SSPW k-means Clustering のセントロイドの分散に基づくスパース化割合の決定

早稲田大学 基幹理工学部 情報通信学科　学部 3 年　1w193077　田淵 智久

概要

本稿では，2021 年度プロジェクト研究 A で行った研究活動内容とその成果について報告する．

1. プロジェクト研究における活動内容の概要

プロジェクト研究 A では，以下の取り組みを行った．

- 論文精読
  - Sparse Simplex Projection Wasserstein k-means Clustering
  - Computational Optimal Transport, §1,2,4
  - An introduction to continuous optimization for imaging, §1,2,3
- SSPW k-means Clustering の再実験
  - 特にスパース化割合の変化によってクラスタリグ性能がどのように変化するかを確認した．
- 入力画像の確率分布の変形についての検討
  - COIL-100 データセットにおいて, 入力画像の背景が画素 24 で埋められており, その点における割合が大きかった．画像における背景の割合は物体の撮影された位置に依存し, 物体の特徴を表す情報を保持していないのではないかと考え, 画素 24 の確率の重みを 0 にして実験した．結果としては, クラスタリング性能は悪化した．
  - COIL-100 データセットにおいて, 入力画像の確率分布を異なるクラスで比較したときに, 物体によって画素のとりうる範囲が少し異なることが確認できた．クラスタリングにおいて高い画素が重要な指標になっているのではないかと考え, 高い画素の確率の重みを何倍か増幅させて実験した．どの範囲を何倍増幅させるかのパラメータをいくつか試したが, ほとんどの場合においてクラスタリング性能が悪化した．
- 最尤推定に基づく正規分布への近似と交差エントロピー正則化項の付与
  - 入力画像の確率分布を正規分布に変換した．画素 24 の割合が極端に大きかったため, 画素 24 の確率の重みを 0 としてから正規分布に近似することも行なった．最尤推定に基づいて単峰性の正規分布に近似したが, 元の確率分布の構造を消してしまっており, クラスタリング性能も大きく悪化した．より良いフィッティング手法として最尤推定時に確率分布の近さを表現できる交差エントロピーを正則化項として付与することを検討した．正則化パラメータを変更しても元の確率分布の構造を保持するような変換はできずクラスタリング性能も大きく悪化した．
- スパース化割合を動的に決定することについての検討
  - SSPW k-means Clustering[1] において, データをスパース化してシンプレックス構造に射影する手法について, スパース化割合を入力画像データの確率分布及びクラスターのセントロイドに応じて決定する手法ついて検討した．

以下では，SSPW k-means Clustering[1] の概要について説明し，また, 交差エントロピーを考慮した最尤推定に基づく正規分布への近似とスパース化割合を動的に決定する手法について説明する．

2. Sparse Simplex Projection Wasserstein k-means Clustering [1]

- 背景（応用，アプリケーション）
  k-means クラスタリングは, シンプルかつ強力なアルゴリズムであり, コンピュータビジョン, 統計, ロボット工学, バイオインフォマティクス, 機械学習, 信号処理, 医用画像工学など, さまざまな分野で広く利用されている．
- 従来の研究動向概要　[1]
  Lloyd の k-means アルゴリズム [2](**Algorithm 1**) は代入ステップと更新ステップからなり, データ点と全ての重心点との距離を調べ, 最も距離が小さい重心点のラベルが割り当てられる．データ点の数を q, 重心点の数を k とすると, データ点と重心点の全ての組み合わせについて距離を計算することから, 計算量は $\mathcal{O}(qk)$ と非常に大きい．また, Lloyd の k-means アルゴリズムではヒストグラムの潜在的な構造を無視してしまう性質がある．ヒストグラムの潜在的な構造を考慮したアプローチとして, Wasserstein 距離を用いた Wasserstein k-means アルゴリズム (**Algorithm 2**) がある．各ピクセルが 0 から 255 の画素をとるグレースケール画像を例にとって説明する．Lloyd の k-means アルゴリズムでは 256 次元のユークリッド空間でのデータ点と重心点との距離, すなわち各画素の確率の重みの差の二乗の総和を考えて, データ点から最も近い重心点を定めた．一方, Wasserstein k-means アルゴリズムでは, コスト関数を l_2 ノルムとすると, データ点の各画素の確率の重みを重心点の確率分布となるように輸送することを考えるので, Wasserstein 距離は色の分布の近さを表現できる．この Wasserstein 距離を用いた Wasserstein k-means アルゴリズムは多くのアプリケーションで高い性能を示し利用されている．しかし, Wasserstein k-means アルゴリズムは各イテレーションで最適化問題を解くことから Lloyd の k-means アルゴリズムよりも高い計算コストを要する．そこで計算コストを抑える方法としてデータのスパース性を用いる手法 (**Algorithm 3**) がある．クラスタリング性能を下げることなくデータサンプル, 重心点, コスト行列をスパースなシンプレックス構造に射影し縮退させることで計算コストを減らす．SSPW k-means Clustering[1] では, スパース化割合を各イテレーションで固定, 増加, 減少させており, スパース化割合を固定させる方法が最も良い結果となったことを示している．
- 本稿の視点・問題点
  SSPW k-means Clustering [1] ではデータをスパースなシンプレックス構造に射影することで計算コストを減らした．この手法では, スパース化割合はデータ点や重心点の確率分布の構造を考慮することなく決定していた．スパース化割合はスパース化する対象の確率分布によって決定されるべきである．確率の重みが確率測度の台の一部分に偏っている場合は, スパース化割合を小さくすることが望ましい．逆に, 確率の重みが一様分布のように確率測度の台の広い範囲にわたって分布している場合はスパース化割合を大きくすることが望ましいと考えられる．ただし, スパース化割合とは重みが 0 でない確率測度の台における, スパース化した後に残る確率測度の台の割合を指している．

