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

Yuanhang Ren[1] and Ye Du[2]

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

{ryuanhang,henry.duye}@gmail.com

Introduction

Among all clustering problems, the k-means problem is probably the most well-known one. Lloyd algorithm [1] is a simple and elegant algorithm that gives a certain local optimum for this problem. It works as follows. First, a set of k centers are initialized using uniform random sampling. Then, each point is assigned to its nearest center, which forms k clusters. Finally, the mean point of each cluster is computed, which is used as the new center of the cluster. In practice, the second and third steps can be repeated for t times. However, this algorithm has two severe drawbacks. 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.

Research Problems

Can we design an algorithm that is efficient and the clustering quality is theoretically guaranteed for the k-means problem?

Backgrounds

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,*

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

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 $\mathcal{X}$.

Definition (Solution Quality 1)**.** *Let $\alpha \geq 1$. A set C of k centers is an α approximation solution of k-means if*

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

$\phi_{OPT}(\mathcal{X})$ 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*

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

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

Uniform Sampling

```
Algorithm 1: Clustering based on uniform sampling [2]
Input: Dataset $\mathcal{X}$, # of clusters $k$, # 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 $C$
```

Theoretical Results

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

Assume that a dataset is sampled i.i.d. according to a probability distribution F [3].

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

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

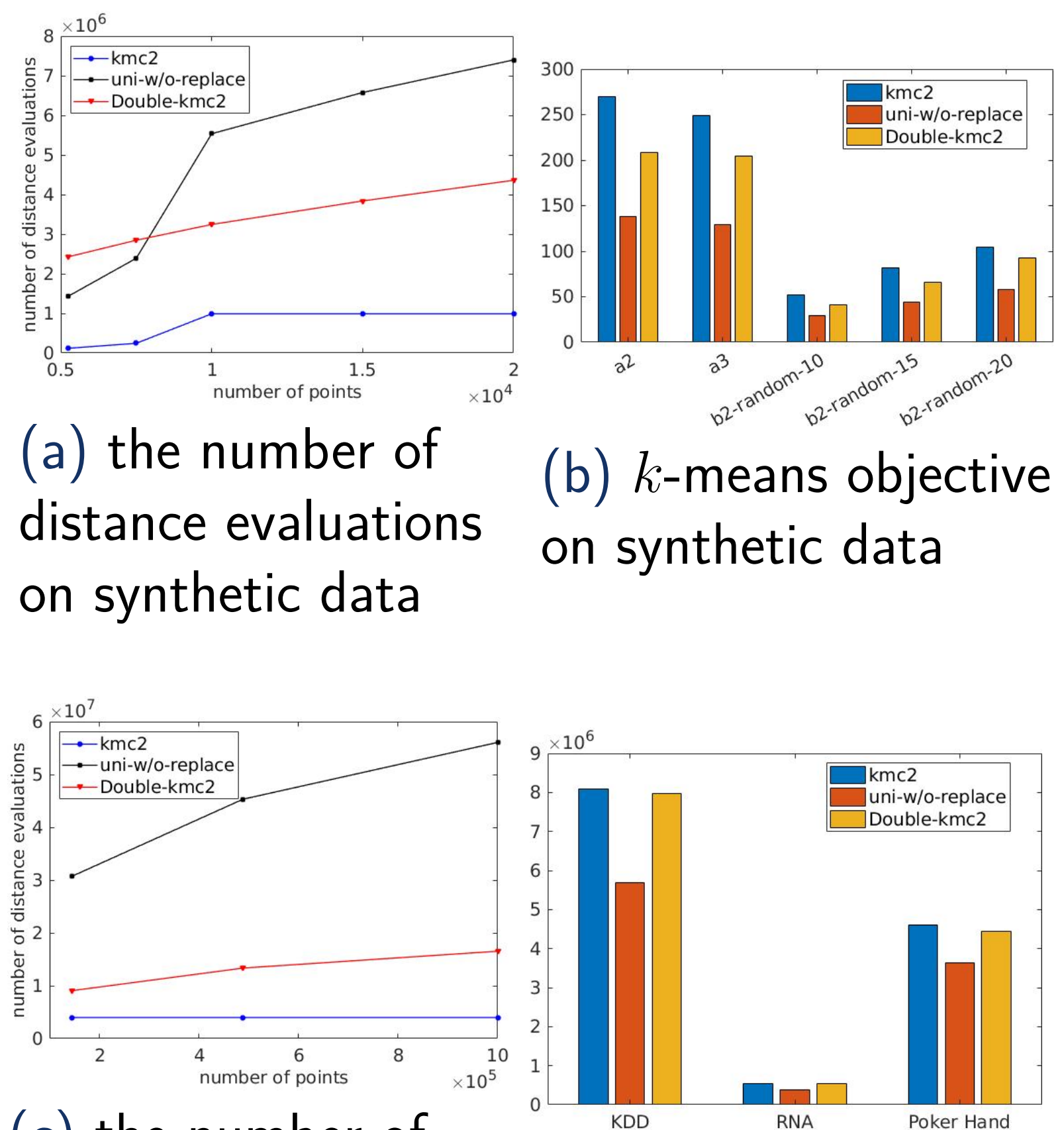
with probability at least $1 - 2\delta$ if we sample $O(\ln(\frac{1}{\delta})\frac{\alpha^2}{\beta^2}k^2\log^4 n)$ points

