Fast Distributed k-Center Clustering with Outliers on Massive Data

Gustavo Malkomes[1], Matt J. Kusner[1], Wenlin Chen[1], Benjamin Moseley[1], Kilian Q. Weinberger[2]

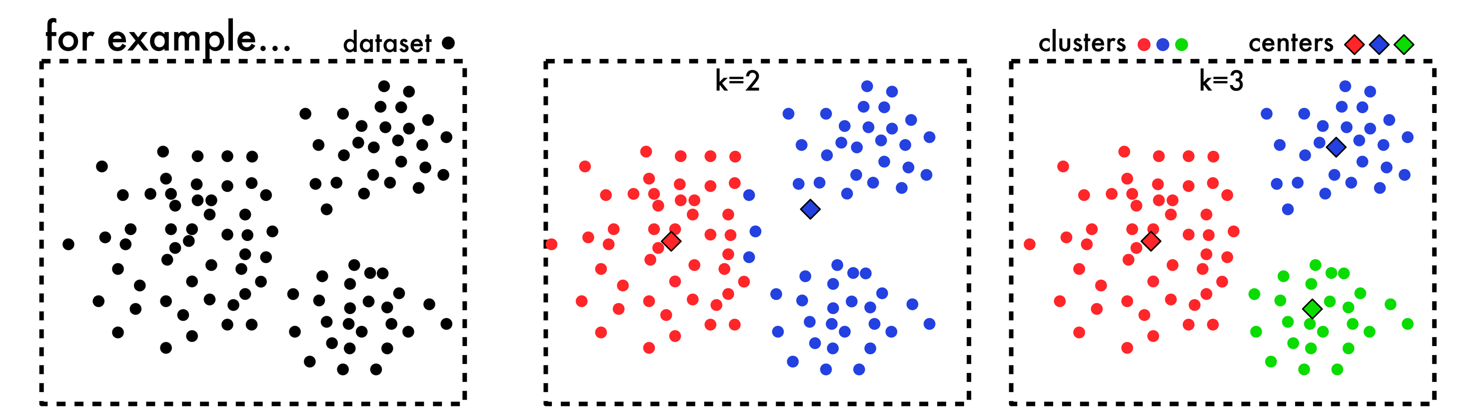
[1]Department of Computer Science & Engineering, Washington University in St. Louis, USA

