータのクラスタリング

清水　信夫　　データ科学研究系　助教

何故シンボリックデータ解析か？

- 解析対象とする多変量データが大規模化かつ多様化
- それらを記述する上で柔軟なデータ構造を定義した枠組みが必要

⇒Didayにより<u>シンボリックデータ</u>(SD)が定義され、
これらを解析する枠組みとして<u>シンボリックデータ解析</u>(SDA)が出現

シンボリックデータの例としてどのようなものがあるのか？

- 1個の量的データ
- 1個の質的データ
- 複数の量的データもしくは質的データの集合
- 区間データ
- 上記の各種データに重みがついたデータの集合（ヒストグラムデータなど）
- 何らかの関係に基づく依存構造をもつデータ集合

など

集約的シンボリックデータ(ASD)とは何か？

- SDAにおいてデータ集合に対しいくつかのグループ分けが既に行われている場合にオリジナルデータではなく<u>各々のグループについての情報</u>に興味がある場合が存在

⇒それらのグループの特徴を分布として表現し、
それを近似的に表現した統計量をデータと考えたものを
<u>集約的シンボリックデータ</u>と呼ぶ

従来のシンボリックデータ解析の特徴と問題点は？

- 主に多変量区間データを考慮し、それらに既存の各種統計手法を拡張

⇒各変数ごとの平均および分散の情報（＝周辺分布の情報）のみを用いた解析

- 分散と同様に2次のモーメントである各変数間の相関係数は解析に利用されず

⇒相関係数の情報も用いてより詳細な解析を行えないだろうか？

ASDの定義

$\mathbf{X}=\{\mathbf{x}_{11},\cdots,\mathbf{x}_{1n_1},\cdots,\mathbf{x}_{g1},\cdots,\mathbf{x}_{gn_g}\}\ (n=\sum_{i=1}^{g}n_i)$

$\{x_{i1},\cdots,x_{in_i}\}$:グループ$i$における$p$次元データ $(i=1,\cdots,g)$

$\varphi_1,\cdots,\varphi_g$：$g$個の$p$変量確率密度関数 （**正規分布**を仮定）

$\boldsymbol{\mu}_i$：φ_i の平均　　$\boldsymbol{\Sigma}_i$：φ_i の分散共分散行列

$\{x_{i1},\cdots,x_{in_i}\}$を φ_i から独立に発生したデータ集合と考えると、各グループのデータ数 n_i が十分大きければ、$\boldsymbol{\mu}_i$ と$\boldsymbol{\Sigma}_i$ は$\{x_{i1},\cdots,x_{in_i}\}$ における平均と分散共分散行列で十分正確に推定される。

以上より ASD:$\{n_i,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\}(i=1,2,...,g)$ を下記の通り定義する。

$$\varphi_i(\mathbf{x}\,|\,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)=\frac{1}{(2\pi)^{p/2}\,|\,\boldsymbol{\Sigma}_i\,|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i(\mathbf{x}-\boldsymbol{\mu}_i)\right]$$

混合正規分布モデルによるクラスタリング

$\mathbf{X}$（＝個々のデータ）から生成される密度関数は混合正規分布

$$\hat{f}(\mathbf{x})=\sum_{i=1}^{g}\frac{n_i}{n}\varphi_i(\mathbf{x}\,|\,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$$

により表される。ここで

$C_1,\cdots,C_m$：m個のクラスター　　a_j：C_j の混合比率 $(\sum_{j=1}^{m}a_j=1)$

$\mathbf{v}_j$：C_jの平均　　$\mathbf{T}_j$：C_j の分散共分散行列

$\boldsymbol{\Psi}=\{\{a_1,\mathbf{v}_1,\mathbf{T}_1\},\cdots,\{a_m,\mathbf{v}_m,\mathbf{T}_m\}\}$

とするとき、

g 個のASDをm個のクラスターに分類するとは、$\hat{f}(\mathbf{x})$ に近い $f_m(\mathbf{x}\,|\,\boldsymbol{\Psi})=\sum_{j=1}^{m}a_j\varphi_j(\mathbf{x}\,|\,\mathbf{v}_j,\mathbf{T}_j)$ を見つけることである。

