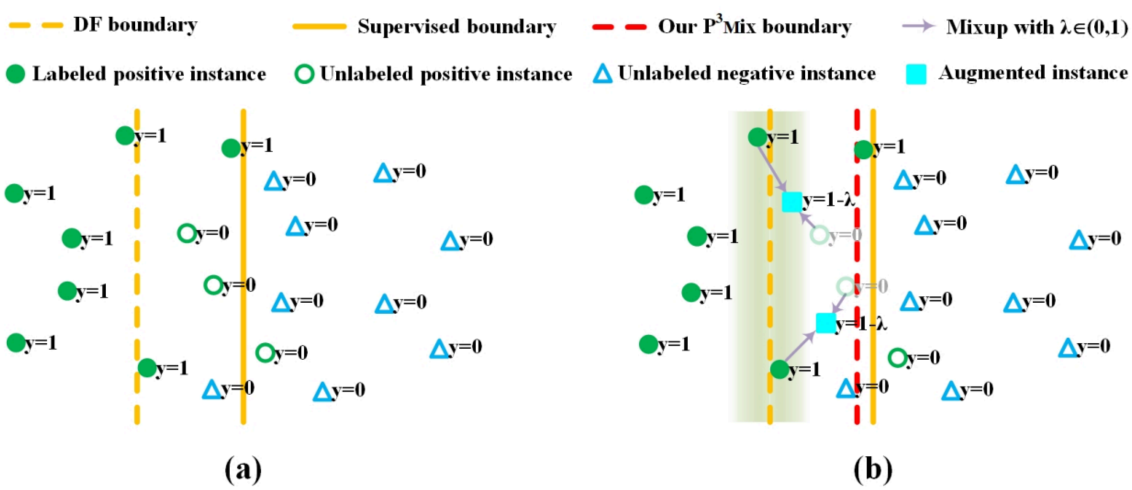
[pdf (openreview.net)]

在pu learning大量使用的cost-sensitive类方法中，模型通常都先将未标记样本当做负样本进行训练，之后进行纠正。对比PU learning和正常全监督学习（PN问题），通过实验可以发现在pu条件下，模型给出的划分正负样本的界限比全监督学习下更靠近正样本附近，这意味着PU条件下模型更倾向于将位于正负样本交界处的，本来应该更可能是正样本的样本划分到负样本去。如图:



为此我们可以思考，能否将同样靠近边界的正样本与这些潜在的假阴性未标记样本融合（mixup），得到新的样本点，这样模型在预测的时候就更可能将他们分到正样本一类，将正负样本的划分线往负样本推，从而提高预测准确率。

核心思想：主要操作针对靠近正负样本分界线的假阴性未标记样本，从靠近边界的标记正样本中按照一定策略取样，两者混合得到新的边界样本点，再送给模型学习。

传统cost-sensitive损失函数如下：

$$\mathcal{L}(\mathcal{X}_p, \mathcal{X}_u; \mathbf{\Theta}) = \frac{1}{|\mathcal{X}_p|} \sum_{(\mathbf{x},y)\in\mathcal{X}_p} \ell\big(f(\mathbf{x};\mathbf{\Theta}), y\big) + \frac{\beta}{|\mathcal{X}_u|} \sum_{(\mathbf{x},y)\in\mathcal{X}_u} \ell\big(f(\mathbf{x};\mathbf{\Theta}), y\big), \tag{1}$$

