Review on ActiveClean from the Perspective of Convex Optimization

Yang Renchi
yang0461@e.ntu.edu.sg
School of Computer Science and Engineering
Nanyang Technological University

I. Introduction

Iterative data cleaning is defined to be the process of cleaning subsets of data, evaluating preliminary results, and then cleaning more data as necessary. ActiveClean [1] is a framework for the iterative cleaning of a dataset for statistical modelling. Essentially, in ActiveClean, the iterative cleaning problem is mainly formulated into a convex optimization problem, and then employ gradient-based methods to search the optimal solutions. ActiveClean ensures global convergence with a provable rate and also supports an important class of models called convex loss models.

In Section II and Section III, I will explore ActiveClean from the perspective of convex optimization and figure out how convex optimization techniques are applied to solve the problems. Section IV summarizes my overview on the paper.

II. Optimal Model Update

A. Problem Definition

Definition 1. *Given a set of data records R ($|R| = n$), for each record $r \in R$, we can map record r to a feature vector x ($|x| = t$) and a label y via a featurizer $F(\cdot)$.*

$$F(r_i) = (x_i, y_i) \tag{1}$$

Find a vector of weights θ ($|\theta| = t$) to minimize the loss function $\phi : \mathbb{R}^t \to \mathbb{R}$, thereby achieving minimal prediction error.

$$\min_{\theta} \sum_{i=1}^{n} \phi(\theta^{\mathsf{T}} x_i, y_i) \tag{2}$$

Let loss function ϕ be a convex loss function (e.g., square loss, hinge loss, logistic loss) in θ, thus the original data cleaning problem is converted into a convex optimization problem, namely, once we find a local optima, we obtain the global optima. Problem (2) is dubbed as *Model Update Problem*.

B. Solutions

Consider problem (2), which is evidently an unconstrained minimization problem, intuitively gradient-based methods can be applied to solve this problem. Hence, the model can be updated via:

$$\theta^{(k+1)} = \theta^{(k)} - \gamma^{(k)} \cdot D^{(k)} \cdot \nabla\phi(\theta^{(k)}) \tag{3}$$

Where $\gamma^{(k)}$ is the step size, $\nabla\phi(\theta^{(k)})$ is the gradient, and $D^{(k)} \cdot \nabla\phi(\theta^{(k)})$ is the descent direction in k-th iteration. The authors set $D^{(k)} = I$, i.e., *Steepest Descent Method.*

$$\theta^{(k+1)} = \theta^{(k)} - \gamma^{(k)} \cdot \nabla\phi(\theta^{(k)}) \tag{4}$$

Gradient method moves the model downhill to reach the bottom. Especially at the optimal point, $\nabla\phi(\theta^{(k)}) = 0$.

The main challenge here is that the true gradient $\nabla\phi(\theta)$ should be obtained over all clean data, which is not available at present.

$$\nabla\phi(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla\phi_i(\theta^{\mathsf{T}} x_i^{(c)}, y_i^{(c)}) \tag{5}$$

The authors propose that we can approximate the true gradient $\nabla\phi(\theta)$ from already cleaned data R_{clean} ($|R_{clean}| = c$) and a sample of dirty data S ($|S| = b$), which is newly cleaned.

$$g(\theta) = \frac{c}{n} \cdot g_c(\theta) + \frac{n-c}{n} \cdot g_s(\theta) \tag{6}$$

$$g_c(\theta) = \frac{1}{c} \sum_{r_i \in R_{clean}} \nabla\phi_i(\theta^{\mathsf{T}} x_i^{(c)}, y_i^{(c)}) \tag{7}$$

$$g_s(\theta) = \frac{1}{b} \sum_{r_i \in S} \nabla\phi_i(\theta^{\mathsf{T}} x_i^{(c)}, y_i^{(c)}) \tag{8}$$

$$R_{clean}^{(k+1)} = R_{clean}^{(k)} + S^{(k)},\ R_{dirty}^{(k+1)} = R_{dirty}^{(k)} - S^{(k)}$$

Therefore, Equation (4) can be approximated by:

$$\theta^{(k+1)} = \theta^{(k)} - \gamma^{(k)} \cdot g(\theta^{(k)}) \tag{9}$$

$\theta^{(0)}$ is initialized over all dirty data R_{dirty}. In the paper, the authors use inexact line search, step size $\gamma^{(k)}$ is set to be diminishing: $\frac{\gamma^{(0)}}{kb}$, $\gamma^{(0)} = \gamma$.

This solution can be seen as a variant of *Mini-batch Stochastic Gradient Descent* [2]. One key difference is that ActiveClean applies a full gradient step on the already cleaned data and averages it with the stochastic gradient step calculated from a sample of dirty data. I will show this modification will not affect the convergence in the following section.

C. Convergence Analysis

The convergence proof of *Model Update* is ommitted in original paper, in this section, I will present the outline that how ActiveClean guarantees the convergence based on my understanding of the algorithm.

