Learning from Incomplete and Inaccurate Supervision

Zhen-Yu Zhang, Peng Zhao, Yuan Jiang, and Zhi-Hua Zhou, *Fellow, IEEE*

Abstract—In plenty of real-life tasks, strongly supervised information is hard to obtain, and thus weakly supervised learning has drawn considerable attention recently. This paper investigates the problem of learning from incomplete and inaccurate supervision, where only a limited subset of training data is labeled but potentially with noise. This setting is challenging and of great importance but rarely studied in the literature. We notice that in many applications, the limited labeled data are with certain structures, which paves us a way to design effective methods. Specifically, we observe that labeled data are usually with one-sided noise such as the bug detection task, where the identified buggy codes are indeed with defects, while codes checked many times or newly fixed may still have other flaws. Furthermore, when there occurs two-sided noise in the labeled data, we exploit the class-prior information of unlabeled data, which is typically available in practical tasks. We propose novel approaches for the incomplete and inaccurate supervision learning tasks and effectively alleviate the negative influence of label noise with the help of a vast number of unlabeled data. Both theoretical analysis and extensive experiments justify and validate the effectiveness of the proposed approaches.

Index Terms—weakly supervised learning, semi-supervised learning, noisy label learning.

1 Introduction

Machine learning has achieved great success in many real-world tasks, especially in supervised learning scenarios. These techniques, such as deep learning [1], typically require a vast number of training data with accurate labels to obtain good performance. However, such strong supervision is not easy to obtain since the labeling process requires human effort expertise. Therefore, it is desired to facilitate the learning system with the capability of preserving satisfactory performance with *weak supervision* [2].

In this paper, we consider the problem of learning from *incomplete and inaccurate* supervision. Specifically, only a small subset of training data is observed with labels while the others remain unlabeled, and meanwhile, the observed labels might be inaccurate. This setting is crucial because it occurs in a variety of real-world applications. For instance, consider the task of medical images annotation in the hospital, there exist amounts of medical images without labels, since the number of doctors is usually limited. Even for those labeled images, they could be wrongly annotated by doctors due to their difficulties. Similar situations also occur in building the learning system from biology data: supervised information of each molecule is not always correct due to limitations of the equipment capability, and the number of labeled molecules is also limited since it is usually too costly to conduct biological experiments for collecting labels.

Learning from incomplete supervision or inaccurate supervision has been studied in the area of Semi-Supervised Learning (SSL) [3], [4] and Noisy Label Learning (NLL) [5], [6], separately. From incomplete supervision, SSL approaches use a vast number of unlabeled data as well as the limited labeled data to construct the model. However, when labeled data are inaccurate, the learning system could be seriously deceived. Under inaccurate supervision, NLL approaches manage to recover the underlying noise-free distribution with noisy labels, in order to learn the predictor which resists the noise. Nevertheless, they typically require a large amount of labeled data and cannot exploit unlabeled data. Therefore, it is very desired to design approaches that can *learn from incomplete and inaccurate supervision simultaneously*. More precisely, we need effective algorithms to handle the task where there are only a limited number of potentially noisy labeled data, and a vast number of unlabeled data.

The problem turns out quite challenging, and it is non-trivial to combine advantages of SSL and NLL approaches to address this problem. For conventional noisy label learning approaches, on the one hand, labeled data are insufficient to estimate the underlying noise-free distribution; on the other hand, these approaches are not able to access label information from unlabeled data, and thus cannot leverage the incomplete supervision to alleviate the label noise. For traditional semi-supervised learning approaches, to handle a vast number of unlabeled data, an underlying assumption is that supervision information should be reliable. Otherwise, these noisy labels can significantly mislead the learning system. For example, in graph-based SSL, if labeled data are not trustworthy, the algorithm probably converges to an arbitrary result because the predicted labels of unlabeled data are propagated depends on these labeled data.

