The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Variance Reduced K-means Clustering

Yawei Zhao, Yuewei Ming, Xinwang Liu, En Zhu
College of Computer
National University of Defense Technology
Changsha, China, 410073

Jianping Yin
Dongguan University of Technology
Dongguan, China, 523000

Abstract

It is challenging to perform k-means clustering on a large scale dataset efficiently. One of the reasons is that k-means needs to scan a batch of training data to update the cluster centers at every iteration, which is time-consuming. In the paper, we propose a variance reduced k-means *VRKM*, which outperforms the state-of-the-art method, and obtain $4\times$ speedup for large-scale clustering. The source code is available on **`https://github.com/YaweiZhao/VRKM_sofia-ml`**.

Motivation

K-means clustering needs to pass over a batch of instances in order to update cluster centers at each iteration, which is computationally intensive. The pioneering work in (Bottou, Bengio, and others 1995) proposes a stochastic gradient descent (SGD) variant of k-means in which one instance is randomly sampled to update cluster centers at each iteration. However, this variant usually brings in stochastic noise[1]. Besides, a mini-batch variant of k-means is proposed in (Sculley 2010) to decrease the stochastic noise while increasing the computational cost in calculating the gradient.

Recently, *SVRG* has been developed to decrease the stochastic noise of SGD via variance reduced gradients(Johnson and Zhang 2013). However, we observe that k-means is sharply divergent at iterations when applying SVRG directly. The reason is that the optimization objective of k-means is jointly dominated by cluster centers and instance partitions. Directly applying SVRG to k-means will first search an optimal decreased direction based on the current instance partition. When the instance partition changes, this direction may not be optimal or even not be a decreased one. It is called the drift of cluster centers, which impedes SVRG to be used into k-means. Moreover, SVRG needs to compute a batch gradient at every epoch, which is time-consuming for a large dataset. Therefore, it is valuable to improve k-means by using SVRG efficiently for a large dataset.

[1]The stochastic noise and variance are equivalent in the paper.

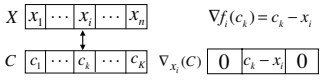

Figure 1: The illustration of the basic data structures and operations where c_k is nearest to x_i.

Symbols and definitions

As illustrated in Figure 1, $X \in \mathbb{R}^{d\times n}$ represents the dataset which contains n instances, and each of instance has d features. X_c denotes the instance set of a cluster $c \in \mathbb{R}^{d\times 1}$, and its size is denoted by v. c represents the center of X_c. x_i with $i \in \{1, 2, ..., n\}$ represents an instance. $\nabla f(c)$ and $\nabla f_i(c)$ represent the batch and stochastic gradients with respect to c, respectively. K is the number of clusters. $C \in \mathbb{R}^{d\times K}$ represents K centers such that $C = [c_1, c_2, ..., c_K]$. $x\{C\}$ represents the nearest center of x. $\nabla_x(C)$ represents the gradient with respect to $x\{C\}$. If x belongs to the cluster c_k, $\nabla_x(C)$ is obtained by using the k-th column of C to subtract x and setting other columns to be 0.

Define 1 *Given a cluster center set $C = [c_1, ..., c_K]$, the nearest cluster center of x is denoted by $x\{C\} \in \mathbb{R}^{d\times 1}$ which is one of the K centers. The index of the center is denoted by $\mathcal{I}_{x\{C\}}$ which is an integer ranging from 1 to K.*

Define 2 *Given a cluster center set $C = [c_1, ..., c_K]$ and an instance x. The gradient with respect to $x\{C\}$ is denoted by $\nabla_x(C) \in \mathbb{R}^{d\times K}$. The $\mathcal{I}_{x\{C\}}$-th column of $\nabla_x(C)$ is $x\{C\} - x$, and the other columns are zeros.*

