Cost-sensitive Blinex SVM Model

Abstract

Index Terms

I. The Cost-sensitive Blinex SVM Model

Consider a supervised learning problem defined on a data set of l labeled augmented samples $\{(\mathcal{X},\mathcal{Y})\} = \{(x_i,y_i)\}_{i=1}^{l}$ $=\{(\mathrm{x}_i;1),y_i)\}_{i=1}^{l}$, where $\mathcal{X}$ is the feature space. Table I summarizes the major notations used in this paper.

Formally, Cost-sensitive Blinex SVM Model in the linear and nonlinear cases can be built as follows.

II. Linear case

Cost-sensitive Blinex SVM Model in the linear case can be built as follows:

$$\begin{aligned} \min_{w,\xi} \quad & \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{l}[1-\frac{1}{1+b(exp(ay_i\xi_i)-ay_i\xi_i-1)}] \\ \text{s.t.} \quad & y_i(w\cdot x_i)) \geqslant 1-\xi_i,\ \xi_i \geqslant 0,\ i=1,2,\cdots,l. \end{aligned} \tag{1}$$

A. Solution for BLINEX-SVM in Linear Case

The primal problem (1) can be reformulated as follows:

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{l} L(\xi_i), \tag{2}$$

where

$$L(\xi_i) = 1 - \frac{1}{1+b(\exp(ay_i\xi_i)-ay_i\xi_i-1)}, \tag{3}$$

and

$$\xi_i = \max(0, 1-y_i(w\cdot x_i)) = I[y_i(w\cdot x_i)<1](1-y_i(w\cdot x_i)). \tag{4}$$

Let

$$f(w) = \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{l} L(\xi_i), \tag{5}$$

then we have

$$\begin{aligned} \frac{\partial f}{\partial w} &= w + c\sum_{i=1}^{l}\frac{\partial L}{\partial \xi_i}\cdot\frac{\partial \xi_i}{\partial w} \\ &= w + \sum_{i=1}^{l}\frac{abc[1-\exp(ay_i\xi_i)]}{[1+b(\exp(ay_i\xi_i)-ay_i\xi_i-1)]^2}\cdot I[y_i(w\cdot x_i)<1]\cdot x_i. \end{aligned} \tag{6}$$

At the t-th iteration with a randomly chosen data point (x_{i_t}, y_{i_t}), $i\in\{1,2,\cdots,l\}$, the objective function $f(w)$ can be formulated as

$$f_t(w_t) = \tfrac{1}{2}\|w_t\|^2 + cL_t(\xi_{i_t}). \tag{7}$$

The gradient of (7) with respect to w_t can be computed as

$$\begin{aligned} \frac{\partial f_t}{\partial w_t} &= w_t + \frac{\partial L_t}{\partial \xi_{i_t}}\cdot\frac{\partial \xi_{i_t}}{\partial w_t} \\ &= w_t + \frac{abc[1-\exp(ay_{i_t}\xi_{i_t})]}{[1+b(\exp(ay_{i_t}\xi_{i_t})-ay_{i_t}\xi_{i_t}-1)]^2}\cdot I[y_{i_t}(w_t\cdot x_{i_t})<1]\cdot x_{i_t}. \end{aligned} \tag{8}$$

TABLE I
LIST OF NOTATIONS

Notation	Description
(x_i, y_i)	ith training point
l	number of training points
$(x_i \cdot x_j)$	inner product between x_i and x_j as $x_i^\top x_j$
w	weight vector
ϕ	mappings from inputs to high-dimensional feature space
$\kappa(x_i, x_j)$	kernel function $(\phi(x_i)\cdot\phi(x_j))$
$a,\ b$	parameters in Blinex loss function
c	penalty parameter
$\|\cdot\|, \|\cdot\|$	1-norm, 2-norm
$I[g]$	a indicator function that return 1 if g is ture, else 0
$\top$	vector with superscript $^\top$ denote transpose

