Uniform and Non-uniform Sampling Methods for Sub-linear Time k-means Clustering

Yuanhang Ren[1]
ryuanhang@gmail.com

Ye Du[2]
henry.duye@gmail.com

[1]University of Electronic Science and Technology of China
[2]Southwestern University of Finance and Economics, China

December 18, 2020

Table of Contents

k-means problem

- It is well-known due to the Lloyd algorithm [Lloyd, 1982] (a.k.a k-means algorithm)
- The k-means problem is NP-hard

Definition (k-means problem)

Given n data points $\mathcal{X} \subseteq \mathbb{R}^d$ and a set of k points $C \subseteq \mathbb{R}^d$, where d is the dimension of the data point. An objective function is defined as follows,

$$\phi_C(\mathcal{X}) = \sum_{x \in \mathcal{X}} d^2(x, C) \tag{1}$$

where $d(x, C) = \min_{c \in C} \|x - c\|$ is the distance of a point to a set. The k-means problem is to find the optimal C such that the $\phi_C(\mathcal{X})$ is minimized given $\mathcal{X}$.

The Solution Quality

Definition (Solution Quality 1)

Let $\alpha \geq 1$. A set C of k centers is an α approximation solution of k-means if

$$\phi_C(\mathcal{X}) \leq \alpha \phi_{\mathsf{OPT}}(\mathcal{X}) \tag{2}$$

$\phi_{\mathsf{OPT}}(\mathcal{X})$ is the minimal objective.

Definition (Solution Quality 2)

Let $\alpha \geq 1$ and $\beta > 0$. A set C of k centers is a β-bad α-approximation solution of k-means if

$$\phi_C(\mathcal{X}) > (\alpha + \beta)\phi_{\mathsf{OPT}}(\mathcal{X}) \tag{3}$$

Otherwise, C is said to be a β-good α-approximation.

Lloyd Algorithm

1. A set of k centers are initialized using uniform random sampling.
2. Each point is assigned to its nearest center, which forms k clusters.
3. The mean point of each cluster is computed, which is used as the new center of the cluster.
4. Repeat the step 2 and 3 multiple times.

However,

- First, there is no theoretical guarantee for the solution quality.
- Second, if the number of points is very large, it could be infeasible to run this algorithm.

Table of Contents

Problems to Investigate

Can we design an efficient algorithm (sublinear time) and the clustering quality is also theoretically guaranteed (constant approximation ratio)?

Table of Contents

Methods Overview

Overall, we use uniform sampling to sample a set of points and we run a quality guaranteed algorithm on this subset to achieve the goal.

Clustering based on Uniform Sampling

Algorithm 1: Clustering based on uniform sampling

Input: dataset $\mathcal{X}$, number of clusters k, number of points to sample s, clustering algorithm $\mathcal{A}_c$

Output: k centers C

$S \leftarrow$ Sample s points uniformly without replacement

$C \leftarrow$ Solve the k-means problem on S with $\mathcal{A}_c$

return k centers C

Theorem (quality of Algorithm 1)

Let $0 < \delta < 1/2$, $\alpha \geq 1$, $\beta > 0$ be approximation parameters. Let C be the set of centers returned by Algorithm 1 and $\mathcal{A}_c$ is an α approximation algorithm. Suppose we sample s points uniformly without replacement such that,

$$s \geq \ln(\frac{1}{\delta})(1 + \frac{1}{n})/(\frac{\beta^2 m^2}{2\Delta^2\alpha^2} + \frac{\ln(1/\delta)}{n})$$

we have

$$\phi_C(\mathcal{X}) \leq 4(\alpha + \beta)\phi_{OPT}(\mathcal{X})$$

with probability at least $1 - 2\delta$, where $\Delta = \max_{i,j} \|v_i - v_j\|^2$ is the squared diameter of the data, $m = \phi_{OPT}(\mathcal{X})/n$ is the average of the optimal objective.

Our Contribution

- A sharper bound for the uniform sampling algorithm is proved.
- A further proof indicates that this algorithm runs in poly-logarithmic time given mild assumptions on datasets.

A Sharper Bound

Theorem (a sharper bound of uniform sampling)

Let $0 < \delta < 1/2$, $\alpha \geq 1$, $\beta > 0$ be approximation parameters. Let C be the set of centers returned by Algorithm 1 and $\mathcal{A}_c$ is an α approximation algorithm. Suppose we sample s points uniformly without replacement such that,

$$s \geq \ln(\frac{1}{\delta})(1 + \frac{1}{n})/(\frac{\beta^2 m^2}{2\Delta^2\alpha^2} + \frac{\ln(1/\delta)}{n})$$

we have

$$\phi_C(\mathcal{X}) \leq (\alpha + \beta)\phi_{OPT}(\mathcal{X})$$

with probability at least $1 - 2\delta$, where $\Delta = \max_{i,j} \|v_i - v_j\|^2$ is the squared diameter of the data, $m = \phi_{OPT}(\mathcal{X})/n$ is the average of the optimal objective.