这种模型下先统一认为未标记样本都是负样本。

经过mixup之后的方法的损失函数为：

$$\mathcal{L}(\widehat{\mathcal{X}}_p, \widehat{\mathcal{X}}_u; \mathbf{\Theta}) = \frac{1}{|\widehat{\mathcal{X}}_p|} \sum_{(\widehat{\mathbf{x}},\widehat{y})\in\widehat{\mathcal{X}}_p} \ell\big(f(\widehat{\mathbf{x}};\mathbf{\Theta}), \widehat{y}\big) + \frac{\beta}{|\widehat{\mathcal{X}}_u|} \sum_{(\widehat{\mathbf{x}},\widehat{y})\in\widehat{\mathcal{X}}_u} \ell\big(f(\widehat{\mathbf{x}};\mathbf{\Theta}), \widehat{y}\big), \tag{2}$$

$$\widehat{\mathcal{X}}_p, \widehat{\mathcal{X}}_u = \mathrm{HeuristicMixup}(\mathcal{X}_p, \mathcal{X}_u, \alpha), \tag{3}$$

混合策略：对于一个边际假阴性未标记样本(x_i, y_i),寻找边际标记正样本点(x_j, y_j)，将二者按照这种策略混合：

$$\widehat{\mathbf{x}}_i = \lambda'\mathbf{x}_i + (1-\lambda')\mathbf{x}_j, \quad \widehat{y}_i = \lambda' y_i + (1-\lambda')y_j, \quad \lambda' = \max(\lambda, 1-\lambda), \\ \lambda \sim \text{Beta}(\alpha, \alpha), \quad \alpha \in (0, \infty), \quad (4)$$

accordingly forming the augmented instance sets $\widehat{\mathcal{X}}_p$ and $\widehat{\mathcal{X}}_u$.

新的样本加入样本集中。用来和边际假阴性样本(mpn)进行混合的样本(cnd)的选取方法：

$$(\mathbf{x}_j, y_j) \sim \begin{cases} \text{Uniform}(\mathcal{X}_{cnd}) & \text{if } (\mathbf{x}_i, y_i) \in \mathcal{X}_{mpn}, \\ \text{Uniform}(\mathcal{X}_p \cup \mathcal{X}_u) & \text{if } (\mathbf{x}_i, y_i) \in \mathcal{X}_p \cup \mathcal{X}_u \setminus \mathcal{X}_{mpn}. \end{cases} \quad (5)$$

至于x_{mpn}和x_{cnd}的计算方法如下：

Marginal pseudo-negative instance estimation. Because the fully supervised boundary is exactly unknown, we have to estimate the set of marginal pseudo-negative instances $\mathcal{X}_{mpn}$ from $\mathcal{X}_u$. In this work, we define them as the "unreliable" pseudo-negative instances measured by the predictive scores with thresholding parameter $\gamma \in [0.5, 1]$:

$$\mathcal{X}_{mpn} = \big\{(\mathbf{x}, y=0) | (\mathbf{x}, y=0) \in \mathcal{X}_u, 1-\gamma \le f(\mathbf{x}; \mathbf{\Theta}) \le \gamma\big\}, \quad (6)$$

where $\gamma = 0.5$ implies $\mathcal{X}_{mpn} = \emptyset$, and $\gamma = 1$ means $\mathcal{X}_{mpn} = \mathcal{X}_u$.

Candidate mixup pool. We maintain a candidate mixup pool $\mathcal{X}_{cnd}$ containing the positive instances around the current learned boundary from $\mathcal{P}$. To be specific, for each positive instance we compute its entropy value of the predictive score, and update the candidate mixup pool with the top-k positive instances as follows:

$$\mathcal{X}_{cnd} = \big\{(\mathbf{x}, y=1) | (\mathbf{x}, y=1) \in \mathcal{P}, \mathcal{H}(f(\mathbf{x}; \mathbf{\Theta})) \in \text{Rank}(\{\mathcal{H}(f(\mathbf{x}_i; \mathbf{\Theta}))\}_{i=1}^{n_p})\big\}, \quad (7)$$

where $\mathcal{H}(\cdot)$ is the entropy, and $\text{Rank}(\cdot)$ outputs a set of positive instances with the top-k maximum entropy values. For efficiency, we update $\mathcal{X}_{cnd}$ per-epoch. The full training procedure is shown in *Algorithm 1*.