weighted quantile sketch algorithm

倪杰

2021 年 5 月 19 日

摘要

XGBoost 的近似算法需要找到合适的分割点。当每个数据的权重都相同时，现有的 quantile sketch 算法就可以解决，但当权重不相同时，没有有理论支撑的可用算法。为此，作者提出了 weighted quantile sketch 算法。这种算法总体上有两种操作，分别是 merge(归并) 和 prune(剪枝). 并且每一步操作都有一定的误差上界.

目录

1 问题

Formally, let multi-set $\mathcal{D}_k = \{(x_{1k}, h_1), (x_{2k}, h_2) \cdots (x_{nk}, h_n)\}$ represent the k-th feature values and second order gradient statistics of each training instances. We can define a rank functions $r_k : \mathbb{R} \to [0, +\infty)$ as

$$r_k(z) = \frac{1}{\sum_{(x,h)\in\mathcal{D}_k} h} \sum_{(x,h)\in\mathcal{D}_k, x<z} h, \tag{8}$$

which represents the proportion of instances whose feature value k is smaller than z. The goal is to find candidate split points $\{s_{k1}, s_{k2}, \cdots s_{kl}\}$, such that

$$|r_k(s_{k,j}) - r_k(s_{k,j+1})| < \epsilon, \quad s_{k1} = \min_i \mathbf{x}_{ik}, s_{kl} = \max_i \mathbf{x}_{ik}. \tag{9}$$

Here ϵ is an approximation factor. Intuitively, this means that there is roughly $1/\epsilon$ candidate points. Here each data point is weighted by h_i. To see why h_i represents the weight, we can rewrite Eq (3) as

$$\sum_{i=1}^{n} \frac{1}{2} h_i (f_t(\mathbf{x}_i) - g_i/h_i)^2 + \Omega(f_t) + constant,$$

2 定义和记号

A.1 Formalization and Definitions

Given an input multi-set $\mathcal{D} = \{(x_1, w_1), (x_2, w_2) \cdots (x_n, w_n)\}$ such that $w_i \in [0, +\infty), x_i \in \mathcal{X}$. Each x_i corresponds to a position of the point and w_i is the weight of the point. Assume we have a total order $<$ defined on $\mathcal{X}$. Let us define two rank functions $r^-_{\mathcal{D}}, r^+_{\mathcal{D}} : \mathcal{X} \to [0, +\infty)$

$$r^-_{\mathcal{D}}(y) = \sum_{(x,w)\in\mathcal{D}, x<y} w \tag{10}$$

$$r^+_{\mathcal{D}}(y) = \sum_{(x,w)\in\mathcal{D}, x\leq y} w \tag{11}$$

We should note that since $\mathcal{D}$ is defined to be a *multiset* of the points. It can contain multiple record with exactly same position x and weight w. We also define another weight function $\omega_{\mathcal{D}} : \mathcal{X} \to [0, +\infty)$ as

$$\omega_{\mathcal{D}}(y) = r^+_{\mathcal{D}}(y) - r^-_{\mathcal{D}}(y) = \sum_{(x,w)\in\mathcal{D}, x=y} w. \tag{12}$$

Finally, we also define the weight of multi-set $\mathcal{D}$ to be the sum of weights of all the points in the set

$$\omega(\mathcal{D}) = \sum_{(x,w)\in\mathcal{D}} w \tag{13}$$

Definition A.3. *ϵ-Approximate Quantile Summary*
Given a quantile summary $Q(\mathcal{D}) = (S, \tilde{r}^+_{\mathcal{D}}, \tilde{r}^-_{\mathcal{D}}, \tilde{\omega}_{\mathcal{D}})$, we call it is ϵ-approximate summary if for any $y \in \mathcal{X}$

$$\tilde{r}^+_{\mathcal{D}}(y) - \tilde{r}^-_{\mathcal{D}}(y) - \tilde{\omega}_{\mathcal{D}}(y) \leq \epsilon\omega(\mathcal{D}) \tag{21}$$

We use this definition since we know that $r^-(y) \in [\tilde{r}^-_{\mathcal{D}}(y), \tilde{r}^+_{\mathcal{D}}(y) - \tilde{\omega}_{\mathcal{D}}(y)]$ and $r^+(y) \in [\tilde{r}^-_{\mathcal{D}}(y) + \tilde{\omega}_{\mathcal{D}}(y), \tilde{r}^+_{\mathcal{D}}(y)]$. Eq. (21) means the we can get estimation of $r^+(y)$ and $r^-(y)$ by error of at most $\epsilon\omega(\mathcal{D})$.