Variance reduced k-means clustering

The formulation of k-means is $\min_C f(C) = 1/2 \min_C \sum_{i=1}^{K} \sum_{x\in X_{c_i}} \| c_i - x \|^2$. The gradient with respect to a center c_i is $\nabla f(c_i) = \sum_{x\in X_{c_i}} c_i - x$. Furthermore, suppose x_{i_t} is randomly picked at the t-th iteration, the stochastic gradient is $\nabla f_{i_t}(c_i) = c_i - x_{i_t}$. When the cluster center $\tilde{c}$ drifts, it is corrected by the average gradient of the instances:

$$\tilde{c}^{\text{new}} \leftarrow \tilde{c} - \tfrac{1}{v}\nabla f(\tilde{c}) = \tilde{c} - (\tilde{c} - \tfrac{1}{v}\sum_{x\in X_{\tilde{c}}} x) = \bar{X}_{\tilde{c}}.$$

The position correction of $\tilde{c}$ guarantees that it is close to the optimum based on the current instance partition.

Algorithm 1 VRKM: variance reduced k-means

Require: The number of clusters K. The dataset X. The constant learning rate η. The epoch size T.

1: Initialize each $c \in \tilde{C}$ with instances picked from X randomly;
2: **repeat**
3: Update the nearest cluster center for every instance x_i with $i \in \{1, 2, ..., n\}$ according to $\tilde{C}$, and thus obtain the instance partitions $\{X_{c_1}, ..., X_{c_K}\}$;
4: $C_0 = \tilde{C}^{\text{new}} = \left(\bar{X}_1, \bar{X}_2, ..., \bar{X}_K\right)$, and let $X_{\tilde{c}_i^{\text{new}}} = X_{c_i}$ for $1 \le i \le K$;
5: Obtain $x_i\{\tilde{C}^{\text{new}}\}$, $\mathcal{I}_{x_i\{\tilde{C}^{\text{new}}\}}$ and $\nabla_{x_i}(\tilde{C}^{\text{new}})$ for every instance x_i with $1 \le i \le n$ based on the instance partition $\{X_{\tilde{c}_1^{\text{new}}}, ..., X_{\tilde{c}_K^{\text{new}}}\}$;
6: **for** $t = 0, 1, ..., T-1$ **do**
7: Pick an index i_t from $\{1, 2, ..., n\}$ randomly;
8: Find the nearest cluster center from C_t for x_{i_t}, and thus obtain $x_{i_t}\{C_t\}$, $\mathcal{I}_{x_{i_t}\{C_t\}}$ and $\nabla_{x_{i_t}}(C_t)$;
9: **if** $x_{i_t}\{C_t\} \neq x_{i_t}\{\tilde{C}^{\text{new}}\}$ or $\mathcal{I}_{x_{i_t}\{C_t\}} \neq \mathcal{I}_{x_{i_t}\{\tilde{C}^{\text{new}}\}}$ **then**
10: $\gamma_t = \nabla_{x_{i_t}}(C_t) - \nabla_{x_{i_t}}(\tilde{C}^{\text{new}})$;
11: $C_{t+1} = C_t - \eta\gamma_t$;
12: **else** $C_{t+1} = C_t$;
13: $\tilde{C} = C_T$;
14: **until** convergence;
15: **return** $\tilde{C}$;

After the position correction, we obtain $\tilde{C}^{\text{new}}$. Then, the variance reduced gradient is:

$$\begin{aligned}\gamma_t &= \nabla_{x_{i_t}}(C_t) - \nabla_{x_{i_t}}(\tilde{C}^{\text{new}}) + \nabla f(\tilde{C}^{\text{new}}) \\ &= \nabla_{x_{i_t}}(C_t) - \nabla_{x_{i_t}}(\tilde{C}^{\text{new}})\end{aligned}$$

because that $\nabla f(\tilde{c}^{\text{new}}) = \sum_{x \in X_{\tilde{c}^{\text{new}}}} (\tilde{c}^{\text{new}} - x) = \mathbf{0}$ holds for every $\tilde{c}$. As with the increase of iterations, the cluster center c_t and c_{t+1} are close to the optimum c_*. We obtain

$$\eta \lim_{t\to\infty} \mathbb{E}\gamma_t = \lim_{t\to\infty} \mathbb{E}C_{t+1} - \lim_{t\to\infty} \mathbb{E}C_t = C_* - C_* = \mathbf{0}.$$