[2]Department of Computer Science, Cornell University, USA

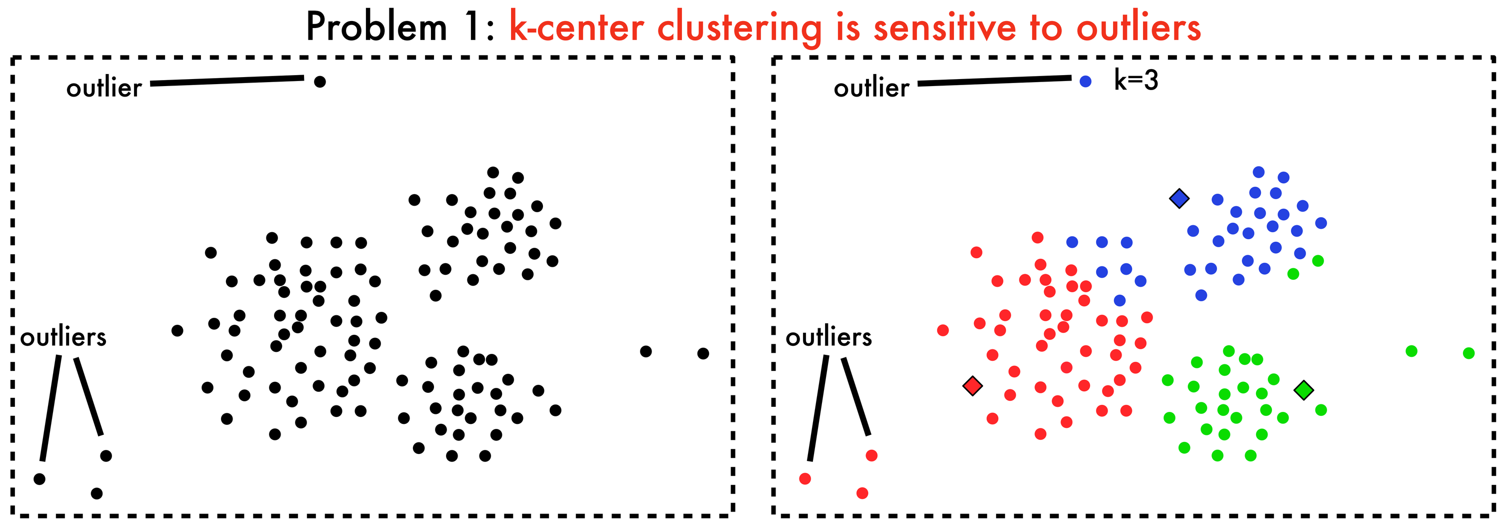
Background

k-center: instance-based clustering

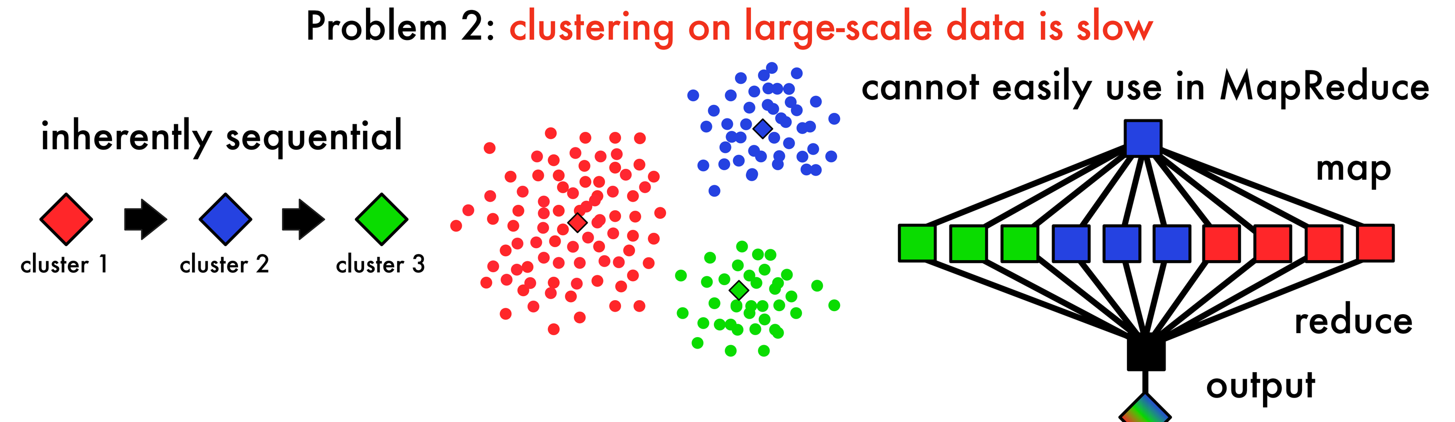
minimize distance from any input to its closest cluster center



Problem 1: k-center clustering is sensitive to outliers



Problem 2: clustering on large-scale data is slow



Prior Work

sequential k-center

notation

- U universe (dataset)
- n size of universe
- k number of clusters
- $d(\cdot,\cdot)$ distance function
- X cluster centers

objective

$$\underbrace{\max_{v \in U} \min_{u \in X} d(u,v)}_{d_X(v)}$$

algorithm

Algorithm 1 Sequential k-center

GREEDY(U,k)

1: $X = \emptyset$
2: Add any point $u \in U$ to X
3: **while** $|X| < k$ **do**
4: $\quad u = \texttt{argmax}_{v \in U}\, d_X(v)$
5: $\quad X = X \cup \{u\}$
6: **end while**

worst-case guarantee [Hochbaum & Shmoys, 1985]

