Linear-Time Algorithms for k-means with Multi-Swap Local Search

Junyu Huang Qilong Feng Ziyun Huang Jinhui Xu Jianxin Wang

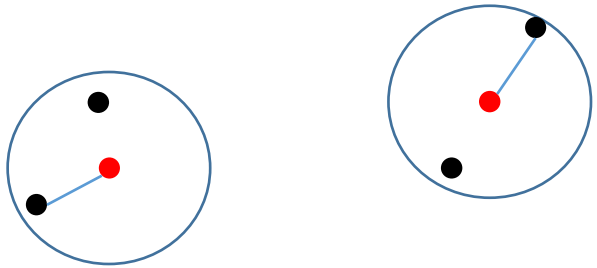
Central South University

NeurIPS 2023

k-means clustering

Input: a dataset $P \subseteq R^d$ and a parameter k

Output: a set $C \subseteq R^d$ of at most k points with minimum clustering cost $\sum_{p \in P} d(p, C)^2$



An instance for k-means for $n = 6,\ k = 2$

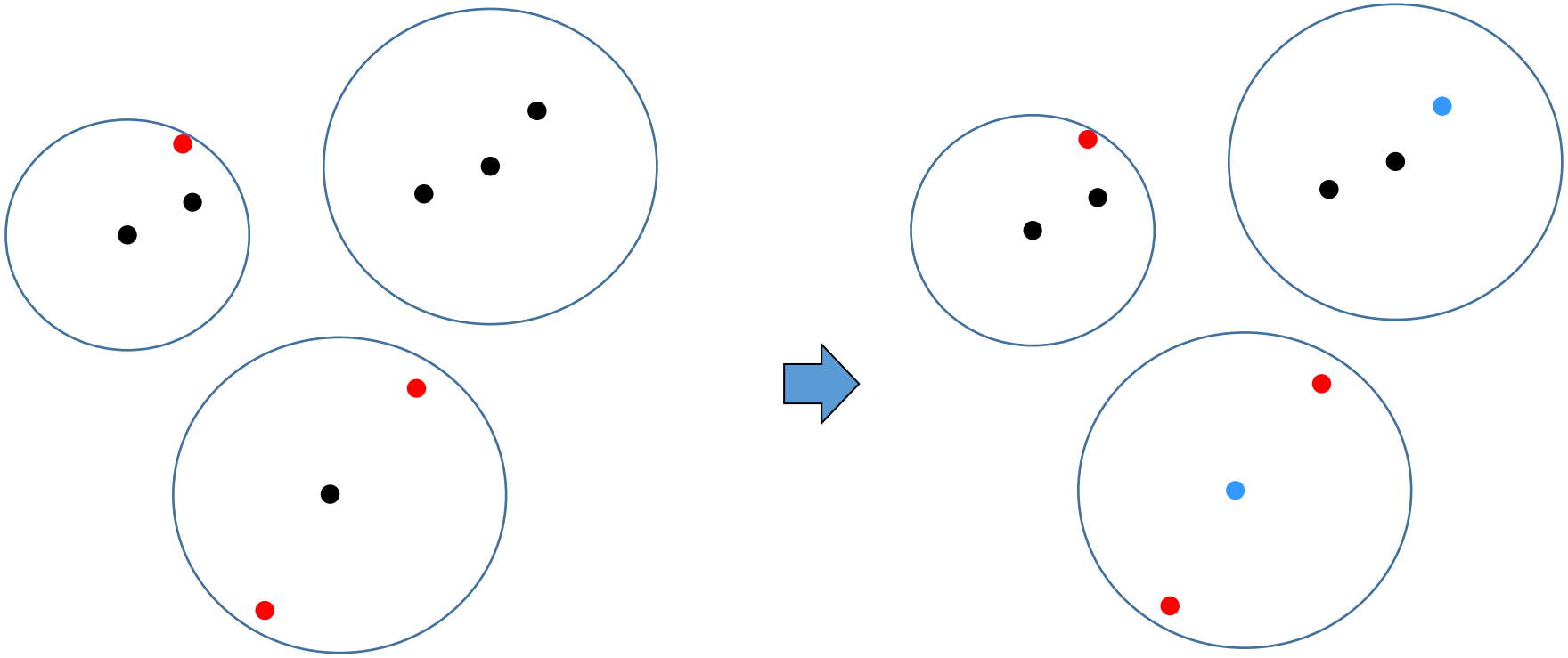
Related works

Method	Approximation	Assumption	Time	Reference
k-means++	$O(logk)$	-	$O(ndk)$	SODA 2007
Multi-Swap Local Search	$(3+2/t)^2$	-	$O(n^{t+1}dk^t log\Delta)$	SOCG 2004
Sampling + Single-Swap Local Search	509	-	$O(ndk^2 loglogk)$	ICML 2019
Sampling + Single-Swap Local Search	$O(1)$	-	$O(ndklogk)$	ICML 2020
Sampling + Single-Swap Local Search	$100+\epsilon$	$\lvert P_h^* \rvert \geq \frac{n\epsilon}{k}$	$O(ndk^2 log\epsilon^{-1})$	IJCAI 2022
Ours (Sampling + Multi-Swap Local Search)	$50(1+1/t)+\epsilon$	-	$O(ndk^{2t+1}\log(\epsilon^{-1}logk))$	NeurIPS 2023

t-Swap Local Search in Linear Time

1) Initialize a set C of k centers using k-means++
2) For $i = 1$ to $O(k^{O(t)})$
3) Sample a set S of data points from P with $|S| = t$, where each $s \in S$ is sampled with probability $\phi(s, C)/\phi(P, C)$
4) If $\exists\, U \subseteq C$ and $V \subseteq S$ s.t. $\phi(P, C\backslash U \cup V) < \phi(P, C)$
5) Do $C = C\backslash U \cup V$

