Vol.2016-DBS-163 No.6
Vol.2016-IFAT-123 No.6
2016/9/13

距離分布の形状分布に基づくオブジェクトの特徴推定

山岸 祐己[1,a)]　斉藤 和巳[1,b)]

概要：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のクラスタリングの解法は種々存在しているが，共通する問題の一つとして，計算コストや必要なメモリ容量がオブジェクト数に依存することが挙げられる．よって，計算過程において探索するべきオブジェクトの特徴，または探索するべきでないオブジェクトの特徴を見出し，それらを事前にオブジェクト集合から得られる情報から推定できれば，計算コストや使用メモリ容量の削減を図ることができると考えられる．今回我々は，各オブジェクトから見たときの他のオブジェクトとの距離総和と，距離分布の歪度と尖度に着目し，クラスタリングにおいて重要となるオブジェクトの特徴を数値化することを試みる．

キーワード：k-medoids, k-median, 距離分布, 歪度, 尖度

Feature Estimation of Objects Based on a Shape Analysis of Distance Distributions

Yuki Yamagishi[1,a)]　Kazumi Saito[1,b)]

Abstract: There are various solutions for clustering problems in cases that objects are given as certain vectors. However, these solutions have a common difficulty which the computational complexity and required memory capacity are depending on a number of objects. Therefore, if we can estimate characteristics of objects which must be searched or not from the prior information, the computational complexity and the required memory capacity will be able to reduce according to the reduction rate of searching objects. Here, we focus on the sum of distances, the skewness, and the kurtosis of the distance distribution of viewpoints of each object, and to attempt to characterize the objects which have key roles in clustering problems.

Keywords: k-medoids, k-median, distance distribution, skewness, kurtosis

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-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-medoids（または k-median ）法と呼ばれており，一般に，k-medoids 問題として定式化すれば，外れ値などに頑健であることが知られている [3]．k-means 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-medoids 法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-medoids 法に着目する．

k-medoids 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．近似解を求める代表的手法は反復アルゴリズムの類であるが，一般に計算量は $O(N^2k)$ の反復回数倍となり，解もその都度一意に求まるわけではないため，大規模な問題になるほど膨大な計

[1] 静岡県立大学
University of Shizuoka, Shizuoka 422–8526, Japan
[a)] yamagissy@gmail.com
[b)] k-saito@u-shizuoka-ken.ac.jp

Vol.2016-DBS-163 No.6
Vol.2016-IFAT-123 No.6
2016/9/13

算コストを要することになる．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一意の近似解により，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．貪欲法のみでは比較的ア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局所改善法も挙げられる．しかし，どちらの手法も根本的には反復手法と同様の探索操作を行うため，貪欲法では $O(N^2k)$，局所改善法では $O(N^2k)$ の数倍程度の計算量が必要であり，依然としてオブジェクト数 N によって計算コストが膨れ上がる恐れがある．よって，今回我々は，探索オブジェクトの大幅な削減を目的として，各オブジェクトのクラスタリングにおける重要性を推定する分析手法を提案する．提案手法は，各オブジェクトから見た他オブジェクトとの距離総和と，それら距離分布の歪度 (skewness) と尖度 (kurtosis) を，各オブジェクトのクラスタリングにおける重要性指標としている．分析結果では，それらを用いた単純な線形回帰で最終的なクラスタリング結果を推定できるか，また，データの規模によってその特徴に変化がないかを検証する．