Algorithm 1 is guaranteed to be at most 2 times worse than the optimal clustering

sequential k-center with outliers

notation

- Z outliers
- z size of outliers
- OPT optimal solution
- G guess of OPT
- $d_{\cdot,\cdot}$ distance

objective

$$\max_{v \in U \setminus Z} d_X(v)$$

algorithm

Algorithm 2 Sequential k-center with outliers

OUTLIERS(U,k,G)

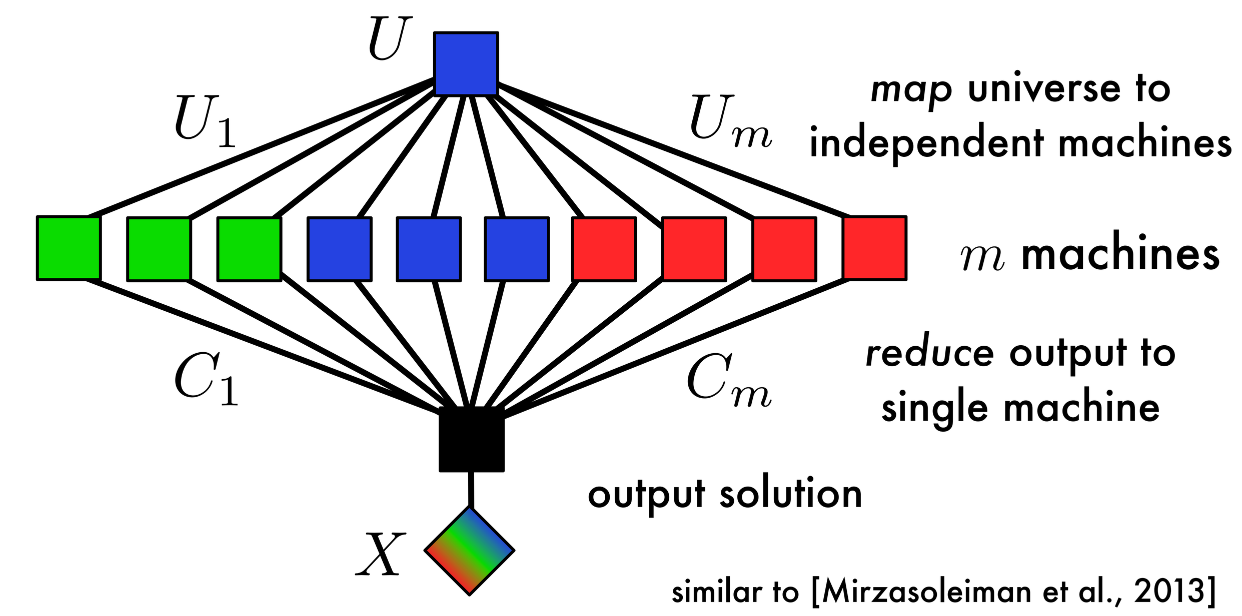
1: $U' = U,\ X = \emptyset$
2: **while** $|X| < k$ **do**
3: $\quad \forall u \in U'$ let $B_u = \{v \mid v \in U', d_{u,v} \le G\}$
4: $\quad$ Let $v' = \texttt{argmax}_{u \in U'} |B_u|$
5: $\quad$ Set $X = X \cup \{v'\}$
6: $\quad$ Compute $B'_{v'} = \{v \mid v \in U', d_{v',v} \le 3G\}$
7: $\quad U' = U' \setminus B'_{v'}$
8: **end while**

worst-case guarantee [Charikar et al., 2001]

Algorithm 2 is guaranteed to be at most 3 times worse than the optimal clustering

Fast Distributed Clustering

k-center



Algorithm 3 Distributed k-center

GREEDY-MR(U,k)