At the $(t+1)$-th iteration, we can update w_{t+1} in the following way

$$\begin{aligned} w_{t+1} &= w_t - \eta_t \frac{\partial f_t}{\partial w_t} \\ &= w_t - \eta_t w_t - \eta_t \cdot \frac{abc[1-\exp(ay_{i_t}\xi_{i_t})]}{[1+b(\exp(ay_{i_t}\xi_{i_t}) - ay_{i_t}\xi_{i_t} - 1)]^2} \cdot I[y_{i_t}(w_t \cdot x_{i_t}) < 1] \cdot x_{i_t} \\ &= (1-\eta_t)w_t + \eta_t \cdot \frac{abc[\exp(ay_{i_t}\xi_{i_t}) - 1]}{[1+b(\exp(ay_{i_t}\xi_{i_t}) - ay_{i_t}\xi_{i_t} - 1)]^2} \cdot I[y_{i_t}(w_t \cdot x_{i_t}) < 1] \cdot x_{i_t}. \end{aligned} \tag{9}$$

Once the optimal w^* of Eq. (1) is achieved, the following formulation is used to predict the label of a new sample (x, y)

$$f{=}sign(f(x)){=}sign(w^{*\top} x). \tag{10}$$

Algorithm 1 NAG Algorithm for the BLINEX-SVM in linear case

Input:

Data set: $\{(x_i, y_i)\}_{i=1}^l$={((x$_i$;1),y_i)}$_{i=1}^l$, $y_i \in \{-1, 1\}$;

Initialize: penalty parameter c, Blinex loss parameters $a,\ b$, maximum iteration number T, error tolerance ϵ, mini-batch size m, learning rate decay factor k, momentum parameter γ; weight w_0, learning rate η_0, velocity v_0;

Output:

Decision function as in Eq. (10);

1: Randomly choosing m samples: $\{(x_i, y_i)\}_{i=1}^m$;
2: Temporary update: $\tilde{w}_t = w_t + \gamma v_t$
3: Computing $grad(\tilde{w}_t) = \frac{\tilde{w}_t}{l} + \frac{abc}{m}\sum_{i=1}^m \frac{[1-\exp(ay_i\xi_i)]}{[1+b(\exp(ay_i\xi_i) - ay_i\xi_i - 1)]^2} I[y_i(\tilde{w}_t \cdot x_i) < 1]x_i$, where $\xi_i = 1 - y_i(\tilde{w}_t \cdot x_i)$;
4: Updating velocity $v_t = \gamma v_t - \eta_t grad(\tilde{w}_t)$;
5: Updating model parameter: $w_{t+1} = w_t + v_t$;
6: Updating learning rate: $\eta_{t+1} = \eta_t e^{-kt}$;
7: Updating current interation number: $t = t + 1$;
8: Repeating above steps until convergence or up to the maximum iteration number T;
9: Return w^*.

III. NONLINEAR CASE

Cost-sensitive Blinex SVM Model in the nonlinear case can be built as follows:

$$\begin{aligned} \min_{w,\xi} \quad & \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{l}[1 - \frac{1}{1+b(exp(ay_i\xi_i) - ay_i\xi_i - 1)}] \\ \text{s.t.} \quad & y_i(w\cdot\phi(x_i)) \geqslant 1 - \xi_i,\ \xi_i \geqslant 0,\ i = 1, 2, \cdots, l. \end{aligned} \tag{11}$$

A. Solution for BLINEX-SVM in Nonlinear Case

Similar to the linear case, updating of w_{t+1} in the $(t+1)$-th iteration can be calculated as

$$w_{t+1} = (1-\eta_t)w_t + \eta_t \cdot \frac{abc[\exp(ay_{i_t}\xi_{i_t}) - 1]}{[1+b(\exp(ay_{i_t}\xi_{i_t}) - ay_{i_t}\xi_{i_t} - 1)]^2} \cdot I(y_{i_t}(w_t \cdot \phi(x_{i_t})) < 1) \cdot \phi(x_{i_t}). \tag{12}$$

