代表点を用いた大規模クラスタリングの近似法とその性質

寺田吉壱 [1,3], 山本 倫生 [2,3]
[1] 大阪大学大学院基礎工学研究科, [2] 大阪大学大学院人間科学研究科
[3] 理化学研究所革新知能統合研究センター

1. はじめに

近年，データの大規模化と複雑化が進み, データから仮説や有益な情報を獲得することが課題となっており, 探索的なデータ解析の中でも教師なし学習の重要性が再認識されている．クラスタリング法は, データの背後のクラスタ構造を明らかにするための教師なし学習の方法であり, 様々な分野で広く応用されている．最も代表的なクラスタリング法として, k-means 法が挙げられる．k-means 法はその簡便性と計算コストの低さから多用されるが, その単純さ故にデータの背後にある複雑なクラスタ構造を十分に捉えられない可能性がある．そのため, spectral clustering (von Luxburg, 2007) など, より柔軟にクラスタ構造を捉えられる方法の利用が望ましい．しかし, これらの複雑な方法は, 一般に計算コストが高く, 計算コストの削減が大きな課題となっている．

カーネル法におけるクラスタリング法に対しては, カーネル法に特化した Nyström 近似や Random Fourier Feature などの計算コスト緩和法が適用出来る．一方で, Yan et al. (2009) では, spectral clustering の k-means 法に基づく近似法 (KASP) が提案されている．KASP では, クラスタ数を多く設定した k-means 法を大規模データに適用し, 得られたクラスタ中心をデータを代表する点とする．そして, 得られた代表点のみに spectral clustering など複雑なクラスタリング法を適用し, 代表点に対するラベルをその代表点に近いデータ点のラベルとする．KASP は, spectral clustering に限らず任意のクラスタリング法に対して適用することができる．この方法の大きな利点は, 安定性, 簡便性, 計算コストの低さ, 汎用性である．また, KASP のアルゴリズムとは本質的に異なるが, 代表点として subsample を用いる spectral clustering の近似法が提案されている (Mohan and Monteleoni, 2017)．しかし, Terada and Yamamoto (2019) の理論から, この方法は subsample に対する normalized cut と等価であり, 汎用的な近似方法ではない．

本発表では, KASP の問題点を明らかにし, その問題点を解決した汎用的な大規模クラスタリングの近似法を提案する．

2. KASP の問題点と提案手法

KASP では, データの代表点として, k-means 法のクラスタ中心を用いる．しかし, k-means 法によって生成した代表点から構成される経験分布は, 母集団分布を代表する点とはならない．具体的には, 代表点の経験分布は, 母集団分布よりも裾が重い分布に収束することが示せる．そのため, KASP を用いて近似を行うと, グラスタリング結果にズレが生じてしまう．

この問題点の最もシンプルな解決法は, 各代表点に適切な重みを与えることである．K-means 法に対応するベクトル量子化は, 母集団分布との L_2-Wasserstein 距離を最小にするような $\#(\mathrm{supp}(Q)) \leq K$ を満たす離散測度を求める問題に対応している．このことから, 代表点に適切な重みを与えることで, KASP の問題点を解消することができる．一方で, 代表点の経験分布のズレから, 母集団分布において密度の低い点も代表点として生成される．それらの代表点には, 低い重みが割り振られるため, 効率が悪い．

この問題を解決するために, 新しい代表点の生成方法である Density-Preserving Vector Quantization (DPVQ) を提案する．DPVQ は重み付き k-means 法の一種であり, 容易に代表点を生成できる．また, DPVQ が生成する代表点の経験分布は, 漸近的にデータの背後の分布へ収束すること

Algorithm 1 $\mathrm{VQ}_n(\mu \,|\, r, K)$ の最適化アルゴリズム

1: $t \leftarrow 0$ とし, クラスタ中心 $\mu_1^{(0)}, \ldots, \mu_K^{(0)}$ を初期化する.

