An Empirical Evaluation of k-Means Coresets*

Vincent Cohen-Addad† Chris Schwiegelshohn Omar Ali Sheikh-Omar‡

Abstract

Coresets are among the most popular paradigms for summarizing data. In particular, there exist many highly performance coresets for clustering problems such as k-means in both theory and practice. Curiously, there exists little work on comparing the quality of available k-means coresets.

In this paper we perform such an evaluation. First, we show that it is computationally hard to compare the quality of not only two different coreset algorithms, but also of two different output of a (randomized) coreset algorithm. To this end, we propose and analyse a benchmark for coreset evaluation. Using this benchmark and real-world data sets, we conduct an exhaustive evaluation of the most commonly used coreset implementations.

1 Introduction

The design an analysis of scalable algorithms has become an important research area over the past two decades. This is particularly important in data analysis, where even polynomial running time might not be enough to handle proverbial *big data* sets. One of the main approaches to deal with the scalability issue is to compress or sketch large data sets into smaller, more manageable ones. The aim of such compression methods is to preserve the properties of the original data, up to some small error, while significantly reducing the number of data points.

Among the most popular and successful paradigms in this line of research are *coresets*. Informally, given a data set A, a coreset $S \subset A$ with respect to a given set of queries Q and query function $f : A \times Q \to \mathbb{R}_{\geq 0}$ approximates the behaviour of A for all queries up to some multiplicative distortion D via

$$\sup_{q \in Q} \max\left(\frac{f(S,q)}{f(A,q)}, \frac{f(A,q)}{f(S,q)}\right) \leq D.$$

Coresets have been applied to a number of problems such as computational geometry [1, 4], linear algebra [12, 14], and machine learning [15, 16]. But the by far most intensively studied an arguably most successful applications of the coreset framework are k-clustering problem.

Here we are given n points A in some metric space with distance function dist and aim to find k centers C such that

$$\text{cost}(C) := \frac{1}{n} \sum_{p \in A} \min_{c \in C} \text{dist}^z(p, c)$$

is minimized. The most popular variant of this problem is probably the k-means problem in d-dimensional Euclidean space where $z = 2$ and $\text{dist}(x, y) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}$. In a long line of work originated more than 15 years ago[2, 3, 5, 7, 8, 10, 9, 11, 3, 13, 17], the size of coresets has been steadily improved with the current state of the art yielding a coreset of $\tilde{O}(k\varepsilon^{-4})$ points for a distortion $D \leq (1+\varepsilon)$ due to Cohen-Addad, Saulpic, and Schwiegelshohn [6].

While we have a good grasp of the theoretical guarantees of these algorithms, our understanding of the empirical performance is somewhat lacking. This is due to two main reasons.

- Experiments are geared towards optimization: Often experiments on coresets are conducted as follows. First, compute coreset(s) with the available algorithm(s). Then run an optimization algorithm. The *best* coreset algorithm is considered to be the one resulting in the clustering with smallest cost.
- Evaluating the quality of a coreset is hard: Given two point sets A and B, it is hard to determine the distortion when considering B as a candidate coreset of A with respect to k-clustering problems in most metrics. Thus, while we can default to the worst case guarantees from theory, it is difficult to compare the output of two coreset algorithms for a given data set.

These two reasons are related. Due to the difficulty of evaluating coresets, comparing the outcome of an optimization algorithm is a simple and reasonable alternative inasmuch as no large cost clustering becomes a low

*The full version of the paper can be accessed at `https://arxiv.org/abs/1902.09310`

†Google Zürich, Switzerland

‡Aarhus University, Denmark.

cost clustering if the coreset computation was successful. Nevertheless, this method of comparison has the drawback that it is more likely to measure the performance of the underlying optimization problem, rather than evaluating coresets.

Thus, the purpose of this paper is to propose a benchmark for k-means coresets in Euclidean spaces and use it to empirically evaluate current coreset algorithms. We argue why this benchmark has properties that make it both hard for all known coreset constructions. In addition, we also show how to efficiently estimate the distortion of a candidate coreset on the benchmark.

2 Coreset Algorithms

Though the algorithms vary in details, coreset constructions come in one of the following two flavours:

3 Hardness of Coreset Evaluation and a Benchmark

References

[1] Pankaj K. Agarwal, Sariel Har-Peled, and Kasturi R. Varadarajan. Geometric approximation via coresets. In *Combinatorial and computational geometry, MSRI*, pages 1–30. University Press, 2005.

[2] Luca Becchetti, Marc Bury, Vincent Cohen-Addad, Fabrizio Grandoni, and Chris Schwiegelshohn. Oblivious dimension reduction for k-means: beyond subspaces and the johnson-lindenstrauss lemma. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019*, pages 1039–1050, 2019.

[3] Vladimir Braverman, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for clustering in excluded-minor graphs and beyond. In Dániel Marx, editor, *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021, Virtual Conference, January 10 - 13, 2021*, pages 2679–2696. SIAM, 2021.

[4] Timothy M. Chan. Dynamic coresets. *Discret. Comput. Geom.*, 42(3):469–488, 2009.

[5] Ke Chen. On coresets for k-median and k-means clustering in metric and Euclidean spaces and their applications. *SIAM J. Comput.*, 39(3):923–947, 2009.

[6] Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. A new coreset framework for clustering. In Samir Khuller and Virginia Vassilevska Williams, editors, *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, pages 169–182. ACM, 2021.

[7] Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pages 569–578, 2011.

[8] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca, and projective clustering. *SIAM J. Comput.*, 49(3):601–657, 2020.

[9] Sariel Har-Peled and Akash Kushal. Smaller coresets for k-median and k-means clustering. *Discrete & Computational Geometry*, 37(1):3–19, 2007.

[10] Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing, Chicago, IL, USA, June 13-16, 2004*, pages 291–300, 2004.

[11] Lingxiao Huang and Nisheeth K. Vishnoi. Coresets for clustering in euclidean spaces: importance sampling is nearly optimal. In Konstantin Makarychev, Yury Makarychev, Madhur Tulsiani, Gautam Kamath, and Julia Chuzhoy, editors, *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*, pages 1416–1429. ACM, 2020.

[12] Piotr Indyk, Sepideh Mahabadi, Shayan Oveis Gharan, and Alireza Rezaei. Composable core-sets for determinant maximization problems via spectral spanners. In Shuchi Chawla, editor, *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms, SODA 2020, Salt Lake City, UT, USA, January 5-8, 2020*, pages 1675–1694. SIAM, 2020.

[13] Michael Langberg and Leonard J. Schulman. Universal ε-approximators for integrals. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010*, pages 598–607, 2010.

[14] Alaa Maalouf, Ibrahim Jubran, and Dan Feldman. Fast and accurate least-mean-squares solvers. In *Advances in Neural Information Processing Systems*, pages 8307–8318, 2019.

[15] Tung Mai, Anup B. Rao, and Cameron Musco. Coresets for classification - simplified and strengthened. *CoRR*, abs/2106.04254, 2021.

[16] Alexander Munteanu, Chris Schwiegelshohn, Christian Sohler, and David P. Woodruff. On coresets for logistic regression. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 6562–6571, 2018.

[17] Christian Sohler and David P. Woodruff. Strong coresets for k-median and subspace approximation: Goodbye dimension. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*, pages 802–813, 2018.