F-016

中心性指標の分布に基づく局所改善クラスタリングの性能評価

A Performance Evaluation of Local Improvement Clustering Based on a Distribution of Centrality Indicator

山岸祐己 * Yuki Yamagishi
斉藤和巳 * Kazumi Saito

1. はじめに

オブジェクト集合のクラスタリングは，統計分析，機械学習，データマイニングなどにおける基本問題である．各オブジェクトがある種のベクトルとして与えられるケースでは，K-平均（K-means）法 [1] が代表的な解法であり，ガウス混合分布の推定問題として定式化すれば EM（Expectation-Maximization）アルゴリズム [2] を用いて，妥当な精度の結果を効率良く求めることができる．また，クラスターの中心を代表オブジェクト集合に限定する枠組みのクラスタリングは K-メディアン（K-median または K-medoids）問題と呼ばれており，一般に，K-メディアン問題として定式化すれば，外れ値などにロバストになることが知られている [3]．K-平均法がクラスターの中心として任意のベクトルが扱えないと適用できないのに対し，K-メディアン問題はそのような状況下でも適用可能な汎用性を有するため，本論文では K-メディアン問題の解法と解品質に着目し，オブジェクト集合の性質との関連性を調べる．

K-メディアン問題を一般に離散最適化の観点で考えれば NP-完全クラスに属するため，大規模になれば妥当な計算時間で厳密解を求めることは困難である．ただし，本問題は劣モジュラ性と呼ばれる性質を持つことが想定できる．劣モジュラ性を持つ離散最適化問題は，潜在的に幅広い応用が存在し，例えば，多変量データから有用な変数集合を選択する問題 [4]，社会ネットワーク上の情報伝播で影響を最大にするノード集合を選択する問題 [5] などが知られている．この最適化問題の重要な特徴は，いわゆる貪欲法で効率良く求まる近似解により，ある程度妥当な精度で最悪ケースの解品質が理論的に保証されている [6] ということである．しかし，貪欲法のみでは比較的アーな局所解にトラップされる危険性が伴うため，この問題の解品質を向上させる手段としては局所改善法が挙げられる．局所改善法を用いれば，高品質の解を安定して求められることが期待できるが，その計算量は大幅に増大する傾向があるため，大規模な問題になるほど，局所改善法を適用する前にその有効性を推定することは重要となってくる．よって，今回我々はオブジェクト集合に対して中心性指標を導入し，その分布の歪度 (skewness) を用いることによって，K-メディアン問題に対する局所改善法の有効性を推定することを試みる．

2. K-medoids クラスタリング

K-medoids 問題は，非階層クラスタリングで有名な K-means 法と同様に，H 個のオブジェクト集合 $\mathcal{H}$ が与えられたとき，オブジェクト集合を K 個のクラスターに分割する手法である．任意のオブジェクトペア $u, v \in \mathcal{H}$ 間に類似度 $\rho(u, v)$ が定義されていれば，オブジェクト集合の中から他のオブジェクトとの類似度の和が高い代表オブジェクトを選定することが可能であるため，最適な代表オブジェクトが選定されれば，類似度の高いオブジェクトペアは同じクラスターに，類似度の低いオブジェクトペアは異なるクラスターに属するように分割されるはずである．このような問題では，一般的に平均（mean）より中央値（median）の方が頑健であることが知られている [3]．ただし，大域最適解を求めるためには $O(H^K)$ の計算量が必要であるため，オブジェクト集合の規模や K がある程度大きくなると，実用的な時間で解を求めることが難しくなる．よって， K-medoids にも局所最適解を求めるための反復法や貪欲法が存在するが，今回は解の一意性が保証される貪欲法に基づく解法を採用する．この解法は，目的関数の劣モジュラ性により，厳密解ではないものの，ある程度妥当な精度で最悪ケースの解品質が理論的に保証されている [6]．貪欲法とは，既に選定した代表オブジェクトを固定し，ある評価関数値を最大にするオブジェクトを求め，目的関数が増加するならば代表オブジェクト集合に追加することで，結果の代表オブジェクト集合を求める方法である．各オブジェクトは，最も類似度の高い代表オブジェクトと同じクラスターに割り当てられる．既に選定した代表オブジェクト集合を $\mathcal{P}$ とし，新たに追加を試みるオブジェクトを w とするとき，ここでは，以下の目的関数を考える．

$$f(\mathcal{P} \cup \{w\}) = \sum_{v \in \mathcal{H}} \max\{\mu(v\,;\mathcal{P}), \rho(v, w)\}. \qquad (1)$$

