3Y-05

高次元データ解析における 測地的 k-平均クラスタリング法の効果の検証

芦沢 未菜†, 呉 湘筠‡, 高橋 成雄‡,

† 東京大学大学院新領域創成科学研究科
‡ 東京大学大学院情報理工学系研究科

1. はじめに

データの意味付けやカテゴリ分類など，その全体的な特徴を把握する際に用いられるクラスタリング手法のひとつに k-平均法がある. これは各サンプル点とクラスタ中心との非類似度を調べ，最も近いクラスタに各サンプル点を分類する手法であるが，サンプル点間の非類似度としてユークリッド距離を用いるため，空間に歪んで配置されたサンプル点集合や，複数の連結成分から成るサンプル点集合等に対し, その特徴を捉えたクラスタリングができない．そこで本研究では，サンプル点集合に対して近接グラフを構築して測地線距離を定義し，それに基づく測地的 k-平均クラスタリング法[1]を導入し，可視化画像の比較によりその効果を検証する．

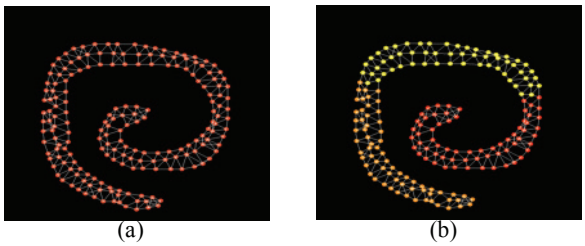
2. 測地的 k-平均法

測地的 k-平均法においては，サンプル点集合に対し近接グラフ(図 1(a))を構築し，サンプル点間の非類似度として，分布や密度を適切に反映する測地線距離を導入し，クラスタリングを行う(図 1(b)). また，この手法の特筆すべき点は，サンプル点とクラスタ中心との測地線距離を，既存のサンプル点間の距離のみを用いて計算することである．これにより，クラスタ中心を近接グラフにその都度取り込むことなく，クラスタリングを実行できる．

2.1. 測地線距離の定義

まず，サンプル点を頂点とし，各頂点とユークリッド距離が近いk − 1 個の頂点をエッジでつなぐことで，図 1(a)のような k 近接グラフを構築する．頂点間の測地的距離は，この k 近接グラフのエッジを辿った距離の総和として定義する．

Study on the effects of geodesic k-means clustering in high dimensional data analysis
†Mina Ashizawa, ‡Hsiang-Yun Wu, ‡Shigeo Takahashi
†Grad. School of Frontier Sciences, The Univ. of Tokyo
‡Grad. School of Information Science and Technology, The Univ. of Tokyo



(a) (b)

図 1: 渦巻き状のサンプル点集合. (a) 近接グラフ. (b) 測地的 k-平均法によるクラスタリング結果.

さらに，k近接グラフにおいて，ある頂点 x_iから$k-1$番目に近い頂点との距離を$R\left(N_k(x_i)\right)$とおく．ここで$Vol\left(R\left(N_k(x_i)\right)\right)$を半径$R\left(N_k(x_i)\right)$の球の体積としたとき，$x_i$ の周りの密度を，

$$\hat{f}(x_i) = \frac{k-1}{n*Vol(R(N_k(x_i)))} \quad (1)$$

と表すことにする．式(1)を用いて，k 近接グラフのエッジで直接結ばれている 2 点x_i, x_j間の測地線距離を，密度を考慮に入れて，

$$d(x_i, x_j) = \exp\left(\frac{1}{2\sigma^2 \max\{\hat{f}(x_i), \hat{f}(x_j)\}}\right) \| x_i - x_j \| \quad (2)$$

と定める．ここで，任意定数 σは密度による距離の変化度合いを表す．これにより疎なところほど，ユークリッド距離と比較したとき測地線距離がより長く定義される．

以上により，サンプル点の分布と密度を考慮した距離測定が可能となる．

2.2. 測地的 k-平均法のアルゴリズム

n個のサンプル点集合をc個のクラスタに分類するときの手順は以下のようになる．

1) サンプル点をc個のクラスタにランダムに分類する．
2) 各サンプル点と各クラスタ中心との測地線距離を求める．
3) 各サンプル点と最も近いクラスタ中心を求め，その所属先を対応するクラスタに変更する．
4) 3)→4)の手順を，クラスタの変更がなくなるまで繰り返す．

ここで，3)について，測地線距離を用いる本手法においては，クラスタ中心を定めることで，

点が増加し，近接グラフが変化し，距離の定義が変わってしまう問題が生じる．ここでは，各サンプル点とクラスタ中心との測地線距離を，文献[1]にならいサンプル点集合間を別のユークリッド距離空間に一時的に射影して間接的に計算する．