2. オブジェクト間距離の特徴

H 次元のベクトルとして与えられる N 個のオブジェクト集合 $\mathcal{N}$ に対し，任意のオブジェクトペア $\alpha, \beta \in \mathcal{N}$ 間の距離 $d(\alpha, \beta)$ を定義すると，任意のオブジェクト $\gamma \in \mathcal{N}$ から見たときの他オブジェクトとの距離総和 $\sum_{n=1}^{N} d(\gamma, n)$，距離分布の歪度 $\sum_{n=1}^{N}(d(\gamma, n) - \mu)^3/N\sigma^3$，距離分布の尖度 $\sum_{n=1}^{N}(d(\gamma, n) - \mu)^4/N\sigma^4$ が求まる．ここで，μ は距離平均 $\sum_{n=1}^{N} d(\gamma, h)/N$ を，σ は距離の標準偏差 $\sqrt{\sum_{n=1}^{N}(d(\gamma, n) - \mu)^2/N}$ をそれぞれ表す．いま，単位 H 次元球内に一様分布するように乱数で発生させたオブジェクト集合を考え，オブジェクト間距離を H 次元ユークリッド距離として定義すると，各オブジェクト視点の他オブジェクトとの距離総和，距離分布の歪度，距離分布の尖度の関係は図 1–12 のようになる．図からわかるように，2 次元や 3 次元のときは，距離総和と歪度と尖度は各オブジェクトに多様な特徴を付与しており，100 次元や 1000 次元のときは，これら 3 指標は特徴としてあまり機能していない．これは，球面集中現象 (concentration on the sphere) [7] によって各オブジェクトの距離分布の特徴が薄れたことによるものであると考えられる．同様の現象は，一様分布以外の分布においても見られるため．裏を返せば，この 3 指標によって特徴が上手く付与できない場合は，クラスタリング問題を設定することが望ましくないことが窺える．