- 本稿の提案内容の概要
  SSPW k-means Clustering [1] ではデータセットをそのまま使用したが, 画像データの確率分布を正規分布に近似する手法を示す. 正規分布同士の間の距離は閉形式で求まることが知られていることから, 最適輸送を考えやすくなる. さらに, 正規分布に近似することでスパース化割合が考えやすい. また, セントロイドの構造を考慮してスパース化割合を決定する手法を提案する. この手法ではセントロイドの確率分布の分散を調べ, 分散が大きければスパース化割合を大きくし, 分散が小さければスパース化割合を小さくすることを考える.
- 表記法の整理
  k-means アルゴリズムにおける重心点をセントロイドと言うことがある. また, 確率測度の台とは確率が定義される確率変数の取る値を指しており, グレースケール画像のデータの場合, それは 0 から 255 のピクセルの画素を表す. データ点の数を q, クラスタの数を k とする.

Algorithm 1 Euclidean k-means

Require: data $\mathbf{X} = \{x_1, \ldots, x_q\} \subset \mathbb{R}^d$, cluster number $k \in \mathbb{N}$.
Initialize randomly centroids $\mathbf{C} = \{\tilde{c}_1, \ldots, \tilde{c}_k\} \subset \mathbb{R}^d$.
repeat
  Find closest centroids (assignment step):
    $s_i = \arg\min_{j=1,\ldots,k} ||x_i - c_j||_2^2, \forall i \in [q]$
  Update centroids (update step):
    c_j =mean$(x \in X | s_i = j), \forall j \in [k]$
until cluster centroids stop changing.
Ensure: tmp

Algorithm 2 Wasserstein k-means

Require: data $\{\nu_1, \ldots, \nu_q\}$, cluster number $k \in \mathbb{N}$, ground cost matrix $\mathbf{C} \in \mathbb{R}^{n \times n}$, maximum number T_{max}, γ_{min}.
Initialize centroids $\{\tilde{c}_1, \ldots, \tilde{c}_k\}$.
repeat
  Find closest centroids (assignment step):
    $s_i = \arg\min_{j=1,\ldots,k} W_p(\nu_i, c_j), \forall i \in [q]$
  Update centroids (update step):
    c_j =barycenter$(\{\nu | s_i = j\})$, $\forall j \in [k]$
until cluster centroids stop changing.
Ensure: cluster centers $\{c_1, \ldots, c_k\}$.

3. 交差エントロピーを考慮した最尤推定に基づく正規分布フィッティング

3.1. 概要

- 最尤推定に基づいて画像の確率分布を正規分布に変換することを考えた. 変換した後の分布を変換する前の分布と比べ, 近似できているかを確認した. 変換したデータを用いてクラスタリング性能が向上するかを調べた.