Here, "$\mathbb{E}$" is the expectation operator on i_t. Thus, $\lim_{t\to\infty}(\mathbb{E}\gamma_t) = 0$ holds when the learning rate η is a constant. Benefiting from this property, a constant learning rate is used to accelerate k-means. It is superior to the traditional methods which use a decaying learning rate. As illustrated in Algorithm 1, every cluster center is corrected by the average of the instances according to Line 4. Lines $10-11$ mean that the variance reduced gradient is used to update the centers with a constant learning rate. VRKM does not need to compute the batch gradient at every epoch according to Line 10, thus yielding a high efficiency.

Empirical studies

In this section, VRKM is compared with the batch k-means denoted by *KM* (Lloyd 1982), the SGD k-means denoted by *SGD-KM* (Bottou, Bengio, and others 1995), the mini-batch k-means denoted by *mini-KM* (Sculley 2010), and the state-of-the-art algorithm denoted by *GB-KM* (Newling and Fleuret 2016). As far as we know, *GB-KM* is the newest variant of k-means which is related to our methods. The dataset is CIFAR-100. The y-axis represents the decrease of the objective function against a baseline. The baseline is obtained by running *KM* for a long given time. The size of a mini-batch is 5000 in *mini-KM*. The epoch size of VRKM is n. The learning rate of VRKM is $5\eta_0$ with $\eta_0 = K/n$.

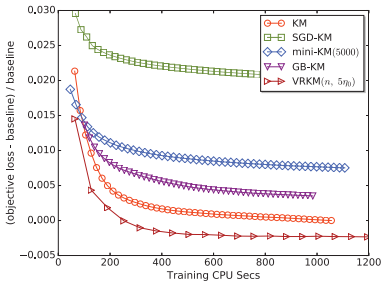

Figure 2: The comparison of the time consumption.

Results. As illustrated in Figure 2, VRKM has an advantage on decreasing the objective loss. Specifically, it yields $4.30\times$ and $3.60\times$ speedups for CIFAR-100, respectively. Additionally, we adopt three metrics: ACC, NMI, and Purity to test the clustering quality. VRKM yields the best clustering solution (ACC: 0.2246, NMI: 0.3742, Purity: 0.2488). Benefiting from the variance reduced gradients and the constant learning rate, VRKM finishes more iterations than the previous methods for the given time, thus yielding the best clustering performance.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Project no. 60970034, 61170287, 61232016, 61672528 and 61671463). We thank for the help provided by Prof. Xinzhong Zhu who is the president of Cixing Research Institute and the Chair Professor of the college of Mathematics, Physics and Information Engineering, Zhejiang Normal University, China. Additionally, we thank Cixing intelligent manufacturing research institute, Cixing textile automation research institute, Ningbo Cixing corporation limited and Ningbo Cixing robotics company limited because of their financial support and application scenarios.

References

Bottou, L.; Bengio, Y.; et al. 1995. Convergence properties of the k-means algorithms. In *Proc. NIPS*, 585–592.

Johnson, R., and Zhang, T. 2013. Accelerating stochastic gradient descent using predictive variance reduction. In *Proc. NIPS*.

Lloyd, S. 1982. Least squares quantization in pcm. *IEEE Transactions on Information Theory* 28(2):129–137.

Newling, J., and Fleuret, F. 2016. Nested Mini-Batch K-Means. In *Proc. NIPS*.

Sculley, D. 2010. Web-scale k-means clustering. In *Proc. WWW*, 1177–1178.