Supplemental material for "Multi-class Boosting for Early Classification of Sequences"

Katsuhiko Ishiguro
ishiguro@cslab.kecl.ntt.co.jp
Hiroshi Sawada
sawada@cslab.kecl.ntt.co.jp
Hitoshi Sakano
keen@cslab.kecl.ntt.co.jp

NTT Communication Science Laboratories
NTT Corporation
Kyoto, 610-0237, Japan

Abstract

In this material, we present the derivation of Earlyboost [3] and the proposed Earlyboost.MH in the standard exponential loss approach [1].

1 Notations

We follow the original notations declared in the paper, but we describe the notations at the head of each section for readers help.

2 Earlyboost

First, we derive the update equations of Earlyboost [3], which first appears in the literature for the best of our knowledge.

Let us denote that i-th sequence $\boldsymbol{x}_i$ has a number of time frame elements $x_{i,t}$: i.e. $\boldsymbol{x}_i = \{x_{i,t} \in \mathbb{R}^d\}, t = 1,2,\ldots,T$. The number of sequences is N: thus $i \in \{1,2,\ldots,N\}$. T is a length of time sequences, and $t \in \{1,2,\ldots,T\}$ is the time index. $y_i \in \{1,-1\}$ is a class label attached to $\boldsymbol{x}_i$. The training data set is denoted by $\mathscr{D} = \{\boldsymbol{x}_i, y_i\}$.

First let us define the strong classifier H:

$$H(\boldsymbol{x}_i) = \sum_{t=1}^{T} \alpha_t h_t(x_{i,t}) \tag{1}$$

$h_t : \mathbb{R} \to \{1,-1\}$ is the t-th weak classifier, which only accepts $\{x_{i,t}\}, i = 1,2,\ldots,N$. We assume the importance weight α_t is strictly positive. Also we consider the t-subset of the above (final) classifier H as follows:

$$H^t(\boldsymbol{x}_i) = \operatorname{sign}\left(\sum_{s=1}^{t} \alpha_s h_s(x_{i,s})\right). \tag{2}$$

Thus $H = H^T$.

Then the empirical loss of the Earlyboost is defined over H^t as follows:

$$J(H^t) = \sum_i \left(\exp\left(-y_i H^t(\boldsymbol{x}_i)\right)\right) = \sum_i \exp\left(-y_i \sum_{s=1}^{t} \alpha_s h_s(x_{i,s})\right). \tag{3}$$

In the learning phase, we iteratively augment H^t from $H^0 = \emptyset$ by adding the weak classifier h_t and the importance weight α_t one by one.

Now let us assume that we have $t-1$ learning process iterations. Thus $H^{t-1} = \sum_{s=1}^{t-1} \alpha_s h_s$ is already given. Given H^{t-1}, we would like to obtain the optimal α_t and h_t to minimize the loss Eq. (3).

For that purpose, we write down the loss of expanded strong classifier.

$$J(H^{t-1} + \alpha_t h_t) = \sum_{i=1}^{N} \left[\exp\left(-y_i (H^{t-1}(\boldsymbol{x}_i) + \alpha_t h_t(x_{i,t}))\right)\right] \tag{4}$$

2.1 Deriving h_t

Performing Taylor expansion of Eq. (4), then we get the following:

$$J(H^{t-1} + \alpha_t h_t) \approx \sum_{i=1}^{N} \left[\exp\left(-y_i H^{t-1}(\boldsymbol{x}_i)\right)\left(1 - y_i \alpha_t h_t(x_{i,t}) + \frac{\alpha_t^2}{2}\right)\right]. \tag{5}$$

Since we assumed $\alpha_t > 0$ for all t, we can rewrite Eq. (5) as follows:

$$\begin{aligned}
\hat{h}_t &= \arg\min_{h_t} J(H^{t-1} + \alpha_t h_t) \\
&= \arg\max_{h} \sum_{i=1}^{N} \left[\exp\left(-y_i H^{t-1}(\boldsymbol{x}_i)\right) y_i \alpha_t h_t(x_{i,t})\right] \\
&\Leftrightarrow \arg\max_{h} \sum_{i=1}^{N} D_t(i) y_i h_t(x_{i,t}).
\end{aligned} \tag{6}$$

Please note that $D_t(i)$ is a positive constant:

$$D_t(i) = \exp\left(-y_i H^{t-1}(\boldsymbol{x}_i)\right) > 0. \tag{7}$$

Therefore, an optimal $\hat{h}_t$ is available by solving Eq. (6):

$$\hat{h}_t(x_{i,t}) = \begin{cases} 1 & P(y_i = 1 | x_{i,t}) > P(y_i = -1 | x_{i,t}) \\ -1 & \text{otherwise} \end{cases} \tag{8}$$

Since h_t is a weak proxy for the true conditional probability $P(y|x_t)$, we can mimic an optimal $\hat{h}_t$ effectively by the following equation.