Proof sketch: Assume that the loss function $\phi(\theta)$ is strongly convex, then we have:

$$\phi(\theta) - \phi(\theta') - \nabla\phi(\theta')^\intercal(\theta - \theta') \leq \frac{M}{2}\|\theta - \theta'\|_2^2 \tag{10}$$

Let $G(\theta) = \nabla\phi(\theta)$, which denotes the gradient computed over full cleaned data, $g_i(\theta) = \nabla\phi_i(\theta^\intercal x_i^{(c)}, y_i^{(c)})$, and $\psi_i = g_i(\theta) - G(\theta))$. Moreover, denote the variance $\mathbb{E}[\|\psi_i\|^2]$ by σ^2. Then for gradient $g(\theta)$ in each iteration, we have:

$$\begin{aligned}
&\mathbb{E}[\|g(\theta) - G(\theta)\|_2^2] \\
&= \mathbb{E}[\|\frac{c}{n}[g_C(\theta) - G(\theta)] + \frac{n-c}{n}[g_S(\theta) - G(\theta)]\|_2^2] \\
&\leq \mathbb{E}[\|\frac{1}{b}\sum_{r_i \in R_{clean} \cup S}(g_i(\theta) - G(\theta))\|_2^2] \\
&= \frac{1}{b^2}\mathbb{E}[\sum_{r_i, r_j \in R_{clean} \cup S}^{i \neq j} \psi_i^\intercal \psi_j] + \frac{1}{b}\mathbb{E}[\|\psi_i\|^2] \\
&= \frac{(c+b)(c+b-1)}{b^2 n(n-1)} \sum_{r_i, r_j \in R_{clean}}^{i=j} \psi_i^\intercal \psi_j + \frac{n-b-c}{b(n-1)}\mathbb{E}[\|\psi_i\|^2] \\
&= \frac{n-b-c}{b(n-1)}\mathbb{E}[\|\psi_i\|^2] \\
&\leq \frac{\mathbb{E}[\|\psi_i\|^2]}{b} = \frac{\sigma^2}{b}
\end{aligned} \tag{11}$$

With inequality (10) and (11), according to [2], it can be derived that the convergence rate is bounded by $O(\frac{\sigma^2}{\sqrt{bK}})$, where K is the maximal iteration times. ■

III. Optimal Sampling Problem

Since the convergence rate of *Model Update* is bounded by $O(\frac{\sigma^2}{\sqrt{bK}})$, a remaining problem is how to take sample S so that the variance is minimized.The below content is the detailed analysis on this problem from my understanding, which is also not fully included in original paper.

A. Problem Definition

Definition 2. *Given a set of gradients $\mathcal{G} = \{g_1, g_2, ..., g_m\}$ of dirty data R_{dirty}, find a sampling probability distribution p on $\mathcal{G}$, such that over all samples S of size b it minimizes the variance of the mean $\bar{g}$ of S:*

$$\min Var[\bar{g}] = \min \mathbb{E}[\bar{g}^2] - \mathbb{E}[\bar{g}]^2 = \min \mathbb{E}[\bar{g}^2] - G(\theta)^2$$

$$\min \mathbb{E}[\bar{g}^2] = \min_p \frac{1}{m^2}\sum_{i=1}^m \frac{g_i^2}{p_i}$$

$$\Longrightarrow$$

$$\min_p \sum_{i=1}^m \frac{g_i^2}{p_i}$$

$$\text{subject to: } p > 0, \ \sum_{i=1}^m p_i = 1 \tag{12}$$

B. Solution

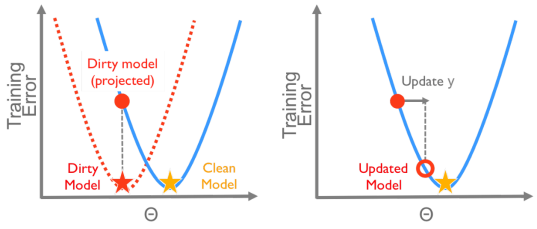
Obviously, this is a problem with simplex constraint. The Lagrangian of problem (12) is $L(p, \lambda, \nu) = (\sum_{i=1}^m g_i^2/p_i - \lambda^\intercal p - \nu^\intercal p) + \nu$. Since $\lambda \geq 0$, Lagrangian dual function $g(\lambda, \nu) = \nu$ if $-g_i^2/p_i^2 - \nu \geq 0 \ \forall i \in [1, m]$, otherwise $g(\lambda, \nu)$ would be unbounded below. Thus, the Lagrangian dual problem of primal problem is:

$$\begin{aligned}
&\max \nu \\
&\text{subject to: } -\frac{g_i^2}{p_i^2} \geq \nu \ \forall i \in [1, m]
\end{aligned} \tag{13}$$

According to complementary slackness, optimal value $p_i^* > 0 \Longrightarrow -g_i^2/p_i^{*2} = \nu \Longrightarrow \nu = -(\sum_{i=1}^m g_i)^2$. So we have $p_i^* = |g_i|/\sum_{i=1}^m |g_i|$, which means the probability distribution with optimal variance is sampling proportionally to the gradients.

IV. Conclusion

In summary, ActiveClean mainly reduce original data cleaning problem to two convex optimization problems: *Model Update* (2) and *Optimal Sampling* (13).



(A) Before ActiveClean Update (B) After ActiveClean Update

Fig. 1: Dirty model vs Clean Model

For *Model Update*, the key insight of ActiveClean is that it introduces convex loss models to measure the performance of data cleaning, which can be trained and cleaned simultaneously. As shown in Fig. 1, ActiveClean moves the dirty model (red dotted line) towards clean model (blue line) by iteratively cleaning data, and at the same time, decides the direction to achieve optimal θ (yellow star) by taking gradient over cleaned data and a sample of dirty data with convergence guarantee.

Regarding *Optimal Sampling*, the crucial idea is using Lagrangian multipliers to transform primal problem into a Lagrangian dual problem, namely a non-convex problem into a convex problem.

These two applications of convex optimization techniques in the paper show me how a problem can be modeled or transformed into a convex optimization problem step by step which we can employ convex optimization methods to solve. This is definitely of great use for my future study and research.

References

[1] Krishnan, Sanjay, et al. "ActiveClean: interactive data cleaning for statistical modeling." *Proceedings of the VLDB Endowment* 9.12 (2016): 948-959.

[2] Dekel, Ofer, et al. "Optimal distributed online prediction using mini-batches." Journal of Machine Learning Research 13.Jan (2012): 165-202.