1: Partition U into m equal sized sets $U_1,\dots,U_m$ where machine i receives U_i.
2: Machine i assigns C_i = GREEDY(U_i,k)
3: All sets C_i are assigned to machine 1
4: Machine 1 sets X = GREEDY$(\cup_{i=1}^m C_i, k)$
5: Output X

communication/storage

Assuming data is already partitioned across machines the communication cost is O(km) and the memory usage on each machine is O(max{n/m,mk})

worst-case guarantee

Algorithm 3 is guaranteed to be at most 4 times worse than the optimal clustering

k-center with outliers

Algorithm 4 Distributed k-center with outliers

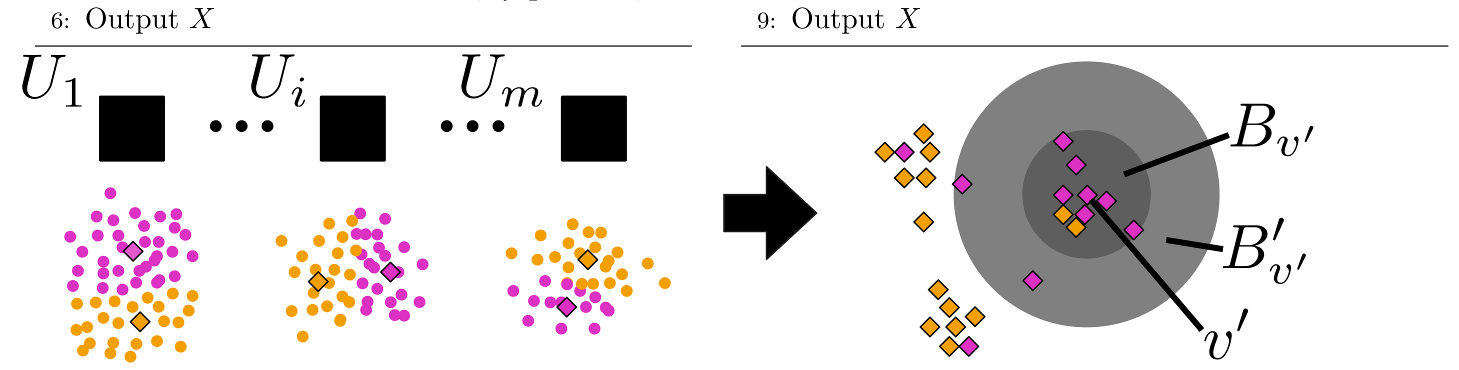
OUTLIERS-MR(U,k,z,G,α,β)

1: Partition U into m equal sized sets $U_1,\dots,U_m$ where machine i receives U_i.
2: Machines i sets C_i = GREEDY$(U_i, k+z)$
3: For each point $c \in C_i$, machine i set $w_c = |\{v \mid v \in U_i, d(v,c) = d_{C_i}(v)\}| + 1$
4: All sets C_i are assigned to machine 1 with the weights of the points in C_i
5: Machine 1 sets X = CLUSTER$(\cup_{i=1}^m C_i, k, G)$
6: Output X

Algorithm 5 Clustering subroutine

CLUSTER(U,k,G)

1: $U' = U,\ X = \emptyset$
2: **while** $|X| < k$ **do**
3: $\quad \forall u \in U$ compute $B_u = \{v \mid v \in U', d_{u,v} \le 5G\}$
4: $\quad$ Let $v' = \texttt{argmax}_{u \in U} \sum_{u' \in B_u} w_{u'}$
5: $\quad$ Set $X = X \cup \{v'\}$
6: $\quad$ Compute $B'_{v'} = \{v \mid v \in U', d_{v',v} \le 11G\}$
7: $\quad U' = U' \setminus B'_{v'}$
8: **end while**
9: Output X



communication/storage

Assuming data is partitioned across machines the communication cost is O(km log n) and the memory usage on each machine is O(max{n/m,m(k+z) log n})

worst-case guarantee

Algorithm 4 is guaranteed to be at most 13 times worse than the optimal clustering

potential applications

Results

clustering datasets

Table 1. The cluster datasets (and their descriptions) used for evaluation.