The big picture of the proof:

1. Show that the sample set S will be a good representative of $\mathcal{X}$.
2. Suppose C is a *bad* solution for $\mathcal{X}$, then the sample set S will be a good representative of $\mathcal{X}$ with a **low** probability.
3. According to 1 and 2, C will be a *good* solution for $\mathcal{X}$.

A Poly-Log Time Algorithm

Assume that a dataset is sampled i.i.d. according to a probability distribution F

- F has finite variance and exponential tails, *i.e.* $\exists c, t$ such that $P[\mathrm{d}(x, \mu(F)) > a] \leq ce^{-at}$, where $\mu(F)$ is the mean of F.
- F's minimal and maximal density on a hypersphere with non zero probability mass is bounded by a constant.

Theorem (efficiency of uniform sampling)

Let $0 < \delta < 1/2$, $\alpha \geq 1$, $\beta > 0$ *be approximation parameters. Assume above hold, and let* C *be the set of centers returned by Algorithm 1, we have the following*

$$\phi_C(\mathcal{X}) \leq (\alpha + \beta)\phi_{OPT}(\mathcal{X})$$

with probability at least $1 - 2\delta$ *if we sample* $O(\ln(\frac{1}{\delta})\frac{\alpha^2}{\beta^2}k^2 \log^4 n)$ *points*

Baseline Algorithms

- Since the uniform sampling algorithm is efficient and provably good, we design experiments to verify this.
- Baselines are K-MC^2 [Bachem et al., 2016] and Double-K-MC^2 sampling.
- In previous works [Bahmani et al., 2012], the sampled points are weighted to obtain a better quality. Hence, we use the K-MC^2 method to sample a set of points as weights and the method is called the Double-K-MC^2.

Algorithm 2: Double-K-MC^2 sampling

Input: dataset $\mathcal{X}$, # of points to sample s, chain length u
Output: k centers C
$S_1 \leftarrow$ Sample s points from V via K-MC^2
$V' \leftarrow$ Remove S_1 from V
$S_2 \leftarrow$ Sample s points from V' via K-MC^2
For point $s_i \in S_1$, let w_i be the number of points in S_2 closer to s_i than to any other points in S_1
Let $w_i + 1$ be the weight of s_i
$C \leftarrow$ Solve the weighted k-means problem on S_1 with an α approximation algorithm
return k centers C

Traditional Clustering

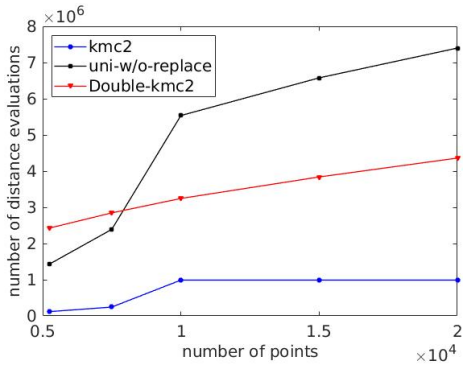
Table 1: data size n, number of clusters k, dimension d

datasets	n	k	d
a2	5250	35	2
a3	7500	50	2
b2-random-10	10000	100	2
b2-random-15	15000	100	2
b2-random-20	20000	100	2
KDD	145751	200	74
RNA	488565	200	8
Poker Hand	1000000	200	10

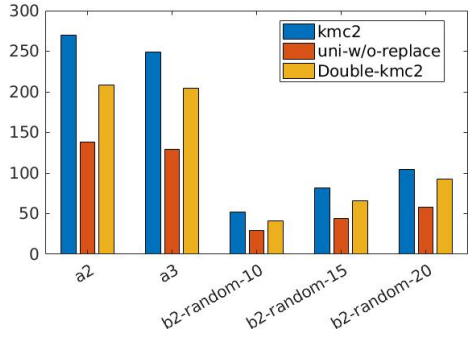
- chain length: $u = 200$
- sampling size: $1.5 \log^2 n$ and $0.7 \log^4 n$ for Double-K-MC^2 and uniform sampling
- α approximation algorithm: (weighted) k-means++ with Lloyd
- evaluation metrics: number of distance evaluations and k-means objective
- algorithms are run 40 times repeatedly with different initial random seeds

Results

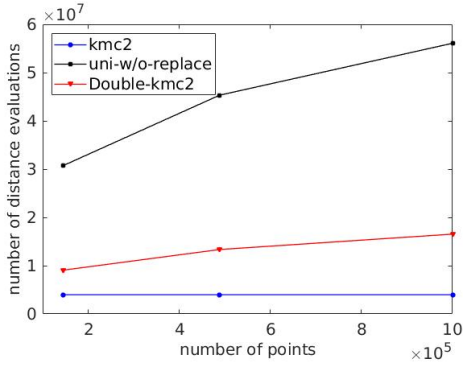
1. The time cost of uniform sampling is about 10 times higher than that of K-MC^2 and it increases slowly with respect to the data size. The k-means objective of uniform sampling is roughly 60% of the objective of K-MC^2.
2. Double-K-MC^2 achieves a better clustering quality compared with K-MC^2 and a lower time cost compared with uniform sampling.
3. Double-K-MC^2 could be the first choice if you prefer a good clustering quality with reasonable time costs. For the best quality, uniform sampling is recommended.