3.2. 詳細

- 最尤推定に基づく正規分布へのフィッティング
  確率変数 $x_i (i = 0, 1, \ldots, 255)$ はそれぞれ 0 から 255 の値を取る. 確率変数 x_0 は平均 μ, 分散 σ^2 の正規分布に従うとする.

$$N(x_0 | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_0 - \mu)^2}{2\sigma^2}\right) \quad (1)$$

Algorithm 3 SSPW k-means

Require: data $\{\nu_1, \ldots, \nu_q\}$, cluster number $k \in \mathbb{N}$, ground cost matrix $\mathbf{C} \in \mathbb{R}^{n \times n}$, maximum number T_{max}, γ_{min}.
Initialize centroids $\{\tilde{c}_1, \ldots, \tilde{c}_k\}$, set $t = 1$.
repeat
  Update sparsity ratio $\gamma(t)$
  Project ν_i to $\hat{\nu}_i$ on sparse simplex Δ_p:
    $\hat{\nu}_i$ =Proj$^{\gamma(t)} \forall i \in [q]$
  Shrink $\hat{\nu}_i$ into $\tilde{\nu}_i$: $\tilde{\nu}_i$ = shrink$(\hat{\nu}_i)$
  Shrink $\hat{c}_i$ into $\tilde{c}_i$: $\tilde{c}_i$ = shrink$(\hat{c}_i)$
  Shrink ground cost matrix $\mathbf{C}$ into $\hat{\mathbf{C}}$: $\hat{\mathbf{C}}$ =Shrink$(\mathbf{C})$
  Find closest centroids (assignment step):
    $s_i = \arg\min_{j=1,\ldots,k} W_p(\nu_i, c_j), \forall i \in [q]$
  Update centroids (update step):
    c_j =barycenter$(\{\nu | s_i = j\})$, $\forall j \in [k]$
until cluster centroids stop changing. StateUpdate the iteration number t as $t = t + 1$
Ensure: cluster centers $\{c_1, \ldots, c_k\}$.

その他の確率変数 $x_i (i = 0, 1, \ldots, 255)$ も平均 μ, 分散 σ^2 で互いに独立である正規分布に従うとする. この時, $x_0, x_1, \ldots, x_{255}$ をサンプルする確率 (尤度関数) は次のようになる. ただし,$n = 255$ とする.

$$L(\mu, \sigma^2) = \prod_{i=0}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \quad (2)$$

ここで得られたサンプル $x_i (i = 0, 1, \ldots, 255)$ の確率分布の平均 μ と分散 σ^2 を推定する. 尤度関数が最大となるような平均 μ が最もらしい値であると言うことができるので, その値を平均 μ の推定値とする. 分散 σ^2 についても同様にして推定することができる. よって, 尤度関数に log を施し, 平均 μ, 分散 σ^2 について偏微分して尤度を最大化するときの平均 μ, 分散 σ^2 を求める.

$$\begin{aligned} l(\mu, \sigma^2) &= \log L(\mu, \sigma^2) \\ &= -\frac{1}{2} \sum_{i=0}^{n} \left(\frac{(x_i - \mu)^2}{\sigma^2} - \log(2\pi) - \log \sigma^2 \right) \end{aligned} \quad (3)$$

平均 μ, 分散 σ^2 それぞれにおいて偏微分して 0 とおけば推定値が求まる.

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i \quad (4)$$

$$\sigma^2 = \frac{1}{n} \sum_{i=0}^{n} (x_i - \mu)^2 \quad (5)$$

- 交差エントロピーを考慮した最尤推定に基づく正規分布フィッティング
  上の正規分布へのフィッティング手法は確率変数 x_i が正規分布に従っていることを仮定したものである. また, 確率測度の台の重みを無視している. 元々の画像データの確率分布の構造を保持しながら正規分布に近似したい. そこで, 2 つの確率分布の近さを表す交差エントロピーを最尤推定するときに正則化項として付与することで確率

測度の重みを考慮した近似ができることが期待される.

$$\begin{aligned}
&\log L(\mu,\sigma^2) - \lambda H(a, N(\mu,\sigma^2)) \\
&= \log L(\mu,\sigma^2) - \lambda \sum_{i=0}^{n} a_i \log N(x_i|\mu,\sigma^2) \\
&= -\frac{1}{2}\sum_{i=0}^{n}(-\lambda a_i + 1)\left(\frac{(x_i-\mu)^2}{\sigma^2} - \log(2\pi) - \log\sigma^2\right)
\end{aligned}$$