ここで，$\mu(v\,;\mathcal{P})$ は既に選定された代表オブジェクトとの類似度の最大値を表し，$\mu(v\,;\mathcal{P}) = \max_{w \in \mathcal{P}}\{\rho(v, w)\}$ で定義される．以下に K-medoids における貪欲アルゴリズムを説明する．

A1-1. $k \leftarrow 1$, $\mathcal{P}_0 \leftarrow \emptyset$，各オブジェクト $v \in \mathcal{H}$ に対し，$\mu(v\,;\emptyset) \leftarrow 0$ と初期化する；

A1-2. 式 1 で $\hat{p}_k = \arg\max_{w \in \mathcal{H} \setminus \mathcal{P}_{k-1}}\{f(\mathcal{P}_{k-1} \cup \{w\})\}$ を求め，$\mathcal{P}_k \leftarrow \mathcal{P}_{k-1} \cup \{\hat{p}_k\}$ とする；

A1-3. $k = K$ ならば $\hat{\mathcal{P}}_K = \{\hat{p}_1, \ldots, \hat{p}_K\}$ を出力し，各オブジェクトを，最も類似度の高い代表オブジェクト $\hat{p}_k \in \hat{\mathcal{P}}$ のクラスター $\mathcal{C}_k$ に割り当て終了する；

A1-4. 各オブジェクト $v \in \mathcal{H}$ に対し， $\mu(v\,;\mathcal{P}_k)$ を求める；

*静岡県立大学, University of Shizuoka

A1-5. $k \leftarrow k+1$ とし，ステップ A1-2. へ戻る．

明らかに，上記のアルゴリズムの計算量は $O(NK)$ となり非常に高速である．しかし，貪欲法に基づく単純な手法であるため，比較的プアーな局所解にトラップされる危険性が伴う．よって，ここからは貪欲アルゴリズムで得た $\hat{\mathcal{P}}_K$ の解品質を向上させるための局所改善アルゴリズムについて述べる．

A2-1. $k \leftarrow 1, h \leftarrow 0$ と初期化する；

A2-2. 式 1 で $p'_k = \arg\max_{w \in \mathcal{H} \setminus \hat{\mathcal{P}}_K \setminus \{\hat{p}_k\}} \{f(\hat{\mathcal{P}}_K \setminus \{\hat{p}_k\} \cup \{w\})\}$ を求める；

A2-3. $p'_k = \hat{p}_k$ ならば $h \leftarrow h+1$ とし，さもなければ $h \leftarrow 0, \hat{\mathcal{P}}_K = \hat{\mathcal{P}}_K \setminus \{\hat{p}_k\} \cup \{p'_k\}$ とする；

A2-4. $h = K$ ならば $\hat{\mathcal{P}}_K$ を出力し，各オブジェクトを，最も類似度の高い代表オブジェクト $\hat{p}_k \in \hat{\mathcal{P}}$ のクラスター $\mathcal{C}_k$ に割り当て終了する；

A2-5. 各オブジェクト $v \in \mathcal{H}$ に対し，$\mu(v; \hat{\mathcal{P}}_K)$ を求め，$k = K$ ならば $k \leftarrow 1$，さもなければ $k \leftarrow k+1$ とし，ステップ A2-2 へ戻る；

明らかに，局所改善アルゴリズムを適用すると，貪欲アルゴリズムだけのときよりも多くの計算量を必要とする．以下，貪欲アルゴリズム単体を A1，貪欲アルゴリズムの後に局所改善アルゴリズムを適用するものを A2 と呼ぶ．

3. データセット

今回の実験で使用したデータセットは，2 次元ユークリッド空間の半径 1 の円上にランダムに生成したオブジェクト集合である．データセットの性質を変化させるため，オブジェクトの生成法として D0 から D5 までの 6 パターンを用意し，それぞれの生成法においてオブジェクト数 $H = 1000$ のデータを 100 個ずつ生成した．任意のオブジェクトが極座標 (r, θ) によって表されるとき，u を $[0,1]$ の一様乱数とすると，生成されるオブジェクト $x \in \mathcal{H}$ の偏角は $\theta = 2\pi u$，動径は以下のデータタイプ毎の設定によって決定される．

D0. $r = 1$;

D1. $r = u^{1/6}$;

D2. $r = u^{1/4}$;

D3. $r = u^{1/2}$;

D4. $r = u$;

D5. $r = u^2$;