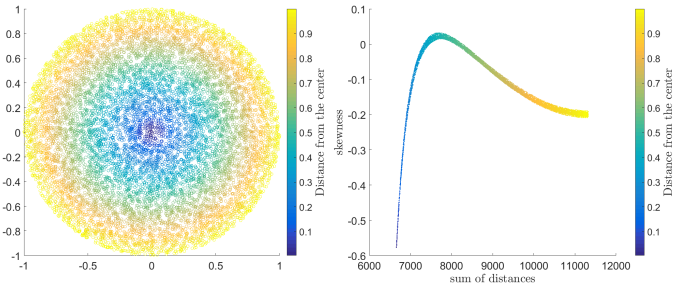

図 1 単位円上に一様分布しているオブジェクトの例

図 2 各オブジェクト視点の距離の総和と距離分布の歪度（単位円）

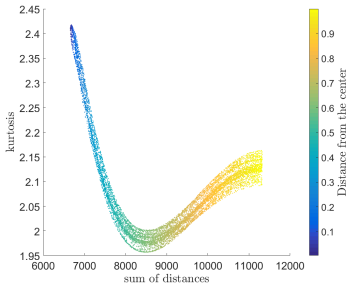

図 3 各オブジェクト視点の距離の総和と距離分布の尖度（単位円）

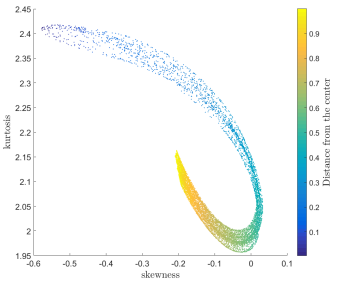

図 4 各オブジェクト視点の距離分布の歪度と尖度（単位円）

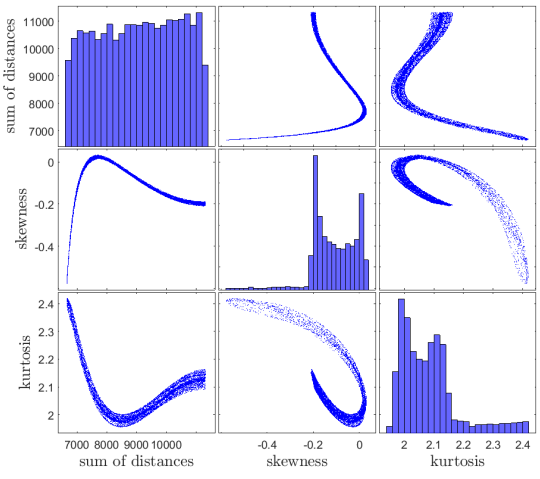

図 5 各オブジェクト視点の距離の総和，距離分布の歪度，尖度の関係（単位円）

Vol.2016-DBS-163 No.6
Vol.2016-IFAT-123 No.6
2016/9/13

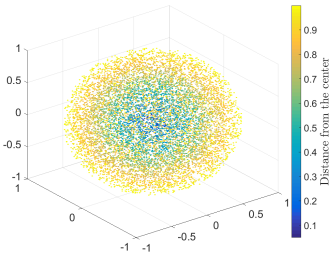

図 6 単位球上に一様分布しているオブジェクトの例

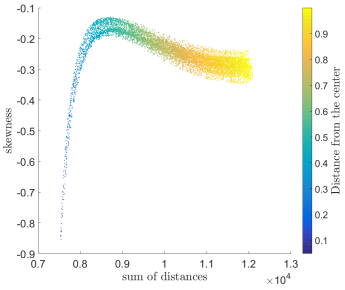

図 7 各オブジェクト視点の距離の総和と距離分布の歪度（単位球）

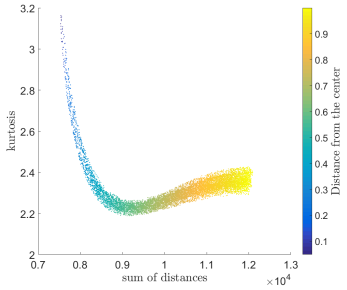

図 8 各オブジェクト視点の距離の総和と距離分布の尖度（単位球）

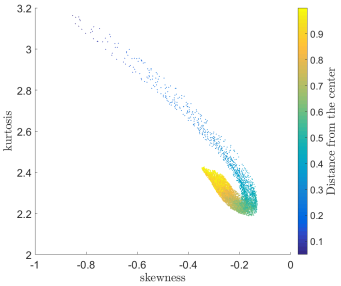

図 9 各オブジェクト視点の距離分布の歪度と尖度（単位球）

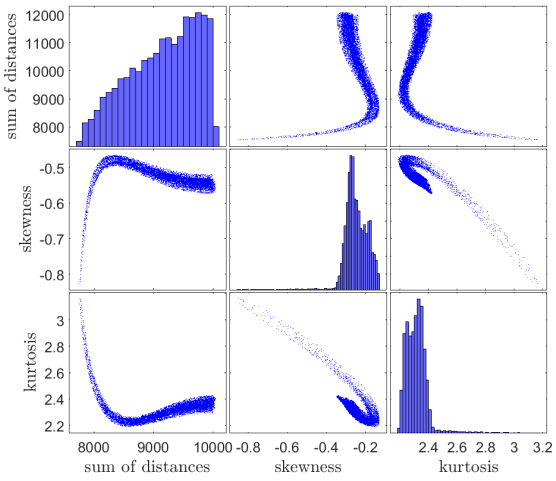

図 10 各オブジェクト視点の距離の総和，距離分布の歪度，尖度の関係（単位球）

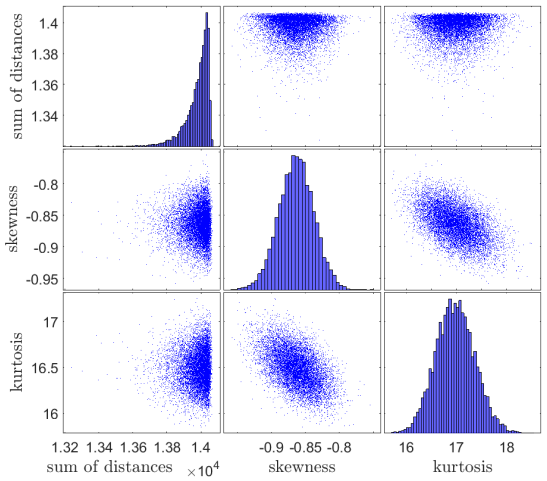

図 11 各オブジェクト視点の距離の総和，距離分布の歪度，尖度の関係（単位 100 次元球）

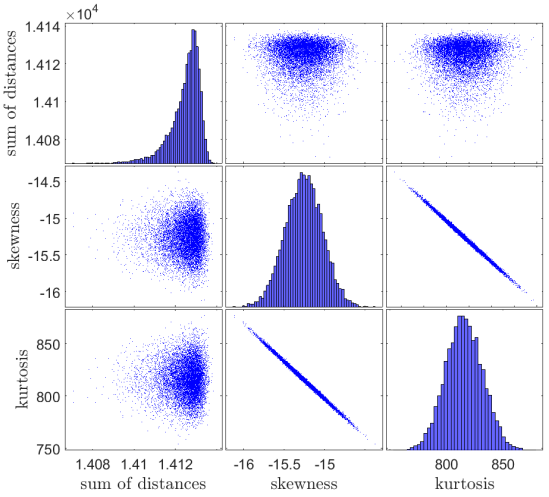

図 12 各オブジェクト視点の距離の総和，距離分布の歪度，尖度の関係（単位 1000 次元球）

3. 分析手法

今回提案する分析手法の大まかな流れは以下となる．

Step 1. k-medoids 問題を近似アルゴリズムで解く

Step 2. 各オブジェクト視点の距離分布の指標を用いて近似解を線形回帰

各ステップの詳細を以下に述べる．

3.1 k-medoids 問題の解法

k-medoids 法は，非階層クラスタリングで有名な k-means 法と同様に，N 個のオブジェクト集合 $\mathcal{N}$ 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 $u, v \in \mathcal{N}$ 間に距離 $d(u, v)$ 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距離の総和が小さ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，距離が近いオブジェクトペアは同じクラスターに，距離が遠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，前述のとおり，k-medoids 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．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，最も距離が近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．

$$f(\mathcal{P} \cup \{w\}) = \sum_{v \in \mathcal{N}} \min\{\mu(v\,; \mathcal{P}), d(v, w)\}. \qquad (1)$$