平均 μ, 分散 σ^2 それぞれにおいて偏微分して 0 とおけば推定値が求まる.

$$\mu = \frac{\sum_{i=0}^{n}(-\lambda a_i + 1)x_i}{\sum_{i=0}^{n}(-\lambda a_i + 1)} \tag{6}$$

$$\sigma^2 = \frac{\sum_{i=0}^{n}(-\lambda a_i + 1)(x_i-\mu)^2}{\sum_{i=0}^{n}(-\lambda a_i + 1)} \tag{7}$$

得られた平均 μ, 分散 σ^2 を用いて, ガウス確率密度関数に代入し, 正規分布を生成した. そして x が 0 から 255 の離散点でプロットした. この後確率の総和が 1 となるように正規化した.

4. セントロイドの構造に依存したスパース化割合の決定

4.1. 概要

- SSPW k-means アルゴリズムではイテレーションごとにセントロイドが更新される. イテレーションごとに, 射影する前にそれぞれのセントロイドについて分散を調べ, 分散に応じてスパース化割合を決定した.

4.2. 詳細

- 分散に応じたスパース化割合の決定
  セントロイドの分散を σ^2, 調整するパラメータを λ とし, それぞれ正の値とする. まず, スパース化割合を次のようにした.

$$\gamma_1 = \exp(-\lambda\sigma^2) \tag{8}$$

これは分散が大きくなるとスパース化割合が小さくなるという性質を持つ. 次にスパース化割合を次のようにした.

$$\gamma_2 = 1 - \exp(-\lambda\sigma^2) \tag{9}$$

これは分散が大きくなるとスパース化割合が大きくなるという性質を持つ. 最後に次のようなスパース化割合に設定して実験した.

$$\gamma_3 = \frac{1}{1+\exp(-\lambda\sigma^2)} \tag{10}$$

スパース化割合が 0 とならないように, スパース化割合が 0.4 より小さくなる場合はスパース化割合を 0.4 と設定した.

5. 数値実験

5.1. 実験条件

- 使用するデータセット
  COIL-100 データセットを使用して実験を行う. このデータセットはオブジェクトが中央にある画像のデータセットであり, 100 クラスのオブジェクトが各クラス 72 枚ずつ計 7200 枚ある. 72 枚の画像はオブジェクトを回転させて, あらゆる方向から撮影された画像となっている. そして画像は 32×32 のピクセルからなり, 各ピクセルは 0 から 255 の画素をとるようなグレースケール画像である. オブジェクトの背景は画素が 24 の値で埋められている. データセットの入手は次の URL [*1]から取得できる.
- 実験条件
  画像の 32×32 の行列を列ごとに分解し結合することで 1024×1 の行列を作る. この 1024×1 の行列から確率測度の台を 0 から 255 とするヒストグラムを作成する. この時, 確率測度の台に定まっている確率の重みの合計が 1 となるように正規化する. 100 クラスのオブジェクトからランダムに 10 クラス選択し, 各クラス 3 枚ずつサンプルして, 合計 30 枚の画像を 10 個のクラスに分類するタスクを行う.
  - 重心の初期位置は Lloyd's の k-means アルゴリズムによって得られた点を使用する.
  - Sparse Simplex Projection アルゴリズムは GSHP を使用する.
  - ラベルの更新をする時, データ点から最も近いセントロイドを探すアルゴリズムは Matlab の線形計画法ソルバー linprog を使用する.
  - barycenter を求めるアルゴリズムは Sinkhorn アルゴリズムを使用する. この時, 正則化パラメータは 0.1000 とする.
  - スパース化割合が 0.4 より小さくなる場合, スパース化割合を 0.4 と設定し, スパース化割合が 0.9 を超える場合, スパース化割合を 0.9 と設定した. パラメータ λ はスパース化割合が 0.4 から 0.9 の間を変動するような値を探して, その周辺の値を使用した.
- 比較方式について
  スパース化割合を 0.7 に固定した場合とスパース化割合のパラメータ λ を 3 あるいは 4 通り変化させた場合の実験を行なった. データは 5 回サンプルし直して実験し, その平均の値を使用する. スパース化割合を 0.7 に固定した場合よりも提案手法がクラスタリング性能を向上させるかを調べる. スパース化割合の設定方法は γ_1, γ_2, γ_3 について実験した.