ここで，我々はオブジェクト x の中心性指標を導入する．オブジェクトペア $a, b \in \mathcal{H}$ 間のユークリッド距離を $d(a,b)$ としたとき，オブジェクト x の中心性 $c(x)$ を $c(x) = \sum_{y \in \mathcal{H}} d(x,y)$ と定義する．各データタイプのサンプルデータと中心性指標 $c(x)$ の分布を図 1 から図 12 に示す．なお，中心性分布の度数は 100 個のデータの平均値である．

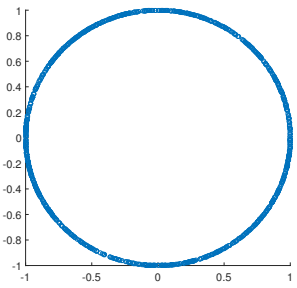

図 1: D0 のサンプル

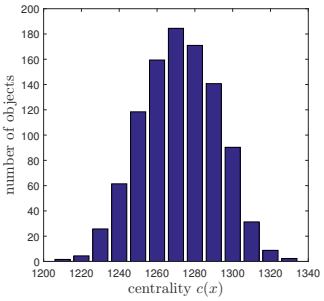

図 2: D0 の $c(x)$ の分布

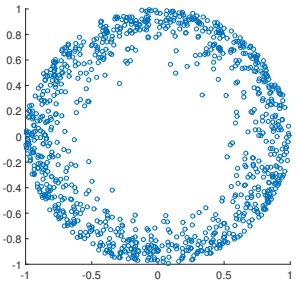

図 3: D1 のサンプル

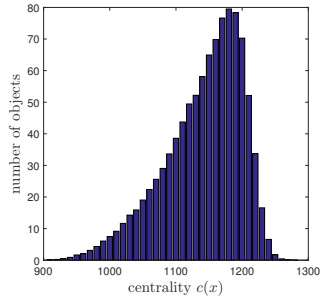

図 4: D1 の $c(x)$ の分布

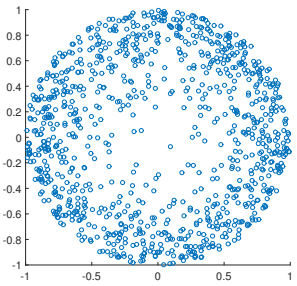

図 5: D2 のサンプル

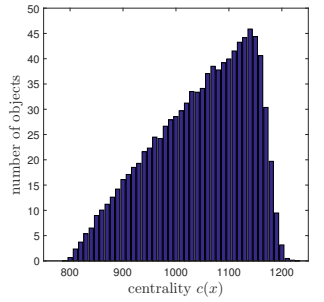

図 6: D2 の $c(x)$ の分布

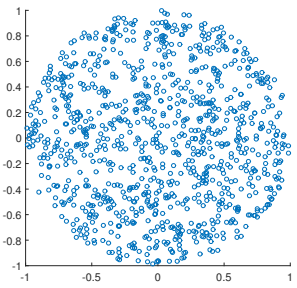

図 7: D3 のサンプル

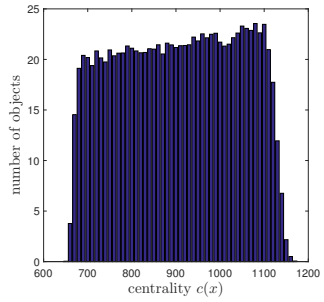

図 8: D3 の $c(x)$ の分布

4. 実験結果

今回のデータセットにおけるオブジェクトペア間の最大ユークリッド距離は 2 であるため，K-medoids クラスタリングにおける任意のオブジェクトペア間の類

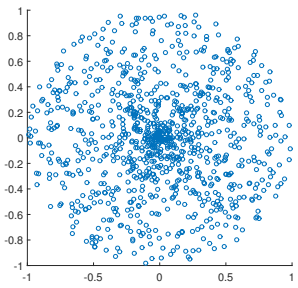

図 9: D4 のサンプル

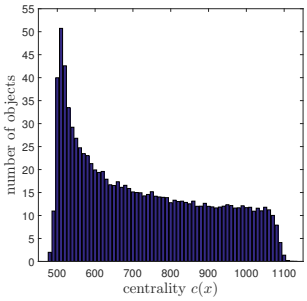

図 10: D4 の $c(x)$ の分布

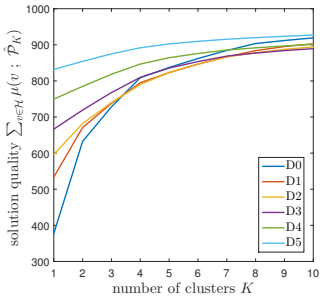

図 13: A1 の解品質