- An instance of 3-means with 2-swap

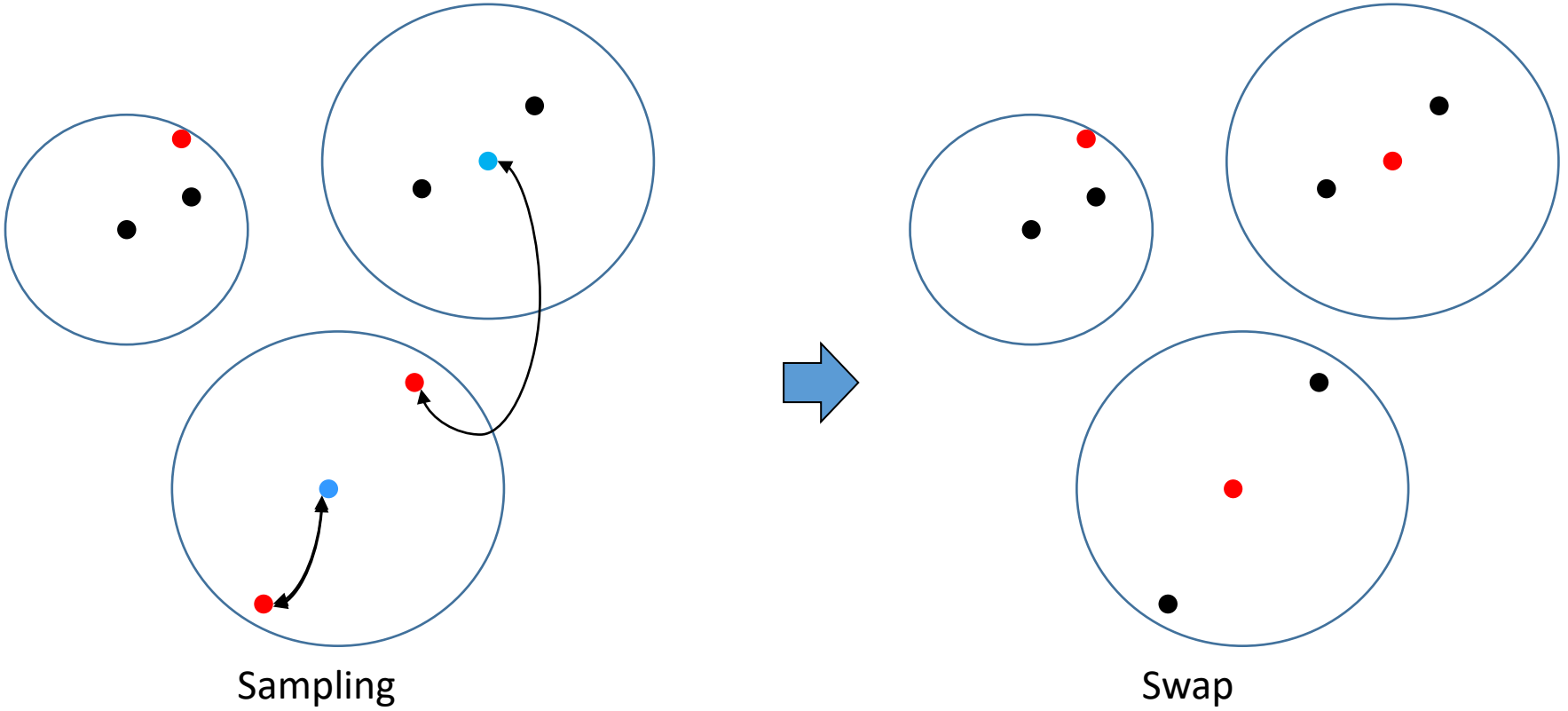


Initial Solution

Sampling

The sampling process guarantees that data points close to a subset of optimal clustering centers can be found

- An instance of 3-means with 2-swap



Large-Scale Clustering Experiments

Datasets: Large-scale Clustering datasets including SUSY (5 million)、 HIGGS (10 million) and SIFT (100 million)

Criteria: Fixed 400 iterations of Swap for each local search algorithm

Dataset Description

dataset	Size
SUSY	5,000,000 * 18
HIGGS	11,000,000 * 28
SIFT	100,000,000 * 128

- Large-Scale Clustering Experiments

Hardware: 72 Intel Xeon Gold 6230 CPUs with 500GB memory

Algorithms: Single-Swap Local Search (LS++, from ICML 2019)、 Sampling-Based Local Search(FLS, from IJCAI 2022)、 Multi-Swap Local Search using acceleration heuristics(MLS)