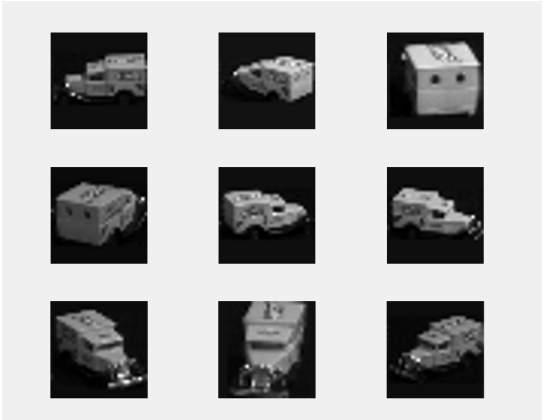
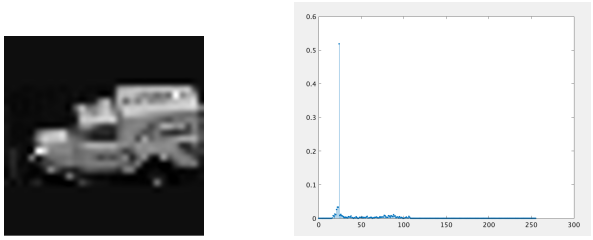


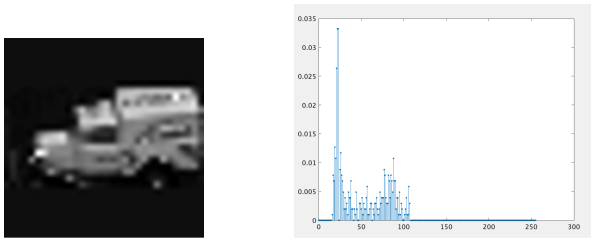
図1: COIL-100 データセット: あるクラスの画像を表している. オブジェクトを回転させた画像となっている.

[*1] https://www.cs.columbia.edu/CAVE/software/softlib/coil-100.php



(a) (b)

図2: 画像とその確率分布: (a) はあるクラスの画像, (b) はその画像の確率分布を表す.



(a) (b)

図3: 画像とその確率分布 (a) はあるクラスの画像, (b) は背景の画素 24 の確率を 0 とした時の画像の確率分布を表す.

5.2. 実験結果と考察

- スパース化割合の設定方法を γ_1, γ_2, γ_3 とした時の結果を表 1, 表 2, 表 3 に示す. この結果をグラフに表したものを図 4, 図 5, 図 6 にそれぞれ示す.

表 1: スパース化割合 γ_1

λ	Purity	NMI	Accuracy	Time
0.0001	0.6867	0.7755	0.6467	220.3664
0.0002	0.6467	0.7519	0.6200	174.4796
0.0003	0.6800	0.7670	0.6533	176.0322
0.0004	0.6467	0.7519	0.6200	123.9319
fix	0.6933	0.7780	0.6600	197.0669

表 2: スパース化割合 γ_2

λ	Purity	NMI	Accuracy	Time
0.0002	0.6733	0.7729	0.6600	135.1927
0.0004	0.6067	0.7326	0.5800	309.5305
0.0006	0.6467	0.7383	0.6267	779.0588
fix	0.6400	0.7499	0.6200	235.3125

表 3: スパース化割合 γ_3

λ	Purity	NMI	Accuracy	Time
0.00006	0.6733	0.7591	0.6533	124.7970
0.00012	0.6733	0.7606	0.6600	167.7670
0.00018	0.6800	0.7601	0.6533	233.5376
0.00024	0.6467	0.7523	0.6333	414.0343
fix	0.6667	0.7561	0.6400	354.9794

- スパース化割合を γ_3 とした時が Purity, NMI, Accuracy ともにスパース化割合を 0.7 と固定した時よりも良くなる傾向があることがわかった. 実行時間についてもスパース化割合を γ_3 とした時の方がスパース化割合を 0.7 で固定したものより速くなった. 重心の分散に基づいてスパース化割合を決定したことによって, 適切なスパース化割合に決定することができた可能性があると考えられる.