ここで行列D，Hをそれぞれ$D_{ij} = d^2(x_i, x_j)$，$H = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} - \frac{1}{n}\begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}$ とすると，$-\frac{1}{2}HDH$ が半正定値のとき，より低次元空間で

$$\| y_i - y_j \| = d(x_i, x_j) \quad (3)$$

かつ，$y_i \in Y$となる点集合Yを見つけることができる．また，集合 Y 内のクラスタC_lの中心$\bar{y}_l$に対応する$\bar{x}_l$を，サンプル点集合での仮想クラスタ中心とする．このときx_iと$\bar{x}_l$との測地線距離は，内積<,>を用いて，

$$d^2(x_i, \bar{x}_l) = \| y_i - \bar{y}_l \|^2 = \langle y_i - \bar{y}_l, y_i - \bar{y}_l \rangle \quad (4)$$

と表すことができる．ここでクラスタC_l内のサンプル点の総数をn_lとし，$\bar{y}_l = \frac{1}{n_l}\sum_{r:x_r \in C_l} y_r$であることに注意すると，$d^2(x_i, \bar{x}_l)$は

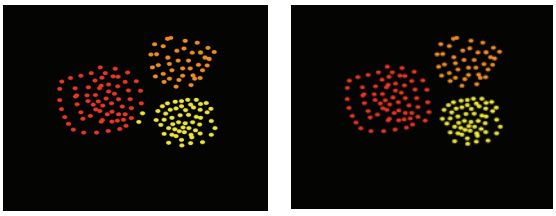
$$d^2(x_i, \bar{x}_l) = \frac{2}{n_l}\sum_{x_r \in C_l} d^2(x_i, x_r) - \frac{1}{{n_l}^2}\sum_{x_r, x_{r'} \in C_l} d^2(x_r, x_{r'}) \quad (5)$$

と，既存のサンプル点のみで表現可能となる．

3. 実験結果

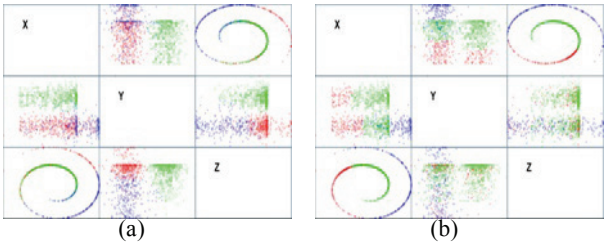
以下に３つの連結成分から成る２次元サンプル点集合(図 2)と，空間に歪んで配置された３次元サンプル点集合(図 3)に対して適用した，一般的な k−平均法(a)と測地的 k−平均法(b)によるクラスタリングの結果を示す．

図 2(a)では，左の赤いクラスタ内に黄色い点が混ざっているのに対し，図 2 (b)では，正確に点集合を３つのクラスタに分離できている．これは，距離の定義に密度を考慮に入れたため，各クラスタ間の距離が実際のユークリッド距離よりも離れていると計算されるためだと考えられる．また，図 3(a)では，青，赤，緑が入り交じっているのに対し，図 3(b)では，渦巻き状のデータが，外側からほぼ青，赤，緑の順にクラスタリングされている．これは，近接グラフのエッジを辿った距離の総和で距離を定義したことにより，クラスタリングの結果に点の接続性



(a) (b)

図 2: 3 つのクラスタ集合 ($n = 161,\ k = 5,\ \sigma = 1.0$).
(a) k−平均法．(b) 測地的 k−平均法．



(a) (b)

図 3:スイスロール ($n = 1200,\ k = 5,\ \sigma = 100.0$).
(a) k−平均法．(b) 測地的 k−平均法．

が考慮された結果だと考えられる．また，測地的 k−平均法において，2 次元よりも 3 次元の事例で，クラスタ数が指定した数以下になる問題が見られた．これは，3 次元空間に対してサンプル点数が少なく，近接グラフが分離し，近傍探索の精度が低下したためだと考えられる．

4. まとめ

本稿では，一般的な k-平均法の拡張として測地線距離を用いたクラスタリング手法を導入し，結果を用いて測地的 k−平均法の特長を検証した．

今後の課題として，さらに次元の高いデータに適応させるため，サンプル点の増加とそれに伴う計算コストの増加の回避法として，具体的には高次元空間内におけるサンプル点をつなぐ適切な近接グラフの選択について検討していくこと[2]が挙げられる．

謝辞 本研究の一部は，文科省科研費25120014と学振科研費25540041, 26730061の助成による．

参考文献

[1] N. Asgharbeygi, and A.Maleki, “Geodesic K-means Clustering,” In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR2008)*, pages 1-4, 2008.
[2] P. Oesterling et al., “Visualization of High-Dimensional Point Clouds Using Their Density Distribution's Topology,” *IEEE Transactions on Visualization and Computer Graphics,* vol. 17, no. 11, pp. 1547-1559, 2011.