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距離の最小値を表し，$\mu(v\,; \mathcal{P}) = \min_{w \in \mathcal{P}}\{d(v, w)\}$ 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．なお，$\cdot \backslash \cdot$ は集合差を表し，最終的なクラスター数は K とする．

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

A1-2. 式 1 で $\hat{p}_k = \arg\min_{w \in \mathcal{N} \backslash \mathcal{P}_{k-1}}\{f(\mathcal{P}_{k-1} \cup \{w\})\}$ 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，最も距離が近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 $v \in \mathcal{N}$ に対し，$\mu(v\,; \mathcal{P}_k)$ を求める；

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 $O(N^2K)$ 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，比較的ア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 ${p'}_k = \arg\min_{w \in \mathcal{N} \backslash \hat{\mathcal{P}}_K \backslash \{\hat{p}_k\}}\{f(\hat{\mathcal{P}}_K \backslash \{\hat{p}_k\} \cup \{w\})\}$ 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 $h \leftarrow 0$, $\hat{\mathcal{P}}_K = \hat{\mathcal{P}}_K \backslash \{\hat{p}_k\} \cup \{{p'}_k\}$ 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，最も距離が近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が，我々の経験上 $O(N^2K)$ の数倍程度であることが分かっている．

3.2 線形回帰の設定

最終的なクラスタリング結果 $\hat{\mathcal{P}}_K$ における，任意のオブジェクト v の最も距離が近い代表オブジェクトとの距離 $\min_{w \in \hat{\mathcal{P}}}\{d(v, w)\}$ を，オブジェクト v 視点の距離総和 $x_{v,1}$ と距離分布の歪度 $x_{v,2}$ と尖度 $x_{v,3}$ から推定できると考えると，以下のような線形回帰問題が設定できる．

$$h(\mathbf{x}_v|\boldsymbol{\theta}) = \theta_0 + \theta_1 x_{v,1} + \theta_2 x_{v,2} + \theta_3 x_{v,3}. \qquad (2)$$

この問題に対して最小二乗モデルを仮定すれば，回帰係数 $\boldsymbol{\theta}$ は正規方程式を解くことで得られる．

4. データセット

分析では，手書き数字画像データセットの MNIST[*1] を使用した．画像の次元数は 784 (28x28 pixels), 各次元の値は 0 から 255 の整数である．今回，結果をデータの規模で比較するため，test set （サンプル数 10,000） と training set （サンプル数 60,000）のそれぞれで分析を行い，オブジェクト間距離 $d(u, v)$ は L1 距離（マンハッタン距離）とコサイン距離（1-コサイン類似度）の 2 つを採用した．各データと各距離定義における各オブジェクト視点の距離の総和，距離分布の歪度，尖度の関係を図 13–16 に示す．図