$$\hat{h}_t = \arg\min_{h_t} \sum_{i: y_i \neq h_t(x_{i,t})} D_t(i) \tag{9}$$

2.2 Deriving α_t

Given the optimal h_t, next we optimize the importance weight α_t. From Eq. (4), we obtain the following:

$$\begin{aligned} J(H^{t-1}+\alpha_t\hat{h}_t) &= \sum_i \left[\exp\left(-y_i(H^{t-1}(\boldsymbol{x}_i)+\alpha_t\hat{h}_t(x_{i,t}))\right)\right] \\ &= \sum_i D_t(i)\exp\left(-y_i\alpha_t\hat{h}_t(x_{i,t})\right) \end{aligned} \tag{10}$$

We split the sequence indices into two sets: $i^+ = \{i : y_i = \hat{h}_t(x_{i,t})\}$ and $i^- = \{i : y_i \neq \hat{h}_t(x_{i,t})\}$. Using these notations, we rewrite Eq. (10) as follows:

$$J(H^{t-1}+\alpha_t\hat{h}_t) \propto \sum_{i\in i^+} D_t(i)\exp(-\alpha_t) + \sum_{i\in i^-} D_t(i)\exp(\alpha_t) \tag{11}$$

Taking the derivative of Eq. (11) with respect to α_t, we obtain the solution of α_t. Please note that ϵ_t in the paper is equivalent to $\sum_{i\in i^-} D_t(i)$.

$$\begin{aligned} \frac{\partial J}{\partial \alpha_t} &= -\sum_{i\in i^+} D_t(i)\exp(-\alpha_t) + \sum_{i\in i^-} D_t(i)\exp(\alpha_t) = 0 \\ \Leftrightarrow \alpha_t &= \frac{1}{2}\log\left(\frac{1-\sum_{i\in i^-} D_t(i)}{\sum_{i\in i^-} D_t(i)}\right). \end{aligned} \tag{12}$$

2.3 Deriving $D_{m,i}$

Using Eq. (4) and the definition of $D_t(i)$, we obtain the weight propagation equation naturally.

$$\begin{aligned} \exp\left(-y_i H^t\right) &= \exp\left(-y_i(H^{t-1}+\alpha_t h_t)\right) \\ &= \exp\left(-y_i H^{t-1}\right)\exp(-y_i\alpha_t h_t(x_{i,t})) \\ \Leftrightarrow D_{t+1}(i) &\propto D_t(i)\exp(-y_i\alpha_t h_t(x_{i,t})) \end{aligned} \tag{13}$$

3 Earlyboost.MH

Then, we derive the update equations of the proposed Earlyboost.MH.

We have a set of training data $\mathscr{D} = \{\boldsymbol{x}_i, y_i\}, i = 1, 2, \ldots, N$. The i-th sequence is denoted as $\boldsymbol{x}_i$, which has T frame elements $x_{i,t} \in \mathbb{R}^d\}$. Because of multi-class classification, class labels range in K values: $y_i \in \{1, 2, \ldots, K\}$ which makes contrast to Adaboost and Earlyboost.

The strong classifier H is defined as follows:

$$H(\boldsymbol{x}_i) = \sum_{t=1}^{T} \alpha_t h_t(x_{i,t}) \tag{14}$$

H returns K-dimensional response given a sequence input. $h_t : R^d \to \{1, -1\}^K$ is the t-th weak classifier which only accepts the observation in the time frame t. $\alpha_t > 0$ is its importance weight.

More conveniently, we assume H consists of K binary classifiers H_k, namely $H = \{H_k\}$. We denote the k-th one vs. all strong classifier as H_k, and its t-th weak classifier as $h_{t,k}$:

$\mathbb{R}^d \to \{1,-1\}$, respectively. A weak classifier $h_{t,k}(x) : \mathbb{R}^d \to \{1,-1\}$ only accepts the samples on the t-th time frames. $h_{t,k}(x)$ returns 1 if x belongs to class k, and returns -1 otherwise. We also define $g_k(y) : \{1,2,\ldots,K\} \to \{1,-1\}$ returns 1 if $y = k$, and returns -1 otherwise.

The definition of H_k ans its t-subset H_k^t is described as follows:

$$H_k(\boldsymbol{x}_i) = \sum_{t=1}^{T} \alpha_t h_{t,k}(x_{i,t}), \tag{15}$$

$$H_k^t(\boldsymbol{x}_i) = \sum_{s=1}^{t} \alpha_s h_s(x_{i,s}). \tag{16}$$

Then we define the following loss function:

$$J(H^t) = \sum_{i=1}^{N}\sum_{k=1}^{K}\left(\exp\left(-g_k(y_i)H_k^t(\boldsymbol{x}_i)\right)\right) = \sum_{i=1}^{N}\sum_{k=1}^{K}\exp\left(-g_k(y_i)\sum_{s=1}^{t}\alpha_s h_{s,k}(x_{i,s})\right). \tag{17}$$

Construction of above loss function is similar to Adaboost.MH [1, 2].

Then, we compute an optimal $h_{t,k}$ and α_t given H_k^{t-1} which consists of $t-1$ elements weak classifiers.