3 扩展到实数域

Definition A.2. *Extension of Function Domains*
Given a quantile summary $Q(\mathcal{D}) = (S, \tilde{r}^+_{\mathcal{D}}, \tilde{r}^-_{\mathcal{D}}, \tilde{\omega}_{\mathcal{D}})$ defined in Definition A.1, the domain of $\tilde{r}^+_{\mathcal{D}}$, $\tilde{r}^-_{\mathcal{D}}$ and $\tilde{\omega}_{\mathcal{D}}$ were defined only in S. We extend the definition of these functions to $\mathcal{X} \to [0, +\infty)$ as follows
When $y < x_1$:

$$\tilde{r}^-_{\mathcal{D}}(y) = 0,\ \tilde{r}^+_{\mathcal{D}}(y) = 0,\ \tilde{\omega}_{\mathcal{D}}(y) = 0 \tag{16}$$

When $y > x_k$:

$$\tilde{r}^-_{\mathcal{D}}(y) = \tilde{r}^+_{\mathcal{D}}(x_k),\ \tilde{r}^+_{\mathcal{D}}(y) = \tilde{r}^+_{\mathcal{D}}(x_k),\ \tilde{\omega}_{\mathcal{D}}(y) = 0 \tag{17}$$

When $y \in (x_i, x_{i+1})$ for some i:

$$\begin{aligned} &\tilde{r}^-_{\mathcal{D}}(y) = \tilde{r}^-_{\mathcal{D}}(x_i) + \tilde{\omega}_{\mathcal{D}}(x_i), \\ &\tilde{r}^+_{\mathcal{D}}(y) = \tilde{r}^+_{\mathcal{D}}(x_{i+1}) - \tilde{\omega}_{\mathcal{D}}(x_{i+1}), \\ &\tilde{\omega}_{\mathcal{D}}(y) = 0 \end{aligned} \tag{18}$$

4 merge 操作

A.3 Merge Operation

In this section, we define how we can merge the two summaries together. Assume we have $Q(\mathcal{D}_1) = (S_1, \tilde{r}^+_{\mathcal{D}_1}, \tilde{r}^-_{\mathcal{D}_1}, \tilde{\omega}_{\mathcal{D}_1})$ and $Q(\mathcal{D}_2) = (S_2, \tilde{r}^+_{\mathcal{D}_1}, \tilde{r}^-_{\mathcal{D}_2}, \tilde{\omega}_{\mathcal{D}_2})$ quantile summary of two dataset $\mathcal{D}_1$ and $\mathcal{D}_2$. Let $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$, and define the merged summary $Q(\mathcal{D}) = (S, \tilde{r}^+_{\mathcal{D}}, \tilde{r}^-_{\mathcal{D}}, \tilde{\omega}_{\mathcal{D}})$ as follows.

$$S = \{x_1, x_2 \cdots, x_k\}, x_i \in S_1 \text{ or } x_i \in S_2 \tag{25}$$

The points in S are combination of points in S_1 and S_2. And the function $\tilde{r}^+_{\mathcal{D}}, \tilde{r}^-_{\mathcal{D}}, \tilde{\omega}_{\mathcal{D}}$ are defined to be

$$\tilde{r}^-_{\mathcal{D}}(x_i) = \tilde{r}^-_{\mathcal{D}_1}(x_i) + \tilde{r}^-_{\mathcal{D}_2}(x_i) \tag{26}$$

$$\tilde{r}^+_{\mathcal{D}}(x_i) = \tilde{r}^+_{\mathcal{D}_1}(x_i) + \tilde{r}^+_{\mathcal{D}_2}(x_i) \tag{27}$$

$$\tilde{\omega}_{\mathcal{D}}(x_i) = \tilde{\omega}_{\mathcal{D}_1}(x_i) + \tilde{\omega}_{\mathcal{D}_2}(x_i) \tag{28}$$

Here we use functions defined on $S \to [0, +\infty)$ on the left sides of equalities and use the extended function definitions on the right sides.