name	description	n	dim.
Parkinsons [5]	patients with early-stage Parkinson's disease	5,875	22
Census [6]	census household information	45,222	12
Skin [6]	RGB-pixel samples from face images	245,057	3
Yahoo [4]	web-search ranking dataset (features are GBRT outputs [7])	473,134	500
Covertype [6]	a forest cover dataset with cartographic features	522,911	13
Power [6]	household electric power readings	2,049,280	7
Higgs [6]	particle detector measurements (the seven 'high-level' features)	11,000,000	7

sequential vs. distributed

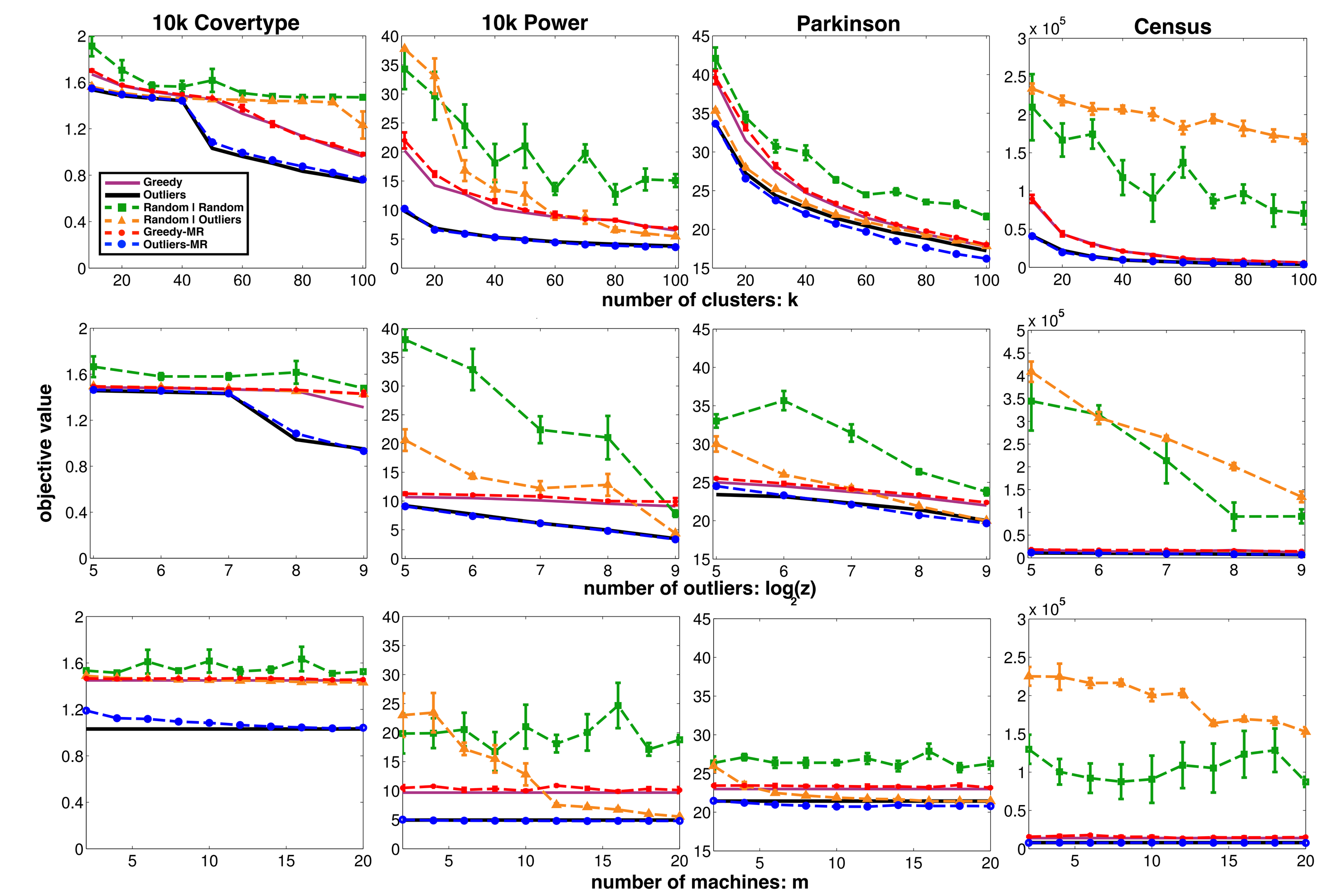


Figure 1. The performance of sequential and distributed methods. We plot the objective value of four small datasets, varying k,m,z.

