Semi-Supervised Learning in the Imbalanced Visual World

Tinghui Zhou, Joseph Lim

1 Model without $\boldsymbol{\beta}$ regularization

$$\min_{\mathbf{w},\boldsymbol{\beta}} \ \frac{1}{2}\|\mathbf{w}\|^2 + C_l \sum_{i \in I_l} \ell(\mathbf{w}, \mathbf{x}_i, y_i) + C_{u+} \sum_{i \in I_u} \beta_i \ell(\mathbf{w}, \mathbf{x}_i, 1) + C_{u-} \sum_{i \in I_u} (1-\beta_i)\ell(\mathbf{w}, \mathbf{x}_i, -1), \quad (1)$$

where $\beta_i \in [0,1]$ indicates how much the i-th example is borrowed as a positive ($\beta_i = 1$ means borrowing as a full positive, and $\beta_i = 0$ means borrowing as a full negative).

We utilize an iterative update scheme to optimize Eq. 1, i.e., after initializing $\mathbf{w}$ using the labeled data, we iteratively update $\mathbf{w}$ and $\boldsymbol{\beta}$ by fixing the other as constant.

1.1 Stochastic Gradient Descent over w

Let F denote the objective function given in Eq. 1, and $\ell(\mathbf{w}, \mathbf{x}_i, s) = \max(0, 1 - s\mathbf{w} \cdot \mathbf{x}_i)$. The gradient over $\mathbf{w}$ is thus given by:

$$\frac{\partial F}{\partial \mathbf{w}} = \mathbf{w} + C_l \sum_{i} g(\mathbf{w}, \mathbf{x}_i, y_i) + C_{u+} \sum_{i \in I_u} \beta_i g(\mathbf{w}, \mathbf{x}_i, 1) + C_{u-} \sum_{i \in I_u} (1-\beta_i) g(\mathbf{w}, \mathbf{x}_i, -1), \quad (2)$$

where

$$g(\mathbf{w}, \mathbf{x}_i, s) = \begin{cases} 0 & \text{if } s\mathbf{w} \cdot \mathbf{x}_i \geq 1 \\ -s\mathbf{x}_i & \text{otherwise} \end{cases}$$

At each iteration, a randomly selected subset of examples is used for update: $\mathbf{w} := \mathbf{w} - \eta \frac{\partial F}{\partial \mathbf{w}}$, where η is the learning rate (needs to be adaptive?).

1.2 Optimizing over $\boldsymbol{\beta}$

Since each β_i is independent[1], we can optimize over each β_i separately. Consequently, we would like to minimize the following objective function ($\mathbf{w}$ is fixed):

$$F_1 = C_{u+}\beta_i \ell(\mathbf{w}\cdot\mathbf{x}_i, 1) + C_{u-}(1-\beta_i)\ell(\mathbf{w}\cdot\mathbf{x}_i, -1).$$

Expanding the second term and getting rid of the term $C_{u-}\ell(\mathbf{w}\cdot\mathbf{x}_i, -1)$ that does not depend on β_i gives the new objective function:

$$F_2 = \beta_i[C_{u+}\ell(\mathbf{w}\cdot\mathbf{x}_i, 1) - C_{u-}\ell(\mathbf{w}\cdot\mathbf{x}_i, -1)].$$

Thus, the optimal β_i can be solved analytically by

$$\hat{\beta}_i = \begin{cases} 0 & \text{if } C_{u+}\ell(\mathbf{w}\cdot\mathbf{x}_i, 1) - C_{u-}\ell(\mathbf{w}\cdot\mathbf{x}_i, -1) \geq 0 \\ 1 & \text{otherwise} \end{cases}$$

2 Model with $\boldsymbol{\beta}$ regularization

$$\min_{\mathbf{w},\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{w}\|^2 + C_l \sum_{i\in I_l} \ell(\mathbf{w}\cdot\mathbf{x}_i, \mathrm{sign}(y_i)) + C_{u+}\sum_{i\in I_u}\beta_i\ell(\mathbf{w}\cdot\mathbf{x}_i, 1) + C_{u-}\sum_{i\in I_u}(1-\beta_i)\ell(\mathbf{w}\cdot\mathbf{x}_i, -1) + \lambda\|\boldsymbol{\beta}\|_1 \tag{3}$$

The additional ℓ_1 regularization over $\boldsymbol{\beta}$ enforces a strong preference of borrowing the unlabeled data as negatives, which incorporates the prior that the vast majority of the visual world consists of negatives.

[1] Joseph: I think we should come up with a way to model the correlation among β_i's.