- Large-Scale Clustering Experiments

Results:

Method	dataset	Size	Min	Mean+ std	Time (s)
LS++	SUSY	5,000,000 * 18	3.2738E+07	3.2875E+07 ± 1.1E+05	827.71
FLS			3.1632E+07	3.1672E+07 ± 2.8E+05	9287.26
MLSP			**3.1575E+07**	**3.1633E+07 ± 3.8E+04**	7462.57
MLS			3.2219E+07	3.2424E+07 ± 1.4E+05	**534.11**
LS++	HIGGS	11,000,000 * 28	1.8604E+08	1.8834E+08 ± 1.4E+06	2424.97
FLS			1.8938E+08	1.8964E+08 ± 1.5E+05	39826.29
MLSP			**1.8373E+08**	**1.8410E+08 ± 1.1E+05**	21928.97
MLS			1.8623E+08	1.8686E+08 ± 5.5E+05	**2037.68**
LS++	SIFT	100,000,000 * 128	1.5886E+13	1.5953E+13 ± 1.9E+11	130248
FLS			1.5824E+13	1.5902E+13 ± 1.8E+11	62628
MLS			**1.5802E+13**	**1.5898E+13 ± 1.1E+11**	**37081**

The clustering cost is reduced by 2.4% and 1.9% compared with LS++, FLS algorithms.

Thanks