Vol.2012-MPS-91 No.38
Vol.2012-BIO-32 No.38
2012/12/7

カーネル法によるDP-meansの非線形化

北口 景子[†1]　吉田 裕亮[†1]

概要：ディリクレ過程混合モデルと k-means 法を組み合わせた DP-means 法というクラスタリング手法がある．本研究では，DP-means 法にカーネル法を用いることにより，クラスタ数を自動推定し，カーネル k-means 法のように初期値に依存しない非線形クラスタリング手法を提案する．

DP-means for non-linear data using kernel technique

Kitaguchi Keiko[†1]　Yoshida Hiroaki[†1]

Abstract: DP-means is a k-means-like clustering method using the Dirichlet process mixture. In this study, we shall apply the kernel technique to DP-means and propose a non-linear clustering method which estimates the number of clusters automatically like as kernel k-means.

1. はじめに

データをまとまりごとに集めてグループ分けをすることは，クラスタリングと呼ばれている．非線形なデータをグループ分けするには，カーネル k-means 法やスペクトラルクラスタリングが一般的であるが，前者は初期値依存が強く適切な解を得るのに何度も初期値を変えて繰り返さなければならない場合がある．また，どちらも事前にユーザがクラスタ数を設定する必要があり，設定されたクラスタ数により結果が大きく変化する．そこで，本研究ではディリクレ過程混合モデルと k-means 法を組み合わせた DP-means 法に対し，カーネル法を用いることで，初期値に依存せず，クラスタ数を推定する非線形クラスタリング手法を提案する．

2. ディリクレ過程混合モデル

ディリクレ過程混合モデルはディリクレ過程 (DP:Dirichlet Process) を各出力モジュールの選択確率の事前分布に利用した混合モデルである．ディリクレ過程は無限次元のディリクレ分布であり，原理的には無限次元の多項分布を生成する．

ディリクレ過程は，基底分布 G_0 と正の集中度パラメータ γ で定義される．ディリクレ過程混合モデルのデータ生成過程は以下である．

(1) $G \sim DP(\gamma, G_0)$
(2) $\theta_i \mid G \sim G, \text{for } i = 1, \ldots, n$
(3) $x_i \sim p(x \mid \theta_i), \text{for } i = 1, \ldots, n$

まず，パラメータ θ の事前分布である基底分布 G_0 から，DP によってパラメータの事前分布 G が生成される．これは $\theta \in \Theta$ を変数に持つ離散分布である．データが生成される際には，そのデータ毎に分布 G から θ_i を生成し出力分布のパラメータとしている．

3. k-means 法

k-means 法は，以下のアルゴリズムでクラスタリングを行う. ここで，データの数を n，クラスタ数を K とする．

(1) 各データ $x_i (i = 1, \ldots, n)$ に対してランダムにクラスタを割り当てる．
(2) 割り当てたデータを元に各クラスタの中心 V_j $(j = 1, \ldots, K)$ を計算する．

[†1] 現在，お茶の水女子大学大学院　人間文化創成科学研究科
Presently with Graduate School of Humanities and Sciences, Ochanomizu University

Vol.2012-MPS-91 No.38
Vol.2012-BIO-32 No.38
2012/12/7

(3) 各 x_i と V_j との距離を求め，x_i を最も近い中心のクラスタに割り当て直す．

(4) 上記の処理で全ての x_i のクラスタの割り当てが変化しなかった場合は処理を終了する．それ以外の場合は新しく割り振られたクラスタから V_j を再計算して上記の処理を繰り返す．

4. DP(Dirichlet Process)-means 法

DP-means 法 [1] は以下のアルゴリズムでクラスタリングを行う手法である．

初期値設定：$k=1$，$l_1=\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ とし，μ_1 は全データの平均である．λ の値は適当に設定しておく．

$\boldsymbol{x}_i(i=1,\ldots,n)$ について各パラメータの値が収束するまで以下の (1)〜(3) を繰り返す．

(1) $\boldsymbol{x}_i(i=1,\ldots,n)$ について，

- $d_{ic}=\| \boldsymbol{x}_i-\boldsymbol{\mu}_n \|^2\ (c=1,\ldots,k)$
- $\min_c d_{ic}>\lambda$ の場合 , $k=k+1, z_i=k, \boldsymbol{\mu}_k=\boldsymbol{x}_i$
- それ以外の場合 $z_i=\mathrm{argmin}_c d_{ic}$