Let

$$v_t = \frac{[\exp(ay_{i_t}\xi_{i_t}) - 1]}{[1+b(\exp(ay_{i_t}\xi_{i_t}) - ay_{i_t}\xi_{i_t} - 1)]^2} \cdot I(y_{i_t}(w_t \cdot \phi(x_{i_t})) < 1) \cdot \phi(x_{i_t}), \tag{13}$$

and

$$\eta_t = \frac{1}{t}, \tag{14}$$

then

$$w_{t+1} = \frac{t-1}{t} w_t + \frac{1}{t} abcv_t. \tag{15}$$

According to (15), we can derive recursively and obtain

$$w_{t+1} = \frac{abc}{t} \sum_{s=1}^{t} v_s. \tag{16}$$

Thus, we have

$$w_{t+1} = \frac{abc}{t} \sum_{s=1}^{t} \frac{[\exp(ay_{i_s}\xi_{i_s}) - 1]}{[1 + b(\exp(ay_{i_s}\xi_{i_s}) - ay_{i_s}\xi_{i_s} - 1)]^2} \cdot I(y_{i_s}(w_t \cdot \phi(x_{i_s})) < 1) \cdot \phi(x_{i_s}). \tag{17}$$

For each t, let $\alpha_{t+1} \in R^l$ be the vector such that $\alpha_{t+1}[j]$ counts how many times the sample (x_j, y_j) has been selected so far and accumulates the loss on them,

$$\alpha_{t+1}[j] = \{\sum_r \frac{[\exp(ay_j\xi_{i_r}) - 1]}{[1 + b(\exp(ay_j\xi_{i_r}) - ay_j\xi_{i_r} - 1)]^2} : r \le t \wedge i_r = j \wedge y_j(w_r \cdot \phi\,(x_j) < 1)\}, \tag{18}$$

where

$$\begin{aligned} \xi_{i_r} &= \max\{0, 1 - y_j(w_r \cdot \phi(x_j)) \\ &= I[y_j(w_r \cdot \phi(x_j)) < 1][1 - y_j(w_r \cdot \phi(x_j))] \\ &= 1 - y_j(w_r \cdot \phi(x_j)). \end{aligned} \tag{19}$$

We can represent w_{t+1} using α_{t+1} as follows

$$w_{t+1} = \frac{abc}{t} \sum_{j=1}^{l} \alpha_{t+1}[j]\phi(x_j). \tag{20}$$

The following formulation is used to predict the label of a new sample (x, y)

$$f(x) = \frac{abc}{t} \sum_{j=1}^{l} \alpha_{t+1}[j]\kappa(x_j, x). \tag{21}$$

Algorithm 2 SGD Algorithm for the BLINEX-SVM in nonlinear case

---

Input:

Data set: $\{(x_i, y_i)\}_{i=1}^{l}$=$\{((\mathrm{x}_i;1),y_i)\}_{i=1}^{l}, y_i \in \{-1, 1\}$;

Initialize: penalty parameter c, Blinex loss parameters a and b, parameter in the kernel function κ, maximum iteration number T, error tolerance ϵ, $\alpha_1 = 0$, iteration number $t = 1$;

Output:

Decision function as in Eq. (21);

1: Randomly choose one sample (x_{i_t}, y_{i_t});

2: For all $j \neq i_t$, set $\alpha_{t+1}[j] = \alpha_t[j]$;

3: Compute

$$\begin{aligned} \xi_{i_t} &= 1 - y_{i_t}(w_t \cdot \phi(x_{i_t})) \\ &= 1 - y_{i_t} \cdot \frac{abc}{t-1} \sum_{j=1}^{l} \alpha_t[j](\phi(x_j) \cdot \phi(x_{i_t})) \\ &= 1 - y_{i_t} \cdot \frac{abc}{t-1} \sum_{j=1}^{l} \alpha_t[j]\kappa(x_{i_t}, x_j); \end{aligned} \tag{22}$$

4: If $\xi_{i_t} > 0$,
then $\alpha_{t+1}[j] = \alpha_t[j] + \frac{\exp(ay_{i_t}\xi_{i_t})-1}{[1+b(\exp(ay_{i_t}\xi_{i_t})-ay_{i_t}\xi_{i_t}-1)]^2}$;
Else
$\alpha_{t+1}[j] = \alpha_t[j]$;

5: Update current iteration number $t = t + 1$;

6: Repeating above steps until convergence or up to the maximum iteration number T;

7: Return $\alpha_{t+1}[j],\ j = 1, 2, \cdots, l$.
