Universal Weak Coreset (Ragesh Jaiswal and Amit Kumar, CSE, IIT Delhi) [AAAI'24]

Alice and Bob meet for a discussion.

Hi. I am a beginner data analyst. I was wondering whether it is possible to compress data for arbitrary data analysis tasks.

No. Loss of information can easily skew the results if you do not know which task you will be performing.

That makes sense… what if we restrict the type of data processing tasks to be performed?

Maybe… but you need to make it more precise. What type of data processing are you interested in?

1

Well. I am interested in data clustering.

Great. Can you make it a bit more precise? We need to be able to compare results on the original and compressed data.

I am interested in solving the k-means* clustering problem. That has a precise mathematical formulation.

Great. In that case, you can consider using an object known as a "coreset".

Be careful while looking up the word… search engines tend to spellcheck it to corset.

It indeed does! Let me not search. Could you please tell me more about Coresets?

**The k-means problem: Given $X \subset \mathbb{R}^d$ and integer k, find a center set $C \subset \mathbb{R}^d$ with $|C| = k$ such that the cost $\Psi(C,X) \equiv \sum_{x \in X} \min_{c \in C} \|x - c\|^2$ is minimized.*

2

Okay… let me define an ε-coreset for the k-means problem.

An ε-coreset for the k-means problem with input X, k is a weighted pointset (Y, w) with $|Y| \ll |X|$ such that for any k center set C:

$$(1-\varepsilon) \cdot \Psi(C, Y, w) \leq \Psi(C, X) \leq (1+\varepsilon) \cdot \Psi(C, Y, w).$$

I see… so a coreset (Y, w) mimics the dataset X in terms of retaining the k-means cost for "any" k-center set. That's awesome!

3

It is pretty cool! Do you see how this would mean that solving the k-means problem on the coreset gives a good solution (within $(1 \pm O(\varepsilon))$) for the original dataset?

Yes. I do see that.

An ε-coreset for the k-means problem with input X, k is a weighted pointset (Y, w) with $|Y| \ll |X|$ such that for any k center set C:

$$(1-\varepsilon) \cdot \Psi(C, Y, w) \leq \Psi(C, X) \leq (1+\varepsilon) \cdot \Psi(C, Y, w).$$

Moreover, the coreset construction can be applied repeatedly to keep compressing streaming data. This gives a "streaming algorithm" for the k-means problem. The idea is called "merge-and-reduce".

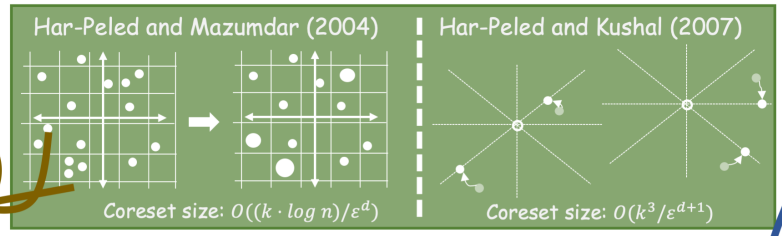
Wow! Coresets have strong properties. Now I am even more curious. How are they constructed? What kind of compression do they give?

4

Universal Weak Coreset (Ragesh Jaiswal and Amit Kumar, CSE, IIT Delhi) [AAAI'24]

Well… there's a bit of history… let me recollect.

The initial construction ideas were geometric…. ideas such as throw a grid around the points and fuse points in a cell… snap points to rays emanating from a good center set, etc.

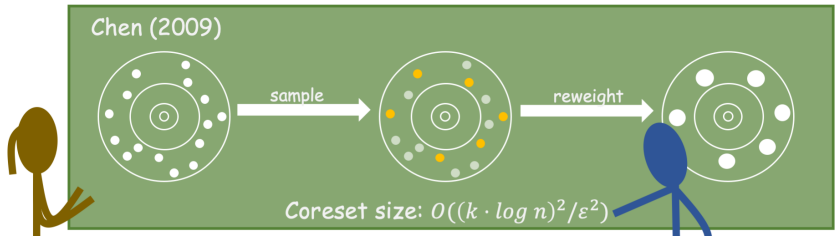


A coreset with size independent of the dataset! That looks impressive! At the cost of sounding greedy, are there such coresets for general metric spaces… I mean non-geometric data as well? The above geometric ideas may not work.

5

For general metric spaces, there are sampling-based constructions. They use ideas such as sample from concentric "rings" around a representative k center set and reweight the sampled points to represent points in the same ring.

Unlike the geometric coreset, I see a dependence on the data size n.