(2) $z_1\ldots,z_k$ に基づき，クラスタ $l_1,\ldots,l_k$ に各データを割り当てる．

(3) $l_j(j=1,\ldots,k)$ について $\boldsymbol{\mu}_j=\frac{1}{|l_j|}\sum_{x\in l_j}\boldsymbol{x}$ を求める．

このとき，k はクラスタ数，l_i は i 番目のクラスタ，μ_i はクラスタ l_i の平均，z_i は $\boldsymbol{x}_i$ が所属するクラスタ番号を表す．また，λ はクラスタ数 k を制御するパラメータであり，一般に λ の値が大きければ k の値は小さく，λ の値が小さければ k の値は大きくなる傾向にある．適切なクラスタリング結果を得るには，適切な λ を選択する必要がある．

5. 提案手法

カーネル法 [2] では，非線形なデータを一度入力空間から高次元の特徴空間上に写像することで解析しやすいデータに変換し，その特徴空間上で線形なモデルを組み立てて，問題を解く．このとき，特徴空間上での内積をカーネル関数を用いて計算することによって，計算量を抑えることができる．

そこで本研究では，DP-means 法におけるデータ間距離 d_{ic} を以下のようにカーネル法に基づき特徴空間上に表現して，DP-means 法の手順でクラスタリングを行う手法を提案する．

$$\begin{aligned} d_{ic} &= \| \phi(x_i)-\mu_c \|^2 \\ &= K(x_i,x_i)-\frac{2}{n_c}\sum_{x_j\in X_c}K(x_i,x_j) \\ &\quad +\frac{1}{n_c^2}\sum_{x_i,x_l\in X_c}K(x_j,x_l) \end{aligned}$$

なお本研究ではカーネル関数にガウスカーネル

$$k(x_i,x_j)=\exp\left(-\frac{\| x_i-x_j \|^2}{\beta^2}\right)$$

を用いることにした．この手法により，カーネル k-means 法のように初期値に依存せず，クラスタ数を自動に推定する非線形なクラスタリングを行うことができる．

6. 実験例

図 1，図 2 のような，線形で分けることの出来ないデータを用意する．図 1 は 1 群が 100 点，2 群が 300 点の合計 400 点のサンプルデータ，図 2 は 1 群が 100 点，2 群が 100 点，3 群が 300 点の合計 500 点のサンプルデータとなっている．これらを提案手法によってクラスタリングを行う．λ の値とガウスカーネルにおける β の値はあらかじめ適当な値に設定しておく．

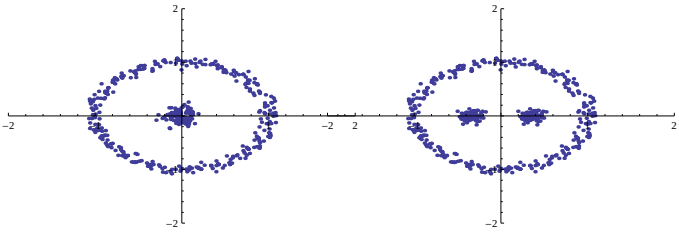

図 1 サンプルデータ 1

図 2 サンプルデータ 2

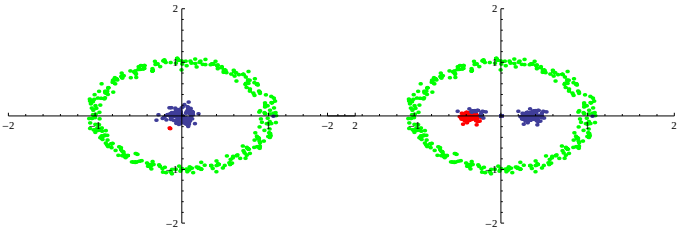

図 3 結果 1

図 4 結果 2

7. 結果とまとめ

サンプルデータ 1 については図 3 の結果が得られた．このとき，λ の値は 0.757 であり，クラスタ数 k は 2 となった．サンプルデータ 2 については図 4 の結果が得られた．このとき，λ の値は 0.6316 であり，クラスタ数 k は 3 となった．

2 つのサンプルデータについて，期待したとおりのクラスタリング結果と k の値は得られたが，実は λ の値により結果が大きく左右し，適切な λ の値を探すのが難しい．今後は適切な λ の値の定め方を検討する課題が残されている．

参考文献

[1] Jordan, M.I. and EDU, B.: *Revisiting k-means: New Algorithms via Bayesian Nonparametrics*, arXiv preprint arXiv:1111.0352(2011).

[2] 赤穂昭太郎：カーネル多変量解析 非線形データ解析の新しい展開，岩波書店．