本報告における提案

- 多次元混合正規分布モデルにEMアルゴリズムを適用することによる各次元ごとの周辺分布だけの利用にとどまらないクラスタリング手法の提案

EMアルゴリズムのASDへの適用

対数尤度関数 $\log L(\boldsymbol{\Psi}\,|\,\mathbf{X})=\sum_{i=1}^{g}\sum_{j=1}^{n_i}\log\sum_{l=1}^{m}a_l\varphi_l(\mathbf{v}_l,\mathbf{T}_l\,|\,\mathbf{x}_{ij})$ を最大化する$\boldsymbol{\Psi}$を解析的に導出するのは困難であることから、**EMアルゴリズム**を用いて対数尤度関数の期待値を繰り返し計算により求める方法がよく用いられている。この方法をASDのクラスタリングに適用すると、k回目の繰り返しの値 $\boldsymbol{\Psi}^{(k)}$ は

$$\boldsymbol{\Psi}_l^{(k)}=\{a_l^{(k)},\mathbf{v}_l^{(k)},\mathbf{T}_l^{(k)}\},\quad U_i(\boldsymbol{\Psi}_l^{(k)})=\left\{\frac{a_l^{(k)}}{(2\pi)^{p/2}\,|\,\mathbf{T}_l^{(k)}\,|^{1/2}}\right\}^{n_i}\exp\left[-\frac{n_i}{2}\sum_{r=1}^{p}\sum_{s=1}^{p}\sigma_{irs}\tau_{lrs}^{(k)}-\frac{n_i}{2}(\boldsymbol{\mu}_i-\mathbf{v}_l^{(k)})'(\mathbf{T}_l^{(k)})^{-1}(\boldsymbol{\mu}_i-\mathbf{v}_l^{(k)})\right],\quad \left(\boldsymbol{\Sigma}_i=(\sigma_{irs}),(\mathbf{T}_l^{(k)})^{-1}=(\tau_{lrs}^{(k)})\right)$$

$$P_{il}^{(k)}=\frac{U_i(\boldsymbol{\Psi}_l^{(k)})}{\sum_{l=1}^{m}U_i(\boldsymbol{\Psi}_l^{(k)})},\quad a_l^{(k+1)}=\frac{1}{n}\sum_{i=1}^{g}\{P_{il}^{(k)}n_i\},\quad \mathbf{v}_l^{(k+1)}=\frac{1}{na_l^{(k+1)}}\sum_{i=1}^{g}\{P_{il}^{(k)}n_i\boldsymbol{\mu}_i\},\quad \mathbf{T}_l^{(k+1)}=\frac{1}{na_l^{(k+1)}}\sum_{i=1}^{g}\{P_{il}^{(k)}n_i[\boldsymbol{\Sigma}_i+(\boldsymbol{\mu}_i-\mathbf{v}_l^{(k+1)})(\boldsymbol{\mu}_i-\mathbf{v}_l^{(k+1)})']\}$$

となり、元々のデータ値である $\mathbf{x}_{ij}$ **を用いずに** $n_i,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i$ **の値のみを用いる**ことにより**<u>クラスタリングを高速に行える。</u>**

ASDの利用に基づく中国60都市の月別平均気温データ(1988年)の混合正規分布モデルクラスタリング例

- 各々の月をそれぞれ変数とみなす
- 各都市における区間データの値から異なる2つの変数間の分散共分散行列を計算

（異なる2変数間の共分散はここでは0とみなす）

- 60都市を5つのクラスターに分類

<u>http://dss.ucar.edu/datasets/ds578.5/</u>

Meteorological Station	January	February	…	December
AnQing	[1.8, 7.1]	[2.1, 7.2]	…	[4.3, 11.8]
BaoDing	[-7.1, 1.7]	[-5.3, 4.8]	…	[-3.9, 5.2]
BeiJing	[-7.2, 2.1]	[-5.9, 3.8]	…	[-4.4, 4.7]
…	…	…	…	…
ChangChun	[-16.9, -6.7]	[-17.6, -6.8]	…	[-15.9, -7.2]
ChanSha	[2.7, 7.4]	[3.1, 7.7]	…	[4.1, 13.3]
ZhiJiang	[2.7, 8.4]	[2.7, 8.7]	…	[5.1, 13.3]

区間データをそのまま用いて解析を行った De Carvalho et al.(2008) による結果とほぼ同様の特徴を表せている