Semi-Supervised Classification based on Positive-Unlabeled Classification

Tomoya Sakai[1,2] **Marthinus Christoffel du Plessis** **Gang Niu**[1] **Masashi Sugiyama**[2,1]

[1]The University of Tokyo, Japan [2]RIKEN, Tokyo, Japan

In real-world machine learning tasks, the size of labeled data is often limited due to laborious manual annotation. In contrast, unlabeled data can be collected more cheaply and abundantly. Based on this fact, various semi-supervised classification methods have been proposed in the past decades.

In an existing semi-supervised classification approach, we often rely on particular assumptions on the data distribution to utilize unlabeled data (Chapelle, Schölkopf, and Zien, 2006). For example, the *cluster assumption* supposes that samples in the same cluster are likely to share the same label. Based on such a distributional assumption, the existing framework leverages unlabeled data to construct a regularizer for a classifier and *biases* the classifier toward a better one if the assumption is correct. However, if the distributional assumption does not agree with the data distribution, the bias adversely affects the performance of the obtained classifier that is even worse than the one obtained with supervised classification algorithms.

Recently, *positive-unlabeled classification* (PU classification), which trains a classifier from only positive and unlabeled data, has been gathering growing attention (du Plessis, Niu, and Sugiyama, 2015; Kanehira and Harada, 2016). In PU classification, unlabeled data is utilized for *risk evaluation*, implying that label information is directly extracted from unlabeled data without specific distributional assumptions, unlike existing methods. Furthermore, state-of-the-art theoretical analysis (Niu et al., 2016) showed that PU classification can outperform positive-negative classification (PN classification, i.e., ordinary supervised classification) under some conditions. Thus, it is expected that combining PN with PU classification can be a promising approach to semi-supervised classification without restrictive distributional assumptions.

More specifically, let $\boldsymbol{x} \in \mathbb{R}^d$ and $y \in \{\pm 1\}$ be equipped with probability density $p(\boldsymbol{x}, y)$, where d is a positive integer, and $\theta_{\mathrm{P}} := p(y=+1)$ and $\theta_{\mathrm{N}} := p(y=-1)$ be the class-priors. Suppose we have three sets of samples $\{\boldsymbol{x}_i^{\mathrm{P}}\}_{i=1}^{n_{\mathrm{P}}}$, $\{\boldsymbol{x}_i^{\mathrm{N}}\}_{i=1}^{n_{\mathrm{N}}}$, and $\{\boldsymbol{x}_i^{\mathrm{U}}\}_{i=1}^{n_{\mathrm{U}}}$ drawn independently from $p(\boldsymbol{x}|y=+1)$, $p(\boldsymbol{x}|y=-1)$, and $p(\boldsymbol{x})$, respectively. Furthermore, let $g\colon \mathbb{R}^d \to \mathbb{R}$ be a decision function for binary classification and $\ell\colon \mathbb{R} \to \mathbb{R}$ be a loss function that imposes penalty on g when a sample is wrongly classified. The goal of classification is to minimize the *true* risk $R(g) := \mathrm{E}_{p(\boldsymbol{x},y)}[\ell(yg(\boldsymbol{x}))]$.

In PN classification, we use its equivalent expression (the PN risk):

$$R_{\mathrm{PN}}(g) := \theta_{\mathrm{P}}\,\mathrm{E}_{\mathrm{P}}[\ell(g(\boldsymbol{x}))] + \theta_{\mathrm{N}}\,\mathrm{E}_{\mathrm{N}}[\ell(-g(\boldsymbol{x}))],$$

where E_{P} and E_{N} are the expectations over $p(\boldsymbol{x}|y=+1)$ and $p(\boldsymbol{x}|y=-1)$, respectively. In contrast, the risk in PU classification (the PU risk), which is equivalent to the PN risk, can be computed from only positive and unlabeled data:

$$R_{\mathrm{PU}}(g) := \theta_{\mathrm{P}}\,\mathrm{E}_{\mathrm{P}}[\widetilde{\ell}(g(\boldsymbol{x}))] + \mathrm{E}_{\mathrm{U}}[\ell(-g(\boldsymbol{x}))],$$

where E_{U} is the expectation over $p(\boldsymbol{x})$ and $\widetilde{\ell}(m) := \ell(m) - \ell(-m)$. In addition to the PU risk, we also define the risk in negative-unlabeled classification (the NU risk) as $R_{\mathrm{NU}}(g) := \theta_{\mathrm{N}}\,\mathrm{E}_{\mathrm{N}}[\widetilde{\ell}(-g(\boldsymbol{x}))] + \mathrm{E}_{\mathrm{U}}[\ell(g(\boldsymbol{x}))]$.

Our idea is to combine the PN risk with the PU/NU risks:

$$R_{\mathrm{PNU}}^{\eta}(g) = \begin{cases} (1-\eta)R_{\mathrm{PN}}(g) + \eta R_{\mathrm{PU}}(g) & (\eta \geq 0), \\ (1+\eta)R_{\mathrm{PN}}(g) - \eta R_{\mathrm{NU}}(g) & (\eta < 0), \end{cases}$$

where $\eta \in [-1, 1]$ is the combination parameter. The empirical risk can be obtained by replacing the expectations with corresponding sample averages. For the proposed empirical risk, we can theoretically guarantee the following properties without the distributional assumptions that are imposed in the existing methods: (i) it is unbiased to the true risk, (ii) the variance is smaller than the plain PN risk, and (iii) the confidence term of the generalization error bound converges with the optimal parametric rate with respect to the number of positive, negative, and unlabeled samples. Through extensive numerical experiments, we analyzed the behavior of the proposed risk and demonstrated the usefulness of the proposed method (see Section 5 in Sakai et al., 2017).

Acknowledgements: TS was supported by JSPS KAKENHI 15J09111. GN was supported by the JST CREST program and Microsoft Research Asia. MCdP and MS were supported by the JST CREST program.

References

Chapelle, O.; Schölkopf, B.; and Zien, A., eds. 2006. *Semi-Supervised Learning*. MIT Press.

du Plessis, M. C.; Niu, G.; and Sugiyama, M. 2015. Convex formulation for learning from positive and unlabeled data. In *ICML*.

Kanehira, A., and Harada, T. 2016. Multi-label ranking from positive and unlabeled data. In *CVPR*.

Niu, G.; du Plessis, M. C.; Sakai, T.; Ma, Y.; and Sugiyama, M. 2016. Theoretical comparisons of positive-unlabeled learning against positive-negative learning. In *NIPS*.

Sakai, T.; du Plessis, M. C.; Niu, G.; and Sugiyama, M. 2017. Semi-supervised classification based on classification from positive and unlabeled data. *arXiv preprint arXiv:1605.06955*.