$\alpha \le 1/7 \quad (1 + 20\alpha)$

Improved Learning Augmented Algorithms for K-mean & K-median

~~Nguyen et al~~ ICLR'23

$\bar{\lambda}$ label error

K-mean

$$\alpha \in [0, \tfrac{1}{2}] \quad 1 + \frac{5\alpha - 2\alpha^2}{(1-2\alpha)(1-\alpha)} \Rightarrow \alpha = 1/7 \Rightarrow 1 + 7.7\alpha$$

$$1 + 20\alpha$$

$|P| = m$

$\bar{\lambda} \quad (P_1 \ldots P_k) \qquad m_i = |P_i|$

$\exists$ Optimal Partition $(P_1^*, P_2^*, \ldots P_k^*)$ Centre $(C_1^*, \ldots C_k^*)$ s.t

$$\sum_{i \in [k]} \text{Cost}(P_i^*, C_i^*) = OPT \qquad m_i^* = |P_i^*|$$

For each Cluster i, denote $P_i^* \cap P_i = Q_i$

Note that $|Q_i| \ge (1-\alpha)\max(|P_i|, |P_i^*|) \quad \alpha < 1/2$

Denote $\bar{X} \to \bar{X} \quad X_i, P_i, X_{ij}, P_{ij}$

Lemma 1:- Consider a Set $X \subset \mathbb{R}^d$ of Size n & $C \in \mathbb{R}^d$

$$\text{Cost}(X, C) = \text{Cost}(X, \bar{X}) + n \cdot \|C - \bar{X}\|^2 \Big\} \text{ Inaba}$$

$Q_i = P_i \cap P_i^*$

Lemma 2 $\text{Cost}(P_i^*, \bar{Q}_i) \le \left(1 + \frac{\alpha}{1-\alpha}\right) \text{Cost}(P_i^*, \overline{P_i^*})$

$$\text{Cost}(P_i^*, C_i^*)$$

Lemma 3:- For any Partition $J_1 \cup J_2$ of a Set $J \subset \mathbb{R}$ of Size n, if $|J_1| \ge (1-\alpha)n$ then $|\bar{J} - \bar{J}_1|^2 \le \frac{\alpha}{n(1-\alpha)} \text{Cost}(J, \bar{J})$

Algorithm

Input :- Dataset P, Partition $P_1 \cdots P_k$, α

for $i \in [k]$ do

for $j \in [d]$ do

Let ω_{ij} be the Collection of all Subset of $(1-\alpha)m_i$ Contiguous Points in P_{ij}

$$Z = \{\ldots\} \quad I_{ij} \leftarrow \arg\min_{Z \in \omega_{ij}} Cost(Z, \bar{Z}) = \arg\min_{Z \in \omega_{ij}} \sum_{z \in Z} z^2 - \frac{1}{|Z|}\left(\sum_{z' \in Z} z'\right)^2$$

Let $\hat{C}_i = (\bar{I}_{ij})_{j \in d}$

Return $\hat{C}_1 \cdots \hat{C}_k$

Lemma 4 For $\forall\, i \in [k]$, $j \in [d]$, let ω_{ij}' be the Collection of all Subsets of $(1-\alpha)m_i$ Points in P_{ij}. Then

$$Cost(I_{ij}, \bar{I}_{ij}) = \min_{Z' \in \omega_{ij}'} Cost(Z, \bar{Z})$$

Proof :- Suppose $I_{ij}' = \arg\min_{Z' \in \omega_{ij}'} Cost(Z', \bar{Z'})$

If $I_{ij}' \in \omega_{ij}$ Done

If $I_{ij}' \notin \omega_{ij}$. Let a & b be the min & max Point I_{ij}'

We know $\exists\, p \in P_{ij} \cap (a, b)$ s.t $p \notin I_{ij}'$.

If $|I_{ij}'| = 2$ then

$$Cost(I_{ij}', \bar{I}_{ij}') = \frac{(b-a)^2}{2} > \frac{(b-p)^2}{2}$$

If $|I'_{ij}| \geq 3$. We know that a and b is the furthest point from $\overline{I'_{ij} \setminus \{a,b\}}$ in the interval $[a,b]$.

$$K_{ij} = \left\{ I'_{ij} \setminus a \right\} \cup P$$

$$Cost(K_{ij} \setminus P, P) = Cost(K_{ij} \setminus \{P, b\}, P) + |P-b|^2$$

$$= Cost(I'_{ij} \setminus \{a,b\}, P) + |P-b|^2$$

$$= Cost(I'_{ij} \setminus \{a,b\}, \overline{I'_{ij} \setminus \{a,b\}}) + |I'_{ij} \setminus \{a,b\}| \cdot |P - \overline{I'_{ij} \setminus \{a,b\}}|^2 + |P-b|^2$$

$$= Cost(I'_{ij} \setminus \{a,b\}, a) + |a-b|^2$$

$$= Cost(I'_{ij} \setminus \{a\}, a)$$

Now,

$$Cost(K_{ij}, \overline{K_{ij}}) = \frac{1}{(1-\alpha)m_i} \sum_{y_1, y_2 \in K_{ij}} |y_1 - y_2|^2$$

$$= \frac{1}{(1-\alpha)m_1} \left(\sum_{y_1, y_2 \in K_{ij} \setminus a} |y_1 - y_2|^2 + Cost(K_{ij} \setminus P, P) \right)$$

$$\leq \frac{1}{(1-\alpha)m_1} \left(\sum_{y_1, y_2 \in I'_{ij} \setminus a} |y_1 - y_2|^2 + Cost(I'_{ij} \setminus a, a) \right)$$

$$= Cost(I'_{ij}, \overline{I_{ij}})$$

$$Cost(X, C) = Cost(X, \overline{X}) + n\|C - \overline{X}\|^2$$

$$\Downarrow$$

$$C = \overline{I_{ij}}, \quad X = P^*_{ij} \qquad \overline{X} = \overline{P_{ij}} \qquad n = m_1^*$$

$$\left| \overline{P^*_{ij}} - \overline{I_{ij}} \right|^2 \quad \frac{Cost(P_{ij}, \overline{P_{ij}})}{m_1}$$

Lemma 5 $\left|\overline{P_{ij}^{*}} - \overline{I_{ij}}\right|^2 \le \left(\frac{\alpha}{1-\alpha} + \frac{4\alpha}{(1-2\alpha)(1-\alpha)}\right) \frac{\mathrm{Cost}\left(P_{ij}^{*}, \overline{P_{ij}^{\alpha}}\right)}{m_i}$

$$\mathrm{Cost}\left(P_{ij}^{\alpha}, \overline{I_{ij}}\right) = \mathrm{Cost}\left(P_{ij}^{\alpha}, \overline{P_{ij}^{\alpha}}\right) + |m_i^{\alpha}| \left|\overline{I_{ij}} - \overline{P_{ij}^{\alpha}}\right|^2$$

$$\le \mathrm{Cost}\left(P_{ij}^{\alpha}, \overline{P_{ij}^{\alpha}}\right)\left(1 + \frac{\alpha}{(1-\alpha)} + \frac{4\alpha}{(1-2\alpha)(1-\alpha)}\right)$$

$$(1 + 7.7\alpha)$$

$$O(dm \log m)$$