ISS-SP-188

# レート歪み理論による
# ディリクレ過程平均法の閾値パラメータ解釈

小林　真佐大†　　渡辺　一帆†
† 豊橋技術科学大学工学部情報・知能工学課程

## 1. はじめに

代表的なクラスタリング手法の一つである k-means 法を拡張する形で，DP-means 法が考案された[1]. DP-means 法は，閾値となるペナルティパラメータを指定すれば，データからクラスタ数を推定することができる. しかし, ペナルティパラメータの変化に対し，推定されるクラスタ数の変化は未だに明かされていない．本研究では，ペナルティパラメータを変化させたときの，クラスタ数の曲線が，データの次元数を大きくする極限でレート歪み曲線に近づくことを示す.

## 2. DP-means 法について

DP-means 法は，データ$\boldsymbol{x}^n = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}\left(\boldsymbol{x}_i = \left(x_i^{(1)}, \dots, x_i^{(L)}\right)\epsilon \boldsymbol{R}^L\right)$が与えられたとき，クラスタ数推定を行いデータ$\boldsymbol{x}^n$に対し順次，クラスタ割り当てを行うものである．$\{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K\}$をクラスタ中心とすると，クラスタが増えるのは，ペナルティパラメータ$\boldsymbol{\lambda}$より，各データ点$\boldsymbol{x}_i$と最も近いクラスタ中心との擬距離の値が大きい，すなわち$\min_k d(\boldsymbol{x}_i, \boldsymbol{\mu}_k) > \lambda$を満たす時である．なお，本研究では，データの各次元は非負整数値である混合二項分布を仮定し，擬距離は，以下とする.

$$d(\boldsymbol{x}, \boldsymbol{\mu}) = \frac{1}{L}\sum_{j=1}^{L}\left\{x^{(j)} \ln \frac{x^{(j)}}{\mu^{(j)}} + \left(N - x^{(j)}\right) \ln \frac{N - x^{(j)}}{N - \mu^{(j)}}\right\}$$

## 3. レート歪み理論における平均歪みと最大歪みの関係

平均歪み，最大歪みは，それぞれ(1)，(2)で定義される．またその時のレート$R$を平均歪み，最大歪みの値が得られた時のクラスタ数$K$を(3)に代入した値とする.

$$D_a = \frac{1}{n}\sum_{i=1}^{n} \min_k d(\boldsymbol{x}_i, \boldsymbol{\mu}_k) \quad (1)$$

$$D_m = \max_i \min_k d(\boldsymbol{x}_i, \boldsymbol{\mu}_k) \quad (2)$$

$$R = \frac{\ln K}{L} \quad (3)$$

データの数が$n \to \infty$であるとき，平均歪み，最大歪みのそれぞれに対応するレートには，次元数$L \to \infty$の極限で達成される同一の下限が存在することが知られ[2]，この下限のことをレート歪み曲線$R(D)$という．DP-means 法において，クラスタ数が$K$となるときのペナルティパラメータ$\lambda$の下限を$\lambda^*$とおくと，$\lambda^* = D_m$がすべての$K$で成り立つ．このことから，次元$L$を上げていった場合，ペナルティパラメータ$\lambda$を変化させた時のクラスタ数の曲線(3)は，レート歪み曲線に近づくと考えられる.

## 4. 数値実験

各データ点の次元$L$を変化させ，次元ごとに，最大値$N = 100$，成功確率$p = 0.3$となる二項分布より，学習データセットとテストデータを作成した．作成した学習データセットに対して，次元ごと，学習データ別にペナルティパラメータを変化させて，DP-means 法を行い，そのときのクラスタ数，学習データとテストデータのそれぞれに対し，平均歪み，最大歪みを求め，それらの学習データセットの出方に関する平均を結果とした．図 1 に次元ごとのクラスタ数の変化とレート歪み曲線を示す.

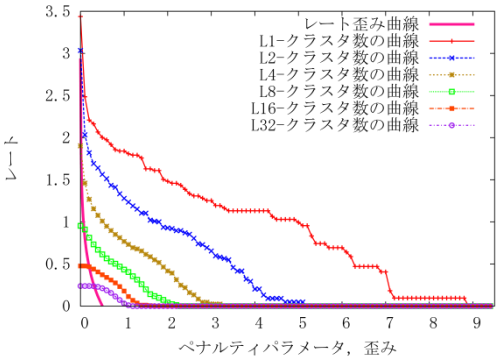


図 1. 次元変化に対するクラスタ数とレート歪み曲線

図 1 よりクラスタ数の曲線は次元を上げると，レート歪み曲線に近づいていることが分かる．しかし，歪みが 0 となる近辺では，レート歪み曲線の下にクラスタ数の曲線が来ている．これはデータ数を有限の値で抑えていることが原因だと考えられる．また，データ数の影響が少ないと考えられる，レートが 0 となる歪み 0.5032 に着目し，次元を上げて調べたところ，0.5032 に近づいていく様子が確認できた.

## 5. まとめ

DP-means 法のクラスタ数変化とレート歪み曲線の間の関係を示した．また，数値実験によりデータの次元数を上げて，ペナルティパラメータを変化させたとき，歪みが大きい領域ではクラスタ数の曲線はレート歪み曲線に近づくことが確認できた.

## 参考文献

[1] B. Kulis and M. I. Jordan, “Revisiting k-means: New Algorithms via Bayesian Noparametrics,” *Proc. ICML* 2012.

[2] 韓太舜, 『情報理論における情報スペクトル的方法』, 1998.