Nonparametric Coreset for Clustering

Rachit Chhaya
IIT Gandhinagar

Jayesh Choudhari
Univ. of Warwick

Anirban Dasgupta
IIT Gandhinagar

Supratim Shit
Technion

k - Clustering

GIven a set **A** of n points in R^d, the solution of the problem is **X*** with k set of centres such that

$$\mathbf{X}^* = \arg\min_{\mathbf{X}} f_{\mathbf{X}}(\mathbf{A})$$

where, $$f_{\mathbf{X}}(\mathbf{A}) = \sum_{\mathbf{a}\in\mathbf{A}} \min_{\mathbf{x}\in\mathbf{X}} \|\mathbf{x}-\mathbf{a}\|^2$$

Coreset

GIven **A** $\in R^{n \times d}$, a coreset **C** is a weighted subsample of **A**, such that |**C**| << |**A**| and with high probability, ∀**X** with k centres.

$$|f_{\mathbf{X}}(\mathbf{C}) - f_{\mathbf{X}}(\mathbf{A})| \leq \epsilon f_{\mathbf{X}}(\mathbf{A})$$

A coreset **C** is non-parametric if its size is independent of both k (#cluster), and it still ensures the above guarantee ∀**X** with at most n centres in R^d.

Such a coreset does not exist!!

Non-Parametric Coreset

Given **A**, there exists a non parametric coreset with small additive error.

$$|f_{\mathbf{X}}(\mathbf{C}) - f_{\mathbf{X}}(\mathbf{A})| \leq \epsilon(f_{\mathbf{X}}(\mathbf{A}) + f_{\varphi}(\mathbf{A}))$$

The additive factor depends on the data. Here, φ = mean(**A**).

To show its existence we rely on importance sampling which is based on sensitivity framework along with barrier potential functions.

For streaming inputs we define online sensitivity scores, that depends on on the points the algorithm have seen so far.

Importance Score

Barrier Potential based Sensitivity Function

$$\sup_{\mathbf{X}} \frac{f_{\mathbf{X}}(\mathbf{a}_i)}{(1+\epsilon)f_{\mathbf{X}}(\mathbf{A}_{i-1}) - f_{\mathbf{X}}(\mathbf{C}_{i-1}) + \epsilon f_{\varphi_i}(\mathbf{A}_i)}$$

$$\sup_{\mathbf{X}} \frac{f_{\mathbf{X}}(\mathbf{a}_i)}{f_{\mathbf{X}}(\mathbf{C}_{i-1}) - (1-\epsilon)f_{\mathbf{X}}(\mathbf{A}_{i-1}) + \epsilon f_{\varphi_i}(\mathbf{A}_i)}$$

Expected Upper Bound

$$\frac{2f_{\varphi_i}^{\mathbf{M}_i}(\mathbf{a}_i)}{\mu_i\epsilon\sum_{j\leq i} f_{\varphi_j}^{\mathbf{M}_j}(\mathbf{a}_j)} + \frac{12}{\mu_i\epsilon(i-1)}$$

Getting a true upper bound is challenging!!

NonParametricFilter Result

Sampling points in **C** based on the above sensitivity scores ensures the following ∀**X** with at most n centres,

$$|f_{\mathbf{X}}(\mathbf{C}) - f_{\mathbf{X}}(\mathbf{A})| \leq \epsilon(f_{\mathbf{X}}(\mathbf{A}) + f_{\varphi}(\mathbf{A}))$$

Coreset Size: $O\left(\frac{\log n}{\mu\epsilon^2}\left(\log n + \log\left(f_{\varphi}^{\mathbf{M}}(\mathbf{A})\right) - \log\left(f_{\varphi_2}^{\mathbf{M}_2}(\mathbf{a}_2)\right)\right)\right)$

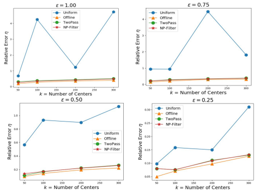
Online Coreset with Deterministic Guarantee!!