2: **for** $t = 0, \ldots, T$ **do**

3: 各 i $(i = 1, \ldots, n)$ に対して, $\left\|x_i - \mu_k^{(t)}\right\|$ を最小にするクラスタ k を割り当て, 帰属行列 $U^{(t)} = \left(u_{ij}^{(t)}\right)_{n \times K}$ を得る.

$$u_{ik}^{(t)} = \begin{cases} 1 & \text{if } \forall j;\ \left\|x_i - \mu_k^{(t)}\right\| \leq \left\|x_i - \mu_j^{(t)}\right\|, \\ 0 & \text{otherwise.} \end{cases}$$

4: クラスタ平均を以下で更新する.

$$\hat{\mu}_k^{(t+1)} = \frac{1}{\sum_{j=1}^n u_{jk}^{(t)} w_{jk}^{(t)}} \sum_{i=1}^n u_{ik}^{(t)} w_{ik}^{(t)} x_i, \quad w_{ik}^{(t)} = \begin{cases} \left\|x_i - \hat{\mu}_k^{(t)}\right\|^{r-2} & \text{if } \left\|x_i - \hat{\mu}_k^{(t)}\right\| > 0, \\ \delta & \text{if } \left\|x_i - \hat{\mu}_k^{(t)}\right\| = 0. \end{cases}$$

ここで, $\delta > 0$ は小さい正の定数である.

5: 収束判定条件を満たせば停止し, 満たさなければ $t \leftarrow t + 1$ とする.

6: **end for**

が示せる. そのため, DPVQ による代表点には平等な重みが割り振られるため, 効率的な近似が期待できる. 一方で, DPVQ は, 密度推定を必要とするため, 高次元データに対しては不安定となる.

そこで, 本発表では, 以下で定義される order r のベクトル量子化器を用いた近似法も提案する.

$$\mathrm{VQ}_n(\mu \,|\, r, K) := \frac{1}{n} \sum_{i=1}^n \min_{1 \leq k \leq K} \|x_i - \mu_k\|^r$$

ここで, $r \in (0, 2]$ は定数, $\|\cdot\|$ は $\mathbb{R}^d$ 上のノルム, $x_1, \ldots, x_n \in \mathbb{R}^d$ は各データ点, $\mu_k \in \mathbb{R}^d$ は k 番目のクラスタ中心, $\mu = (\mu_1, \ldots, \mu_K)$ である. 本発表では, 計算の簡便性のために, $\|\cdot\|$ は Euclid ノルムとする. $r = 2$ とすれば $\mathrm{VQ}_n(\mu \,|\, r, K)$ は k-means 法と一致するが, r を小さくとることで代表点の経験分布と母集団分布のズレを小さくすることができる. 本発表では, $\mathrm{VQ}_n(\mu \,|\, r, K)$ に対する最適化問題を高速に解くために, k-means like なアルゴリズムを提案する. Algorithm 1 において, $\mathrm{VQ}_n(\mu^{(t)} \,|\, r, K) > \mathrm{VQ}_n(\mu^{(t+1)} \,|\, r, K)$ という単調減少性が成り立つため, 停留点への収束性が保証できる. 提案手法を用いた場合の spectral clustering の近似方法, 詳細な理論的性質, DPVQ や $r < 2$ としたベクトル量子化器を用いた提案手法と KASP や subsample を用いた既存の近似手法の数値実験による比較は当日報告する.

参考文献

Mohan, M. and Monteleoni, C. (2017). Beyond the Nystrom Approximation: Speeding up Spectral Clustering using Uniform Sampling and Weighted Kernel k-means. *In Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2494–2500.

Terada, Y. and Yamamoto, M. (2019). Kernel Normalized Cut: a Theoretical Revisit. *In Proceedings of the 36th International Conference on Machine Learning*, PMLR **97**, 6206–6214.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, **17**, 395–416.

Yan, D., Huang, L., and Jordan, M. I. (2009) Fast approximate spectral clustering. *In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 907–916.