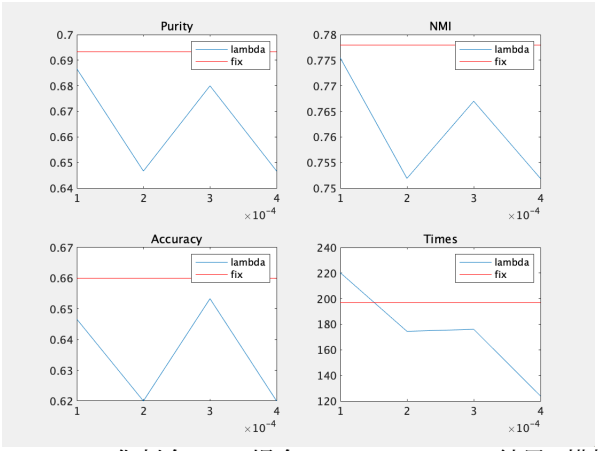


図4: スパース化割合 γ_1 の場合のパフォーマンス結果: 横軸はパラメータ λ の値を表す

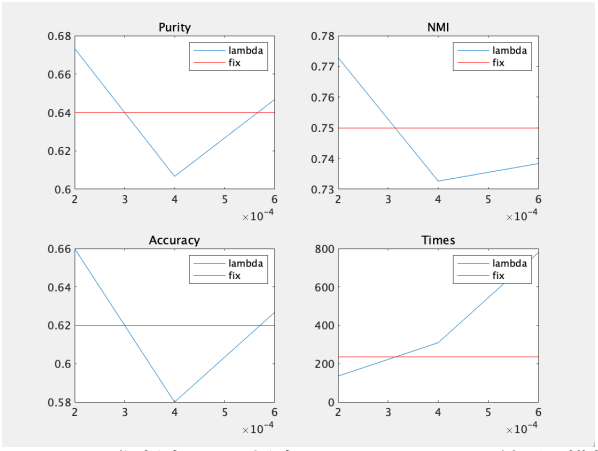


図5: スパース化割合 γ_2 の場合のパフォーマンス結果: 横軸はパラメータ λ の値を表す.

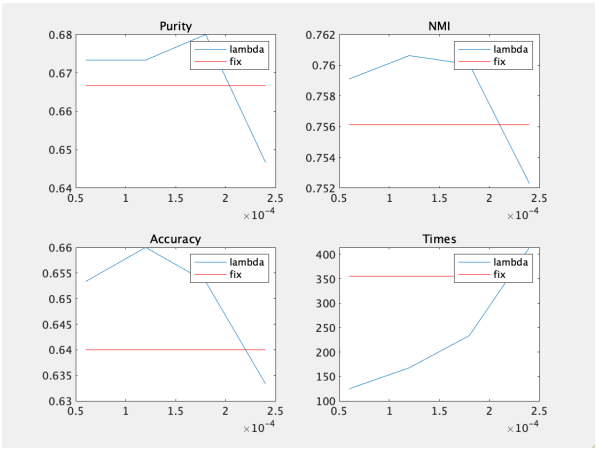


図6: スパース化割合 γ_3 の場合のパフォーマンス結果: 横軸はパラメータ λ の値を表す.

6. まとめ

- 交差エントロピー正則化を考慮した最尤推定に基づく正規分布への近似は元の確率分布の構造を保持するような変換ができなかった. 原因としては単峰性の正規分布を用いた近似をしようとしたことが考えられる. 混合正規

分布を用いたより良い近似手法についても考える．

- スパース化割合をセントロイドの分散に基づいて決定するとクラスタリング性能が向上する場合がある．特にスパース化割合 γ_3 の場合にパフォーマンスが良くなった．ただし, あらゆる λ について実験する必要がある．
- 分散は確率分布の特徴を表す代表的な指標であるが, 分布の構造を捉える上で分散だけでは十分ではないと考えられる．他に良い指標がないか考える．
- セントロイドの分散に基づいてスパース化割合を決定したが, 射影して縮退するとき, 重心とデータ点のサイズが同じになるように縮退する．この時, セントロイドの構造については考慮したが, データ点の構造については考慮されていない．セントロイドの構造とデータ点の構造を同時に考慮できる手法を今後考える．

参考文献

[1] T. Fukunaga and H Kasai. Wasserstein k-means with sparse simplex projection. *ICPR*, 2020.

[2] S. Lloyd. Least squares quantization in pcm. *IEEE*, 1982.