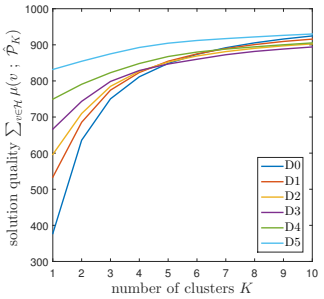

図 14: A2 の解品質

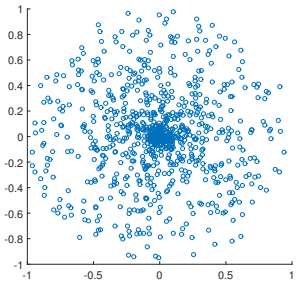

図 11: D5 のサンプル

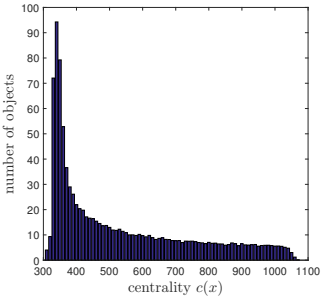

図 12: D5 の $c(x)$ の分布

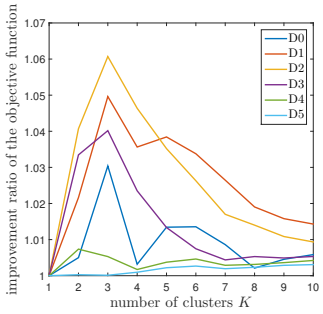

図 15: 改善率の比較

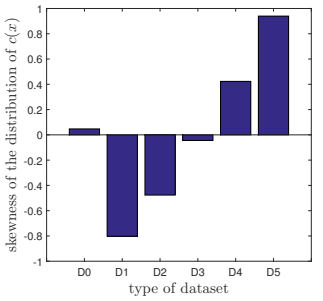

図 16: $c(x)$ の分布の歪度

似度は $\rho(u,v) = 1 - d(u,v)/2$ とした．また，ここで示す図の縦軸の値は全て 100 個のデータにおける実験結果の平均値である．

まず，各データタイプにおける A1 と A2 の解品質を図 13, 14 に示す．ここで解品質としている図の縦軸は，各クラスターにおける代表オブジェクトと各オブジェクトの類似度の最大値の総和 $\sum_{v\in\mathcal{H}}\mu(v\,;\,\hat{\mathcal{P}}_K)$ である．両図だけでは局所改善の適用による解品質の変化が分かりにくいため，解品質の比 (A2 / A1) をとって改善率とし，その比較を図 15 に示す．さらに，各データタイプのオブジェクト中心性 $c(x)$ の分布の歪度を図 16 に示す．図より，$c(x)$ の分布の歪度が大きく負（右に偏っている）となっている D1 と D2 は局所改善の効果が最も高く，$c(x)$ の分布の歪度が 0（左右対象）に近い D0 と D3 は局所改善が中程度に効いており，$c(x)$ の分布の歪度が大きく正（左に偏っている）となっている D4 と D5 は局所改善の効果が最も低いことが分かる．

次に，各データタイプにおける A1 と A2 のクラスタリング結果の例 ($K = 5$) を図 17 から 28 に示す．図より，先程の図 15 における改善率が高いデータタイプの代表オブジェクトは A1 と A2 で大きく違っており，逆に改善率が低いデータタイプの代表オブジェクトは A1 と A2 であまり変化がないことが見て取れる．