3.1 Deriving $h_{t,k}$

Following the derivation of h_t in Earlyboost, we perform Taylor-expansion of the loss of $H^{t-1} + \alpha_t h_t$.

$$\begin{aligned} J(H^{t-1} + \alpha_t h_t) &= \sum_{i=1}^{N}\sum_{k=1}^{K}\left[\exp\left(-g_k(y_i)(H_k^{t-1}(\boldsymbol{x}_i) + \alpha_t h_{t,k}(x_{i,t}))\right)\right] \\ &\approx \sum_{i=1}^{N}\sum_{k=1}^{K}\left[\exp\left(-g_k(y_i)H_k^{t-1}(\boldsymbol{x}_i)\right)\left(1 - g_k(y_i)\alpha_t h_{t,k}(x_{i,t}) + \frac{\alpha_t^2}{2}\right)\right] \end{aligned} \tag{18}$$

Assuming $\alpha_t > 0$, we obtain the following equations:

$$\begin{aligned} \hat{h}_{t,k} &= \arg\min_{h_{t,k}} J(H^{t-1} + \alpha_t h_t) \\ &\Leftrightarrow \arg\max_{h_{t,k}} \sum_{i=1}^{N} D_t(i,k) y_{i,k} h_{t,k}(x_{i,t}) \end{aligned} \tag{19}$$

$$D_t(i,k) = \exp\left(-y_{i,k}H_k^{t-1}(\boldsymbol{x}_i)\right) \tag{20}$$

Note that this is equivalent to the equation in the paper, defined by $r_{t,k}$.

3.2 Deriving α_t

From the definition,

$$\begin{aligned} J(H^{t-1} + \alpha_t h_t) &= \sum_{i}\sum_{k=1}^{K}\left[\exp\left(-g_k(y_i)H_k^{t-1}(\boldsymbol{x}_i) - g_k(y_i)\alpha_t \hat{h}_{t,k}(x_{i,t})\right)\right] \\ &= \sum_{k}\sum_{i} D_t(i,k)\exp\left(-g_k(y_i)\alpha_t \hat{h}_{t,k}(x_{i,t})\right) \end{aligned} \tag{21}$$

As in the case of Earlyboost, we split the sequence indices in two sets, for each class k. Let us define $i^{k+} = \{i : g_k(y_i) = \hat{h}_{t,k}(x_{i,t})\}$ and $i^{k-} = \{i : g_k(y_i) \neq \hat{h}_{t,k}(x_{i,t})\}$. Using these notations, we rewrite Eq. (21) as follows:

$$J(H^{t-1} + \alpha_t h_t) = \sum_k \sum_{i \in i^{k+}} D_t(i,k) \exp(-\alpha_t) + \sum_k \sum_{i \in i^{k-}} D_t(i,k) \exp(\alpha_t) \tag{22}$$

Using

$$r_{t,k} = \sum_{i \in i^{k+}} D_t(i,k) - \sum_{i \in i^{k-}} D_t(i,k), \tag{23}$$

and

$$1 = \sum_k \sum_{i \in i^{k+}} D_t(i,k) + \sum_k \sum_{i \in i^{k-}} D_t(i,k), \tag{24}$$

we can derive the equation for an optimal α_t.

$$\begin{aligned}
\frac{\partial J}{\partial \alpha_t} &= -\sum_k \sum_{i \in i^{k+}} D_t(i,k) \exp(-\alpha_t) + \sum_k \sum_{i \in i^{k-}} D_t(i,k) \exp(\alpha_t) = 0 \\
\Leftrightarrow \exp(2\alpha_t) &= \frac{\sum_k \sum_{i \in i^{k+}} D_t(i,k)}{\sum_k \sum_{i \in i^{k-}} D_t(i,k)} = \frac{1 + \sum_k r_{t,k}}{1 - \sum_k r_{t,k}} \\
\alpha_t &= \frac{1}{2} \log\left(\frac{1 + \sum_k r_{t,k}}{1 - \sum_k r_{t,k}}\right).
\end{aligned} \tag{25}$$

3.3 Deriving $D_t(i,k)$

The update equation for $D_t(i,k)$ is easy to derive:

$$\begin{aligned}
\exp\left(-g_k(y_i) H_k^t(\boldsymbol{x}_i)\right) &= \exp\left(-g_k(y_i)(H_k^{t-1}(\boldsymbol{x}_i) + \alpha_t h_{t,k}(x_{i,t})\right) \\
&= \exp\left(-g_k(y_i H_k^{t-1}(\boldsymbol{x}_i)\right) \exp(-g_k(y_i)\alpha_t h_{t,k}(x_{i,t})) \\
\Leftrightarrow D_{t+1}(i,k) &\propto D_t(i,k) \exp(-g_k(y_i)\alpha_t h_{t,k}(x_{i,t})).
\end{aligned} \tag{26}$$

References

[1] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting (with discussion). *Annals of Statistics*, 28(2):337–407, 2000.

[2] R. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999.

[3] S. Uchida and K. Amamoto. Early recognition of sequential patterns by classifier combination. In *Proceedings of the 19th International Conference on Pattern Recognition (ICPR)*, 2008.