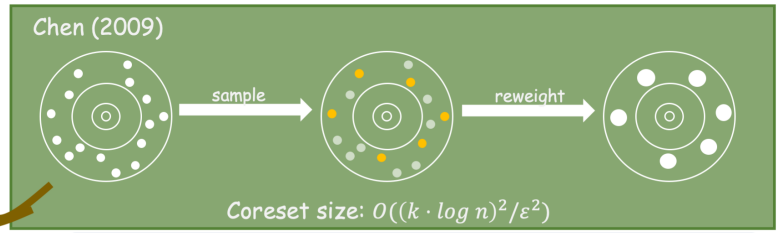


Oh …that comes from the union-bound… let me explain.

The Metric k-means problem: Let $(\mathcal{X}, D)$ be a metric. Given $X \subset \mathcal{X}$ and k, find a center set $C \subset \mathcal{X}$ with $|C| = k$ such that the cost $\Psi(C, X) \equiv \sum_{x \in X} \min_{c \in C} D(x, c)^2$ is minimized.

6

It's a randomized construction. The argument works like this: for a fixed k center set C, (sampling+reweighting) $(k/\varepsilon)^2$ points satisfies the coreset property with high probability. However, we want it to work for "every" k center set and there could be n^k of them. For the union-bound to work, we must blow up the sample size by a factor of (log n).

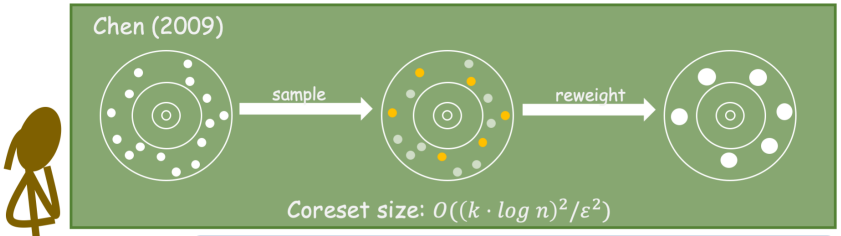


I see… if it were geometric data, one could use something like an ε-net to reduce the number of possibilities for the union bound from n^k to something much smaller… but I guess that is not possible for an arbitrary metric space.

Precisely…

7

In fact, for specific metric spaces, such as the Euclidean space, these ε-set, ε-centroid set ideas have been used to obtain small sized coresets… and that has been the agenda of most research on coresets in the recent past. However, for general metric spaces, the (log n)-barrier exists. This is also supported by a lower bound.



Bummer! Perhaps coreset is too strong a notion and we can appropriately weaken it to break the (log n) barrier without compromising too much on the coreset properties.

Hmmm…that's an interesting take on this… let's discuss.

8

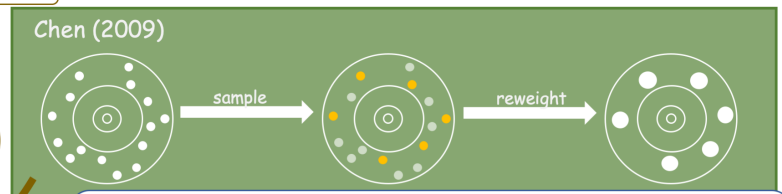
Universal Weak Coreset (Ragesh Jaiswal and Amit Kumar, CSE, IIT Delhi) [AAAI'24]

Hold on… there is another issue that has been bothering me about coresets.

What?

A coreset allows data compression when the objective is fixed. In our case, it is solving the k-means problem.

Yes.



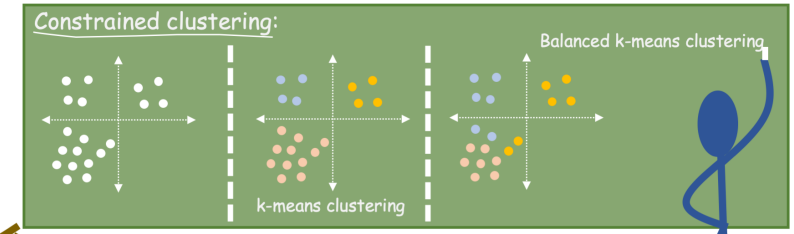
…but if I know that all I am ever going to do with the data is solve the k-means problem, then I can as well solve the problem on the data and then save the solution while throwing away the data. So, what is the point of a coreset?

Well… you are forgetting the streaming application. However, you do raise a pertinent point regarding coresets from a purely compression viewpoint.

9

Discussions on coresets have lately drifted completely towards the streaming application, even though it was initially designed as a data compression tool.