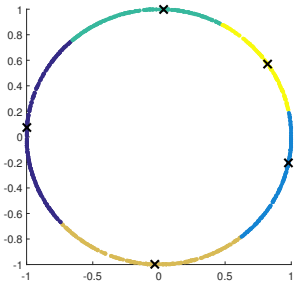

図 17: D0 と A1 の例

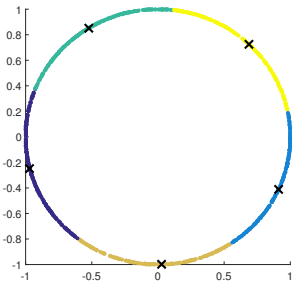

図 18: D0 と A2 の例

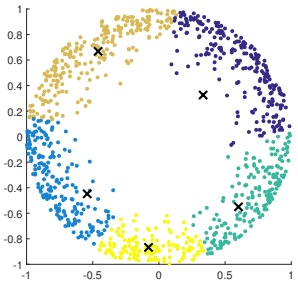

図 19: D1 と A1 の例

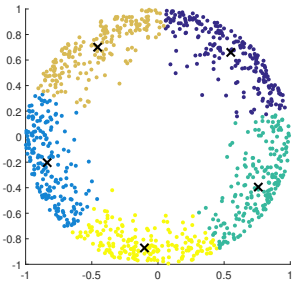

図 20: D1 と A2 の例

5. おわりに

K-medoids クラスタリングにおける局所改善の有効性の推定を目的として，オブジェクト集合に対してオブジェクトの中心性指標を導入し，その分布の歪度を有効性の指標とすることを試みた．今回の実験ではその関係性が明確に見られたが，2 次元の円上という限られた条件下であったため，今後は多次元や多様な分布におけるデータを用いて実験を行う予定である．

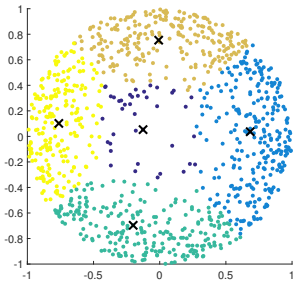

図 21: D2 と A1 の例

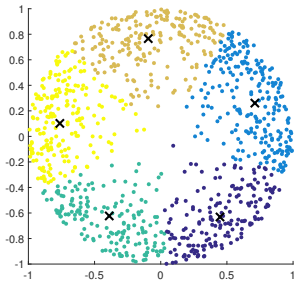

図 22: D2 と A2 の例

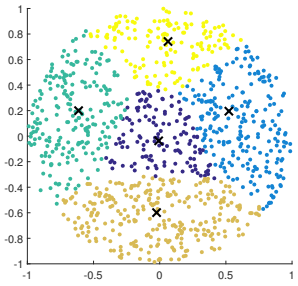

図 23: D3 と A1 の例

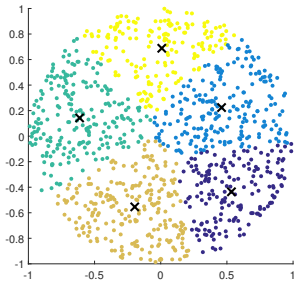

図 24: D3 と A2 の例

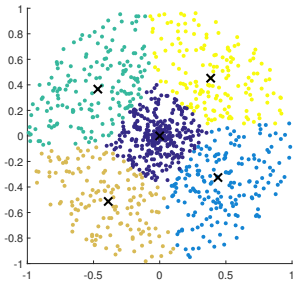

図 25: D4 と A1 の例

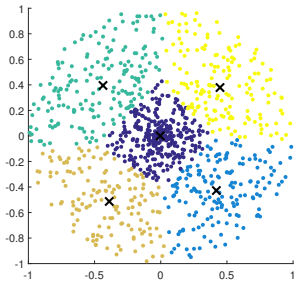

図 26: D4 と A2 の例

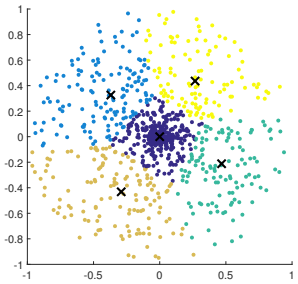

図 27: D5 と A1 の例

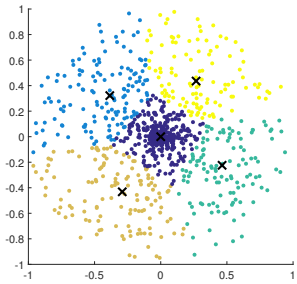

図 28: D5 と A2 の例

謝辞

本研究は， JSPS 特別研究員奨励費 15K00311 の支援を受けて行ったものである．

参考文献

[1] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis*. John Wiley & Sons (1973).

[2] A. P. Dempster, N. M. Laird and D.B. Rubin. Maximum likelihood from incomplete data via EM algorithm, *J. Royal Statist. Soc. Ser. B (methodology)* , Vol. 39, pp. 1–38 (1977).

[3] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press (1996).

[4] A. Krause and C. Guestrin. Near-optimal Nonmyopic Value of Information in Graphical Models, *Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence*, pp. 324–331 (2005).

[5] D. Kempe and J. Kleinberg and E. Tardos. Maximizing the spread of influence through a social network, *Proceedings of the 9th International Conference on Knowledge Discovery and Data Mining*, pp. 137–146 (2003).

[6] G. Nemhauser, L. Wolsey and M. Fisher. An analysis of the approximations for maximizing submodular set functions, *Mathematical Programming*, **14** pp. 265–294 (1978).