Experimental Results

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

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

Table: Clustering Datesets

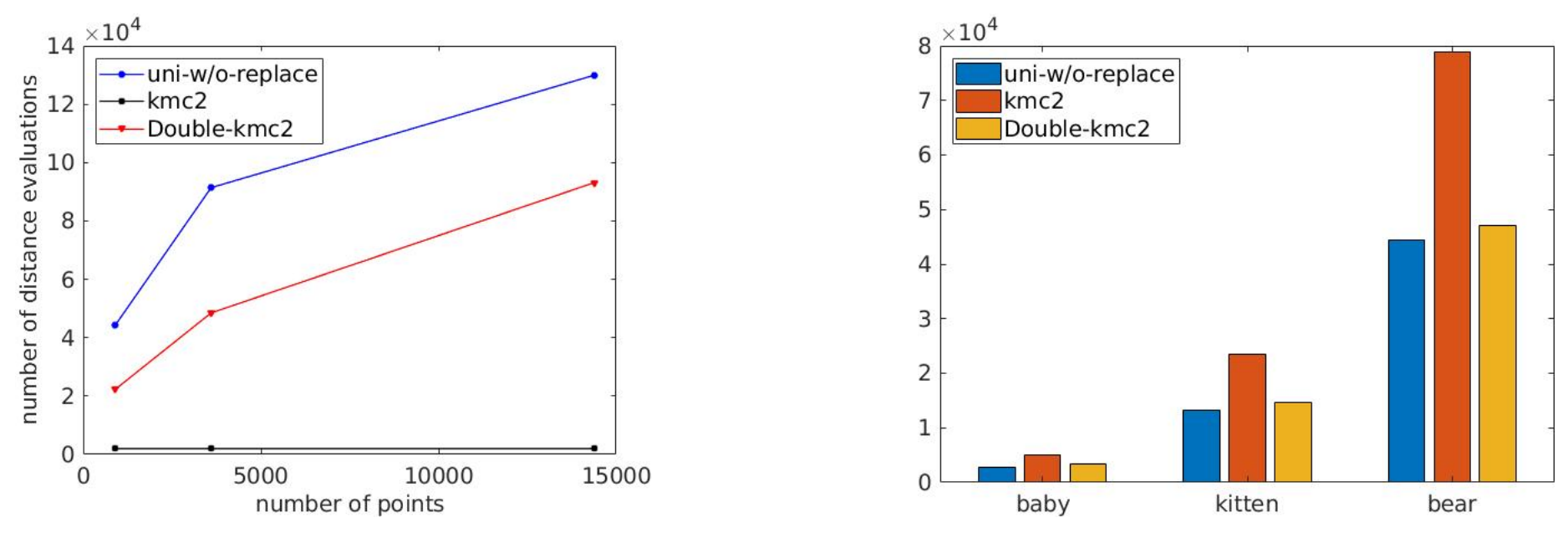


(a) the number of distance evaluations on synthetic data

(b) k-means objective on synthetic data

(c) the number of distance evaluations on real data

(d) k-means objective on real data



(a) the number of distance evaluations on image data

(b) kernel k-means objective on image data

Non-uniform Sampling

```
Algorithm 1: 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 $\alpha$ approximation algorithm
return $k$ centers $C$
```

Conclusions

- We improve 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 propose 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-MC2 almost runs as efficiently as K-MC2 and the solution quality is slightly better.

References

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

[2] Mahesh Mohan and Claire Monteleoni. 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*, IJCAI'17, pages 2494–2500. AAAI Press, 2017.

[3] Olivier Bachem, Mario Lucic, S Hamed Hassani, and Andreas Krause. Approximate k-means++ in sublinear time. 2016.

Codes and Datasets

https://github.com/ryh95/uniform-double-kmc2-sampling