I have an application in mind that may justify the data compression aspect. Lately, constrained clustering has gained importance where the clustering must satisfy certain constraints in addition to optimizing the k-means cost.



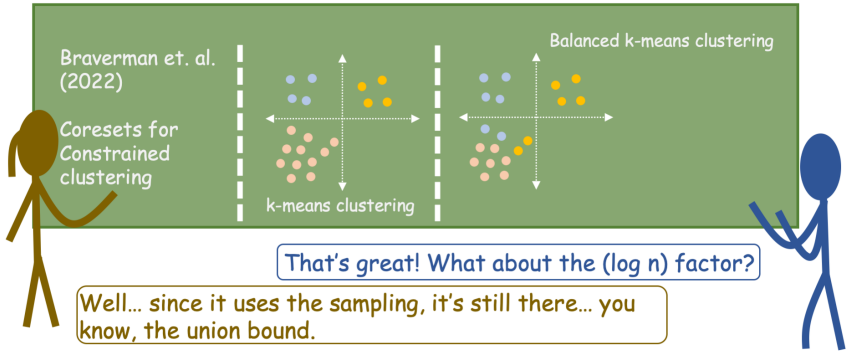
One example is balanced clustering, where the clusters should have roughly equal size. Other examples include various fairness notions.

I'm aware… could you elaborate on the coreset aspect?

10

If we can compress while being oblivious of the specific constraints the clusters need to satisfy, then computing the coreset makes sense. Many times, the constraints are not known at the time of access to the data.

You are in luck! There has been a recent development in coreset construction using Chen's uniform sampling idea that works for a wide range of constrained clustering problems.



That's great! What about the (log n) factor?

Well… since it uses the sampling, it's still there… you know, the union bound.

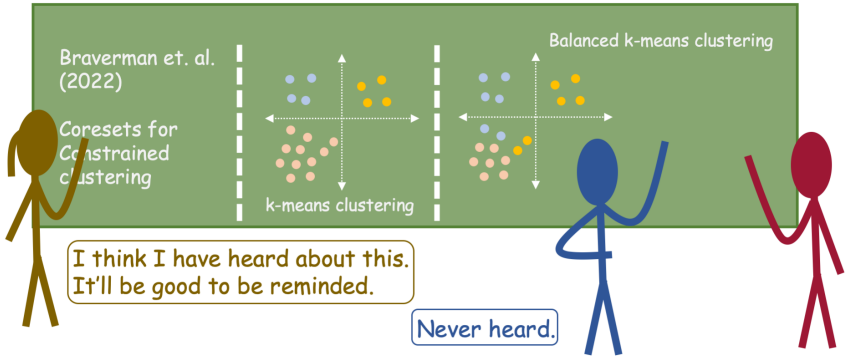
11

Charlie joins the discussions.

Hi Charlie!

Hi Charlie!

Hi guys! I couldn't help overhearing you guys discussing constrained clustering. Do you guys know about this line of research where they design a meta-algorithm that works for any constrained k-means problem?



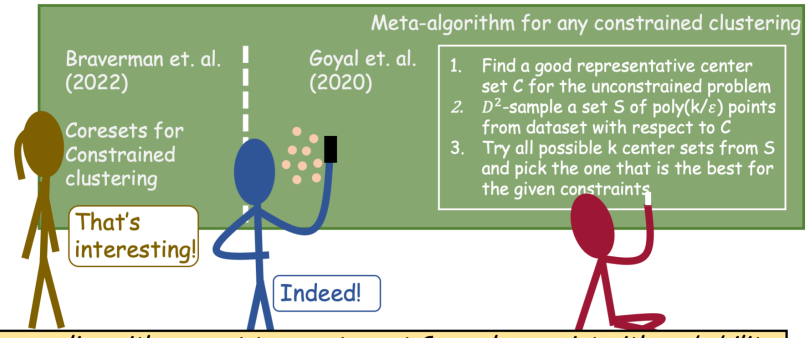
I think I have heard about this. It'll be good to be reminded.

Never heard.