large-scale experiments

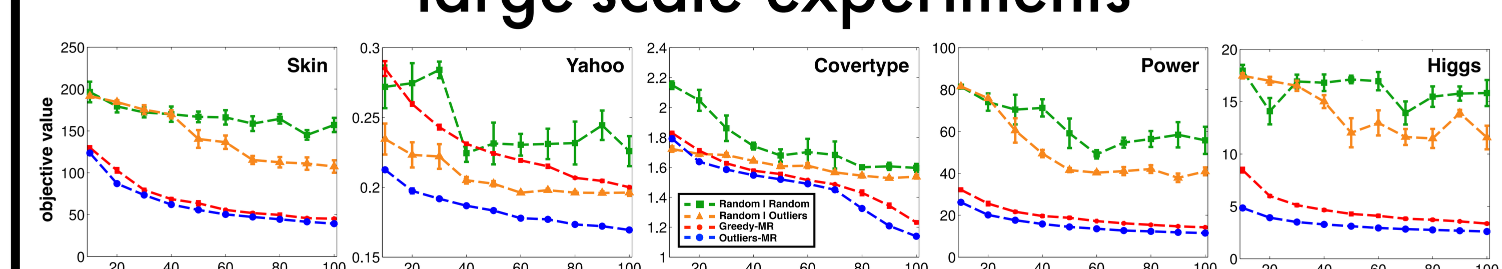


Figure 2. The objective value of five large-scale datasets, for varying k.

speedup

Table 2. The speedup of the distributed algorithms, run sequentially, over their sequential counterparts on the small datasets.On the largest of these datasets (Census) OUTLIERS-MR is more than 677x faster than OUTLIERS. This large speedup is due to the fact that we cannot store the full distance matrix for Census, thus all distances need to be computed on demand.

	k-center	outliers
10k Covertype	3.6	6.2
10k Power	4.8	9.4
Parkinson	4.9	4.4
Census	12.4	677.7

References

[1] Hochbaum, D. S. and Shmoys, D. B. A best possible heuristic for the k-center problem. Mathematics of Operations Research, 10(2): 180-184, 1985.
[2] Charikar, M., Khuller, S., Mount, D. M., and Narasimhan, G. Algorithms for facility location problems with outliers. In SODA, pages 642-651, 2001.
[3] Mirzasoleiman, B., Karbasi, A., Sarkar, R., and Krause, A. Distributed submodular maximization: Identifying representative elements in massive data. In NIPS, pages 2049-2057, 2013.
[4] Chen, M., Xu, Z., Weinberger, K. Q., Chapelle, O., Kedem, D. Classifier cascade for minimizing feature evaluation cost. In AISTATS, pages 219-226, 2012.
[5] Tsanas, A., Little, M. A., McSharry, P. E., Ramig, L. O. Enhanced classical dysphonia measures and sparse regression for telemonitoring of parkinson's disease progression. In ICASSP, pages 594-597. IEEE, 2010.
[6] Lichman, M. UCI machine learning repository [https://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science
[7] Tyree, S., Weinberger, K. Q., Agrawal, K. Parallel boosted regression trees for web search ranking. WWW, pg. 387-396, 2011.
[8] Marfil, C.F., Camadro, E.L., Masuelli, R.W. Phenotypic instability and epigenetic variability in a diploid potato of hybrid origin, Solanum ruiz-lealii. BMC Plant Biology, 2009.

Acknowledgement

GM was supported by CAPES/BR; MJK and KQW were supported by the NSF grants IIA-1355406, IIS-1149882, EFRI-1137211; and BM was supported by the Google and Yahoo Research Awards.