[*1] http://yann.lecun.com/exdb/mnist/

Vol.2016-DBS-163 No.6
Vol.2016-IFAT-123 No.6
2016/9/13

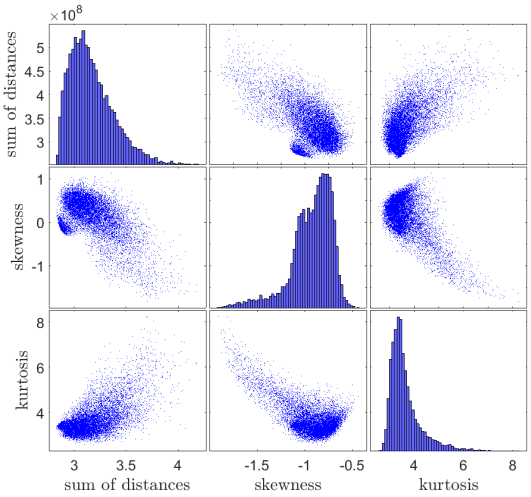

図 13 各オブジェクト視点の距離の総和，距離分布の歪度，尖度の関係（MNIST test set, L1 distance）

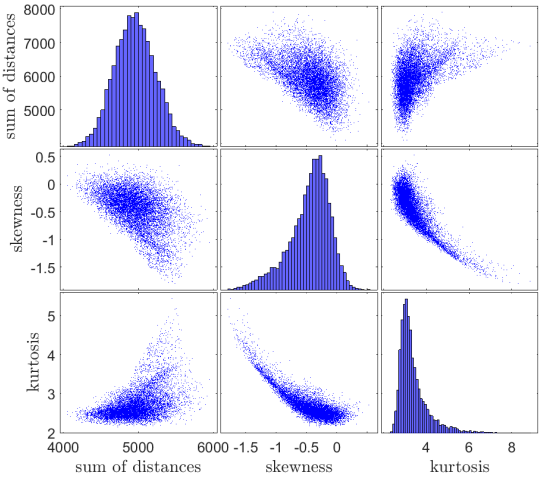

図 14 各オブジェクト視点の距離の総和，距離分布の歪度，尖度の関係（MNIST test set, cosine distance）

より，これら 3 指標の関係は規模によってあまり変化することがないことが分かる．更に，L1 距離の各散布図ではオブジェクトが一極集中しやすい場所が見受けられるが，コサイン距離の各散布図ではそういった傾向があまり見られないため，今回の場合はコサイン距離の方が L1 距離よりも各オブジェクトに特徴を持たせるのに適していることが窺える．

5. 分析結果

提案手法による分析結果を図 17–24 に示す．まず，データの規模によって結果に大きな変化がないことは明らかである．また，どの結果においても，距離の総和と距離分布の歪度の重要性は高く，距離分布の尖度の重要性が低いことが分かる．代表オブジェクトからの最短距離の推定においては，どの結果においても概ね再現ができているように