12

Universal Weak Coreset (Ragesh Jaiswal and Amit Kumar, CSE, IIT Delhi) [AAAI'24]

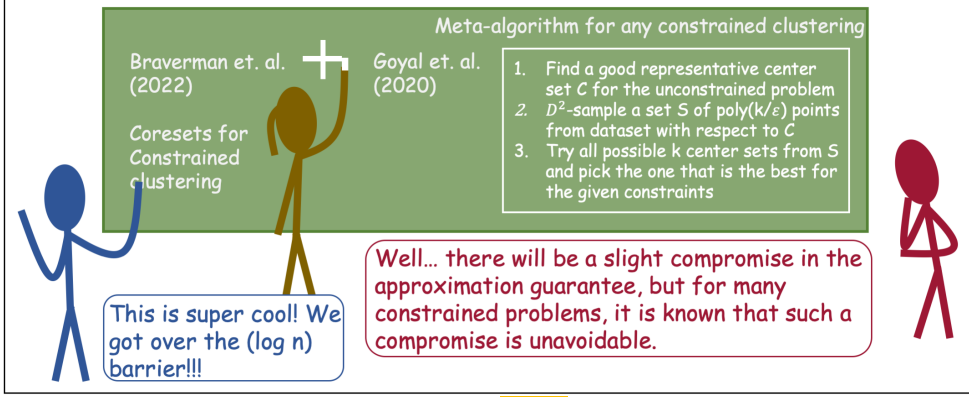
The main idea of the meta-algorithm is to D^2-sample* a set S of $poly(k/\varepsilon)$ points with respect to a decent representative center set C, and then argue that for any clustering, S contains good centers for that clustering with high probability. The "goodness" is in terms of an approximation guarantee.



D^2-sampling with respect to a center set C samples a point with probability proportional to its squared distance from the nearest center in C.

13

Wait a minute. Doesn't this mean that if we simultaneously use the meta-algorithm AND construct the sampling-based coreset, then we would only need to take the union bound over only $|S|^k = \left(\frac{k}{\varepsilon}\right)^{O(k)}$ possibilities. This would give a coreset size independent of n, even if you include the size of the set S.

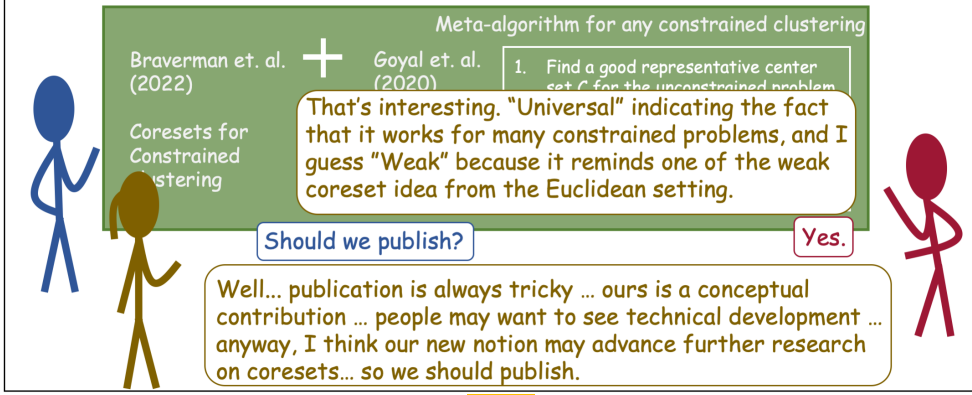


14

This new coreset notion, even though interesting, raises many more interesting questions, including one about streaming construction.

That's true. However, I'm excited! What should we call our coreset?

How about "Universal Weak Coreset?"



15

References

- (Chen, 2009) Chen, K. 2009. On Coresets for k-Median and k-Means Clustering in Metric and Euclidean Spaces and Their Applications. *SIAM Journal on Computing*, 39(3): 923-947.
- (Braverman et. al., 2022) Braverman, V.; Cohen-Addad, V.; Jiang, H.; Krauthgamer, R.; Schwiegelshohn, C.; Toftrup, M.; and Wu, X. 2022. The Power of Uniform Sampling for Coresets. *In 2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, 462-473.
- (Goyal et. al., 2020) Goyal, D.; Jaiswal, R.; and Kumar, A. 2020. FPT Approximation for Constrained Metric k-Median/Means. *In Cao, Y.; and Pilipczuk, M., eds., 15th International Symposium on Parameterized and Exact Computation (IPEC 2020)*, volume 180 of Leibniz International Proceedings in Informatics (LIPIcs), 14:1-14:19.
- (Har-Peled and Kushal, 2007) Har-Peled, S.; and Kushal, A. 2007. Smaller coresets for k-median and k-means clustering. *Discrete & Computational Geometry*, 37(1): 3 - 19.
- (Har-Peled and Mazumdar, 2004) Har-Peled, S.; and Mazumdar, S. 2004. On Coresets for k-Means and k-Median Clustering. *In Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing, STOC '04*, 291-300.

16