Experiments

Compare performance of *NonParametricFilter* with baselines on real world data.

After getting the coreset **C**
- Run k-clustering on **C** and **A**
- Use these centres and report

$$\eta = \frac{|f_{\mathbf{X_C}}(\mathbf{A}) - f_{\mathbf{X_A}}(\mathbf{A})|}{f_{\mathbf{X_A}}(\mathbf{A})}$$



Algorithm 2 `NonParametricFilter`

Require: $\mathbf{a}_i, i = 1,\ldots n; t > 1; \epsilon \in (0,1)$
Ensure: $(\mathbf{C},\Omega)$
$c^u = 2/\epsilon + 1; c^l = 2/\epsilon - 1; \varphi_0 = \emptyset; S = 0; \mathbf{C}_0^1 = \ldots = \Omega_0^t = \emptyset$
$\lambda = \|\mathbf{a}_1\|_{\min}; \quad \nu = \|\mathbf{a}_1\|_{\max}$
while $i \leq n$ **do**
  $\lambda = \min\{\lambda, \|\mathbf{a}_i\|_{\min}\}; \nu = \max\{\nu, \|\mathbf{a}_i\|_{\max}\}$
  Update $\mathbf{M}_i$; $\mu_i = \lambda/\nu$
  $\varphi_i = ((i-1)\varphi_{i-1} + \mathbf{a}_i)/i; S = S + f_{\varphi_i}^{\mathbf{M}_i}(\mathbf{a}_i)$
  if $i = 1$ **then**
    $p_i = 1$
  else
    $\hat{l}_i^u = \frac{2f_{\varphi_i}^{\mathbf{M}_i}(\mathbf{a}_i)}{\epsilon\mu_i S} + \frac{12}{\epsilon\mu_i(i-1)}$
    $\hat{l}_i^l = \frac{2f_{\varphi_i}^{\mathbf{M}_i}(\mathbf{a}_i)}{\epsilon\mu_i S} + \frac{12}{\epsilon\mu_i(i-1)}$
    $p_i = \min\{1, (c^u\hat{l}_i^u + c^l\hat{l}_i^l\}$
  end if
  for $\forall j \in [t]$ **do**
    $(\mathbf{c}_i^j, \omega_i^j) = \begin{cases} (\mathbf{a}_i, 1/(tp_i)) & \text{w. p. } p_i \\ (\emptyset, 0) & \text{else} \end{cases}$
    $(\mathbf{C}_i^j, \Omega_i^j) = (\mathbf{C}_{i-1}^j, \Omega_{i-1}^j) \cup (\mathbf{c}_i^j, \omega_i^j)$
  end for
  $(\mathbf{C},\Omega) = (\cup_{j\leq t}\mathbf{C}_{i-1}^j, \cup_{j\leq t}\Omega_{i-1}^j)$
end while
Return $(\mathbf{C},\Omega)$

Future Scope

- Computing the actual upper bound of the sensitivity scores.
- Coresets with deterministic guarantee for general loss functions.
- Showing lower bound of coresets for any loss functions.

References

- Lucic, Mario, Olivier Bachem, and Andreas Krause. "Strong coresets for hard and soft Bregman clustering with applications to exponential family mixtures." In Artificial intelligence and statistics, pp. 1-9. PMLR, 2016.
- Bachem, Olivier, Mario Lucic, and Andreas Krause. "Scalable k-means clustering via lightweight coresets." Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018.
- Batson, Joshua, Daniel A. Spielman, and Nikhil Srivastava. "Twice-ramanujan sparsifiers." SIAM Journal on Computing 41.6 (2012): 1704-1721.
- Feldman, Dan, and Michael Langberg. "A unified framework for approximating and clustering data." Proceedings of the forty-third annual ACM symposium on Theory of computing. 2011.
- Banerjee, Arindam, Srujana Merugu, Inderjit S. Dhillon, Joydeep Ghosh, and John Lafferty. "Clustering with Bregman divergences." Journal of machine learning research 6, no. 10 (2005).