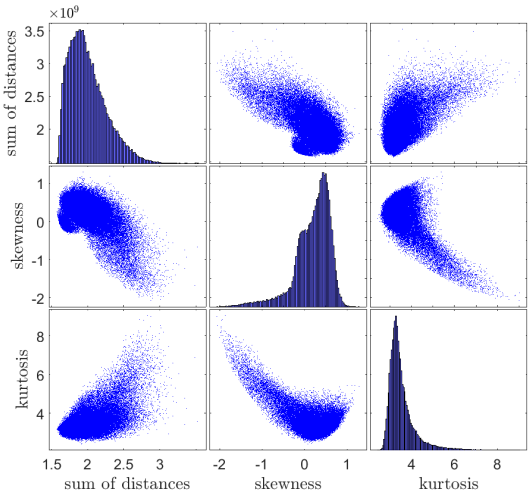

図 15 各オブジェクト視点の距離の総和，距離分布の歪度，尖度の関係（MNIST training set, L1 distance）

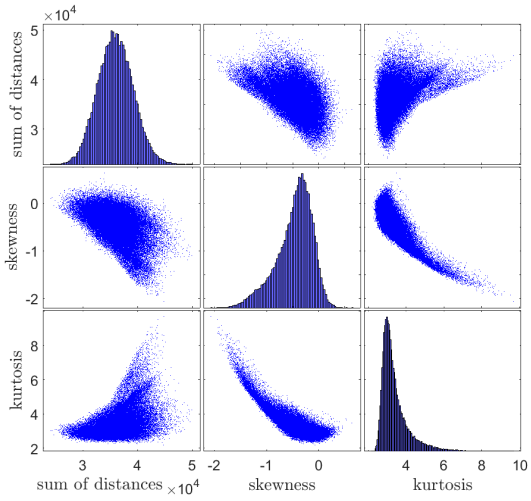

図 16 各オブジェクト視点の距離の総和，距離分布の歪度，尖度の関係（MNIST training set, cosine distance）

見えるが，L1 距離を採用したものは，縦軸上にプロットされている実際の代表オブジェクトの推定が不安定であることが見て取れる．

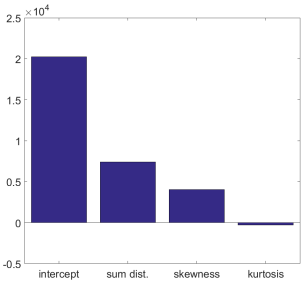

図 17 正規化後の回帰係数（mnist test set, L1 distance）

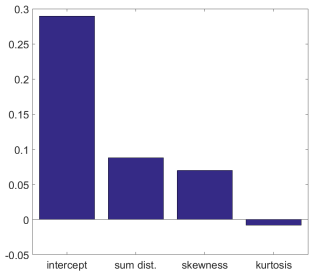

図 18 正規化後の回帰係数（mnist test set, cosine distance）

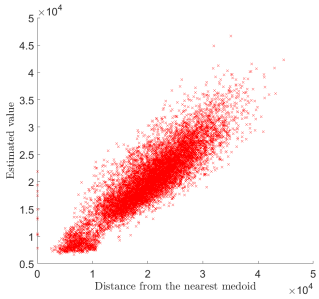

図 19 代表オブジェクトからの最短距離の推定（mnist test set, L1 distance）

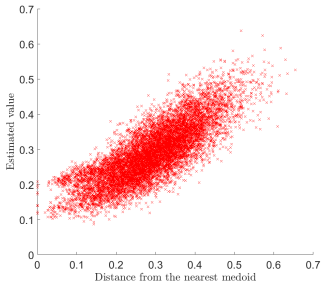

図 20 代表オブジェクトからの最短距離の推定（mnist test set, cosine distance）

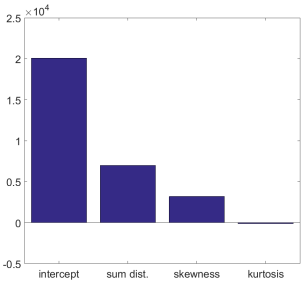

図 21 正規化後の回帰係数（mnist training set, L1 distance）

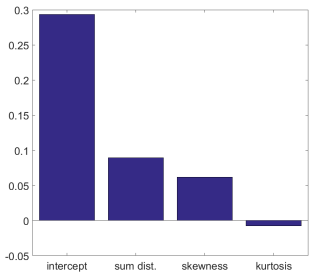

図 22 正規化後の回帰係数（mnist training set, cosine distance）

6. まとめ

各オブジェクトのクラスタリングにおける重要性を推定するため，各オブジェクトから見た他オブジェクトとの距離総和と，それら距離分布の歪度と尖度を用いた分析手法を提案した．提案分布手法の結果から，実際のクラスタリング結果の推定において，距離総和と距離分布の歪度の重要性は高いことが分かった．また，クラスタリング結果の全体的な再現は実現できたが，距離定義によっては代表オブジェクトの推定にばらつきが生じることも分かった．今後は，多様なデータに対して同様の分析を行い，クラスタリング結果の再現性を高めるべく，提案手法を改良する予定である．

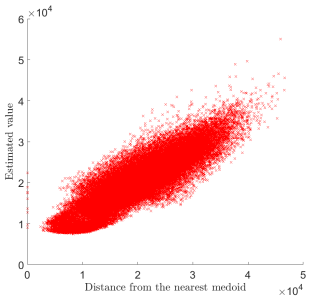

図 23 代表オブジェクトからの最短距離の推定（mnist training set, L1 distance）

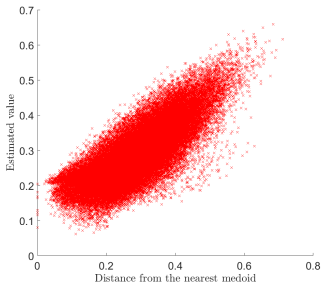

図 24 代表オブジェクトからの最短距離の推定（mnist training set, cosine distance）

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).

[7] A. A. Giannopoulos, and V. D. Milman. Concentration Property on Probability Spaces *Advances in Mathematics*, Vol. 156, issue 1, pp. 77–106 (2000)