With only noisy labeled data and unlabeled data, it is almost impossible to learn from such incomplete and inaccurate supervision, particularly when limited labeled data are arbitrarily corrupted. Fortunately, in many real-world tasks, we have some certain side information on the structure of the observed labeled data. Specifically, we are concerned with the circumstance where the limited labeled data are with *one-sided instance-dependent* noise. Namely, only

• *Z.-Y Zhang, P. Zhao, Y. Jiang and Z.-H. Zhou are with National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China.* (*Corresponding author: Yuan Jiang.*)
E-mail: : {zhangzy, zhaop, jiangy, zhouzh}@lamda.nju.edu.cn

Manuscript received XXX XX, 20XX; revised XXX XX, 20XX.

labels in one category may flip into the other category with an unknown, instance-dependent noise rate while the other category is clean. Such a scenario is quite common in real-world tasks. For example, in the bug detection task, we aim to identify the buggy code from a large number of code files. The codes reported with bug issues by the senior engineers are surely buggy (clean label). Nevertheless, some codes that have been check many times or fixed recently could remain with bugs (noisy label) due to the complexity of the software system. Moreover, plenty of codes are not labeled since it is hard to check them one-by-one entirely (unlabeled).

Besides the one-sided label noise, we further consider the case where the label noise simultaneously occur on both positive and negative sides. This is evidently much more challenging than the one-sided label noise scenario, because there is no reliable label information. To overcome the difficulty caused by two-sided label noise, we exploit additional statistical information of the incomplete supervisions for this learning task, that is, we leverage the power of unlabeled data with class-priors as the side statistical information. Recall the aforementioned example of primary screening scenario, when the equipment is of low quality, it may cause misdiagnosis on both healthy and ill people, which leads to the label noise occurring on both positive and negative sides (two-sided noisy label). Since the data of residents (unlabeled) usually come from different communities, our proposal is to exploit some official statistics, e.g., the class-priors, which associate with these communities as the side information. Therefore, we can leverage unlabeled data with side information to cope with the two-sided label noise.

This paper extends our preliminary study [7]. In this paper, we investigate a popular but challenging learning problem, namely, the *Learning from Incomplete and Inaccurate SuPervision* (LIISP), which accommodates a variety of real-world applications. We propose a novel semi-supervised learning method, leveraging the incomplete supervision to alleviate the negative effect caused by inaccurate supervision, and thus step towards learning from incomplete and inaccurate supervision simultaneously. The main idea is to rewrite the true risk of the underlying noise-free distribution in the importance weighting form. Enlightened by the recent advance of positive-unlabeled learning [8], [9], [10], we use the marginal distribution extracted from the incomplete supervision (unlabeled data) along with accurate labels to estimate the weights, and thus construct the risk minimizer for the incomplete and one-sided inaccurate supervision. Furthermore, for the more challenging scenario where the label noise occurs on both positive and negative sides, we additionally exploit the class-priors of two discrepant unlabeled datasets to resist the two-sided label noise. Inspired by the unlabeled-unlabeled learning [11], [12], we expand our method to handle the inaccurate supervision with the help of discrepant incomplete supervisions. Both theoretical justifications and empirical studies demonstrate the benefit of unlabeled data and noisy labeled data, and thereby we can obtain the optimal convergence rate and remarkable performance improvement.

We summarize our main contributions as follows.

(1) We introduce and investigate the problem of Learning from Incomplete and Inaccurate SuPervision (LIISP), which accommodates many real-world applications but is rarely considered in the literature.
(2) We propose novel learning algorithms, which alleviate the noisy labeled data with the help of unlabeled data. We theoretically justify the effectiveness of unlabeled and noisy data via the excess risk analysis.
(3) We conduct extensive empirical evaluations on synthetic, benchmark datasets, and real-world applications to demonstrate the superiority and robustness of our proposed methods.

In the following, we first briefly review related work in Section 2. Then, we introduce some preliminary background knowledge in Section 3. Next, we provide a detailed description of our proposed methods in Section 4 and Section 5, with detailed proofs in the supplemental material. Experimental results on synthetic, benchmark, and real-world datasets are in Section 6. Finally, we conclude the paper in Section 7.