Due to additive nature of r^+, r^- and ω, which can be formally written as

$$\begin{aligned} r^-_{\mathcal{D}}(y) =& r^-_{\mathcal{D}_1}(y) + r^-_{\mathcal{D}_2}(y), \\ r^+_{\mathcal{D}}(y) =& r^+_{\mathcal{D}_1}(y) + r^+_{\mathcal{D}_2}(y), \\ \omega_{\mathcal{D}}(y) =& \omega_{\mathcal{D}_1}(y) + \omega_{\mathcal{D}_2}(y), \end{aligned} \tag{29}$$

5 merge 的性质

THEOREM A.1. *If $Q(\mathcal{D}_1)$ is ϵ_1-approximate summary, and $Q(\mathcal{D}_2)$ is ϵ_2-approximate summary. Then the merged summary $Q(\mathcal{D})$ is* $\max(\epsilon_1, \epsilon_2)$*-approximate summary.*

PROOF. For any $y \in \mathcal{X}$, we have

$$\begin{aligned}
&\tilde{r}^+_{\mathcal{D}}(y) - \tilde{r}^-_{\mathcal{D}}(y) - \tilde{\omega}_{\mathcal{D}}(y) \\
=&[\tilde{r}^+_{\mathcal{D}_1}(y) + \tilde{r}^+_{\mathcal{D}_2}(y)] - [\tilde{r}^-_{\mathcal{D}_1}(y) + \tilde{r}^-_{\mathcal{D}_2}(y)] - [\tilde{\omega}_{\mathcal{D}_1}(y) + \tilde{\omega}_{\mathcal{D}_2}(y)] \\
\le&\epsilon_1\omega(\mathcal{D}_1) + \epsilon_2\omega(\mathcal{D}_2) \le \max(\epsilon_1, \epsilon_2)\omega(\mathcal{D}_1 \cup \mathcal{D}_2)
\end{aligned}$$

6 prune 操作

Now we are ready to introduce the prune operation. Given a quantile summary $Q(\mathcal{D}) = (S, \tilde{r}^+_{\mathcal{D}}, \tilde{r}^-_{\mathcal{D}}, \tilde{\omega}_{\mathcal{D}})$ with $S = \{x_1, x_2, \cdots, x_k\}$ elements, and a memory budget b. The prune operation creates another summary $Q'(\mathcal{D}) = (S', \tilde{r}^+_{\mathcal{D}}, \tilde{r}^-_{\mathcal{D}}, \tilde{\omega}_{\mathcal{D}})$ with $S' = \{x'_1, x'_2, \cdots, x'_{b+1}\}$ where x'_i are selected by query the original summary such that

$$x'_i = g\left(Q, \frac{i-1}{b}\omega(\mathcal{D})\right).$$

The definition of $\tilde{r}^+_{\mathcal{D}}, \tilde{r}^-_{\mathcal{D}}, \tilde{\omega}_{\mathcal{D}}$ in Q' is copied from original summary Q, by restricting input domain from S to S'. There could be duplicated entries in the S'. These duplicated entries can be safely removed to further reduce the memory cost. Since all the elements in Q' comes from Q, we can verify that Q' satisfies all the constraints in Definition A.1 and is a valid quantile summary.

THEOREM A.2. *Let $Q'(\mathcal{D})$ be the summary pruned from an ϵ-approximate quantile summary $Q(\mathcal{D})$ with b memory budget. Then $Q'(\mathcal{D})$ is a $(\epsilon + \frac{1}{b})$-approximate summary.*