一般化エネルギー関数に基づくクラスター分析

成蹊大学　小森　理,　統計数理研究所　江口　真透

今回の発表では非階層クラスタリングを考える．代表的なものに予めクラスターの数を決めた上で一番近いクラスターの平均からの距離を最小化する k-means 法と，メンバーシップ関数で特徴付けられる fuzzy-c 法がある．前者はハードクラスタリングとも呼ばれ，どのクラスター所属するかは明確に規定される一方，後者はソフトクラスタリングと呼ばれどのクラスターに所属するかはメンバーシップ関数の大きさで連続的に表現される．これら二つの異なるクラスタリングの手法を今回一般化エネルギー関数の枠組みで統一的に議論する．

まず始めに ϕ を単調増加関数とすると，データ $\{x_1,\dots,x_n\}$ の K 個のクラスターの中心 $\mu=(\mu_1,\cdots,\mu_K)^\top$ に対し，一般化エネルギー関数は

$$L_\phi(\mu)=\sum_{i=1}^{n}\phi^{-1}\Big(\sum_{k=1}^{K}\pi_k\phi(\|x_i-\mu_k\|^2)\Big)$$

のように定義される．但し $\pi_k\geq 0\ (k=1,\dots,K)$ は $\sum_{k=1}^{K}\pi_k=1$ を満たすクラスターの混合比とする．このエネルギー関数は二乗ロスの Kolmogorov-Nagumo 平均 (一般化平均) とみなすことができる [2, 1]．クラスターの中心の推定値 $\hat{\mu}$ はこのエネルギー関数を最小にするように求める．ここでさらに ϕ として次のようにパレート分布の分布関数 F を考えると，

$$F(s)=1-(1+\beta s)^{-\frac{1}{\beta}},\qquad F^{-1}(t)=\frac{(1-t)^{-\beta}-1}{\beta}$$

一般化エネルギー関数は

$$L_{\beta,\tau}(\mu)=\frac{1}{\tau}\sum_{i=1}^{n}\frac{1}{\beta}\Big[\Big\{\frac{1}{K}\sum_{k=1}^{K}(1+\tau\beta\|x_i-\mu_k\|^2)^{-\frac{1}{\beta}}\Big\}^{-\beta}-1\Big]$$

と表すことができる．但し $\beta>0$ かつ $\tau>0$ とする．推定アルゴリズムは以下の反復法を用いる．

---

1. クラスターの中心の初期値を定める $\mu^{(0)}=(\mu_1^{(0)},\cdots,\mu_K^{(0)})^\top$.
2. $t=1,\dots,T$ に対し，以下のステップを繰り返す.

   (a)
   $$w_k^{(t)}(x_i,\tau,\beta)=\left\{\frac{(1+\tau\beta\|x_i-\mu_k^{(t-1)}\|^2)^{-\frac{1}{\beta}}}{\sum_{\ell=1}^{K}(1+\tau\beta\|x_i-\mu_\ell^{(t-1)}\|^2)^{-\frac{1}{\beta}}}\right\}^{1+\beta},\ k=1,\dots,K$$

   (b)
   $$\mu_k^{(t)}=\frac{\sum_{i=1}^{n}w_k^{(t)}(x_i,\tau,\beta)x_i}{\sum_{i=1}^{n}w_k^{(t)}(x_i,\tau,\beta)},\ k=1,\dots,K$$

3. $\hat{\mu}=(\mu_1^{(T)},\cdots,\mu_K^{(T)})^\top$ とする.

ここで $\tau\to\infty$ かつ $\beta\to 0$ のとき $L_{\beta,\tau}(\mu)$ は k-means 法の損失関数に，$\tau\to\infty$ かつ $\beta=m-1(m>1)$ とすると fuzzy-c 法の損失関数に帰着することが示される．但し m は fuzz-c のメンバーシップ関数の重み指数とする．よってこの 2 つのパラメータ (τ,β) をうまく選ぶことにより，k-means 法や fuzzy-c 法では捉えることができないクラスター構造をもうまく推定できることが期待できる．当日はいくつかの数値実験を用いて提案法の統計的性質を議論する．

[1] Eguchi, S. and Komori, O. Path Connectedness on a Space of Probability Density Functions. , *Geometric Science of Information: Second International Coference, GSI 2015*, Springer, 2015. p. 615.

[2] Naudts, J. (2011). *Generalised Thermostatistics*, London: Springer.