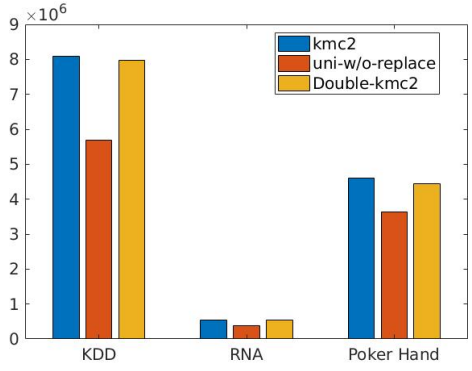
(A) the number of distance evaluations on synthetic data



(B) k-means objective on synthetic data



(C) the number of distance evaluations on real data



(D) k-means objective on real data

FIGURE 1: k-means objective and time cost versus the number of points

Image Segmentation

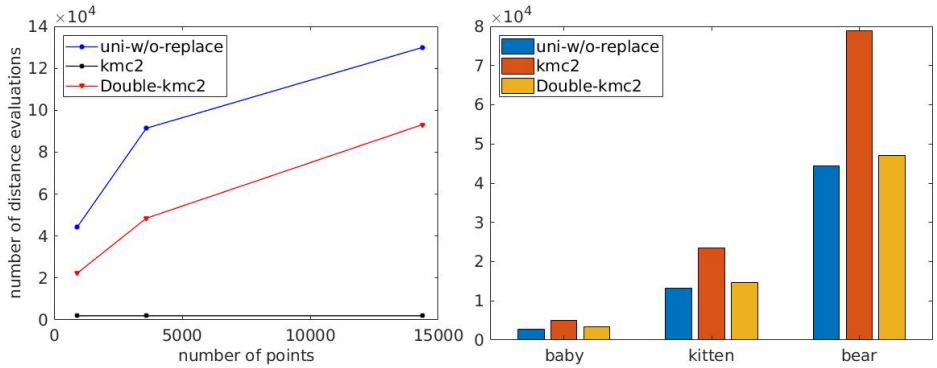
Table 2: data size n, number of clusters k

datasets	n	k
baby	900(30 * 30)	5
kitten	3600(60 * 60)	5
bear	14400(120 * 120)	5

- The kernel versions of uniform sampling, Double-K-MC^2, and K-MC^2.
- Construct an affinity matrix A via the approach in Stella and Shi [2003] and find the nearest positive definite matrix K as the kernel.
- chain length: $u = 200$
- sampling size: $0.25 \log^2 n$ and $0.4 \log^4 n$ for Double-K-MC^2 and uniform sampling
- α approximation algorithm: (weighted) kernel k-means++ with kernel Lloyd
- evaluation metric: number of distance evaluations and kernel k-means objective
- algorithms are run 30 times repeatedly with different initial random seeds

Results

1. The kernel uniform sampling has the best clustering quality while the growth of the time cost is not too rapid.
2. The kernel Double-K-MC^2 has a similar clustering quality with much lower time cost compared with the kernel uniform sampling.
3. Thus, we recommend using kernel Double-K-MC^2 if the quality is your major concern. For a more efficient result, the kernel K-MC^2 is a better choice.



(A) the number of distance evaluations on image data (B) kernel k-means objective on image data

Figure 2: kernel k-means objective and time cost versus the number of points

Table of Contents

Summary

- We improved the analysis of uniform sampling based k-means clustering algorithm by two folds. First, a sharper bound of solution quality is derived. Second, the algorithm runs in poly-log time given mild assumptions of datasets. We then proposed Double-K-MC^2 sampling to weigh sample points.
- Experiments demonstrate that the uniform sampling based algorithm achieves a much better clustering quality while not spend too much time. The Double-K-MC^2 almost runs as efficient as K-MC^2 and the solution quality is slightly better.
- Codes and Datasets:
  `https://github.com/ryh95/uniform-double-kmc2-sampling`

Questions?

References

O. Bachem, M. Lucic, S. H. Hassani, and A. Krause. Approximate k-means++ in sublinear time. 2016.

B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii. Scalable k-means++. *Proceedings of the VLDB Endowment*, 5(7):622–633, 2012.

S. Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

X. Y. Stella and J. Shi. Multiclass spectral clustering. In *null*, page 313. IEEE, 2003.