2 Related work

Starting from the pioneering work of learning with noisy labels [13], a variety of studies on inaccurate supervision have been proposed in the theoretical community. For instance, Aslam et al. [14] studied the learnability of noise tolerant learning in finite VC-dimension. Apart from theoretical findings, various practical approaches are also proposed to avoid the drawback caused by inaccurate supervision, for example, perceptron algorithms [15], [16], robust loss [17], [18], unbiased loss [5], [19], importance-reweighting on training samples [20], [21], etc. Following the line of noisy label learning, instance-independent noise is firstly investigated [5], [20]. These primary studies provide guarantees for risk minimization under random classification noise in the general setting of convex surrogates. In practice, instance-dependent noise [6], [22], [23] is much closer to the realistic situation, where label noise depends on the intrinsic nature of instances. This setting is arguably more complicated than the instance-independent label noise scenario. Preliminary research shows that the optimal classifiers can be recovered from the noisy distribution under certain assumptions [6]. However, noisy label learning mainly focuses on supervised learning field, how to deal with limited labeled data and large amounts of unlabeled data has not yet been well studied.

To take advantage of incomplete supervision, semi-supervised learning algorithms are proposed to utilize unlabeled data along with limited labeled data to construct the predictor. Theoretical analysis shows that, provided with a reasonable assumption on unlabeled data, like the cluster assumption or the manifold assumption [24], [25], unlabeled data can be used to regularize the hypothesis space and thus reduce the searching complexity. Plenty of practical approaches have been proposed over the decades, e.g., graph-based methods [4], [26], S3VMs [3], and disagreement-based methods [27], [28]. In recent years, due to the powerful feature representation ability of deep neural networks [29], some deep-SSL approaches have also been proposed [30], [31]. In traditional SSL, the supervision information should be accurate, which usually does not hold in practice.

A different point of view in semi-supervised learning is formulated as Positive-Unlabeled Learning (PU Learning) [8], [32]. Different from using unlabeled data as the

regularizer of hypothesis space, PU learning assumes the unlabeled data are generated from the same joint distribution as labeled data, but their labels cannot be observed. To deal with the semi-supervised learning task, they linearly combine PU and NU (Negative-Unlabeled) and give theoretical analysis [10]. Nevertheless, PU learning requires sufficient positive data to simulate the effect of the negative part along with unlabeled data, which cannot be satisfied under incomplete supervision. A recent breakthrough in semi-supervised learning shows that with necessary statistical information, the optimal classifier can be obtained by two unlabeled datasets with different class priors [11], [12]. However, they do not exploit the labeled data, which usually contain considerably important supervised information.

Note that disagreement-based SSL approaches also exploit pseudo-labels of unlabeled data [28], and to handle misleading pseudo-labels, some strategies such as data editing [33] or one-sided noisy label learning [34] have been incorporated. These can be seen as early studies considering both incomplete supervision and inaccurate supervision, though the inaccurate supervision was generated during the SSL procedure, rather than the label noise in the initial training data. Some recent studies about Safe-SSL [35], [36], [37] also have inherent mechanisms to handle the label noise, though these mechanisms are implicit. In recent years, there are some other studies [38], [39], [40], which tried to improve the robustness of SSL, but they were mostly heuristic and did not consider structural properties.

3 PRELIMINARY

In this section, we first review the notations for learning from complete and accurate supervision, namely, conventional supervised learning. Then, we introduce preliminary knowledge for learning from incomplete supervision, which is one of the typical scenarios in weakly supervised learning.

3.1 Learning from Complete and Accurate Supervision

In this scenario, we observe the ground-truth label for each instance. Let $\mathcal{D}$ be the underlying true distribution from which the training data $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ are independently and identically sampled, where $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} = \{-1, +1\}$. Given n_P positive data $\{(\mathbf{x}_i, +1)\}_{i=1,\ldots,n_P}$ and n_N negative data $\{(\mathbf{x}_j, -1)\}_{j=1,\ldots,n_N}$, our purpose is to learn a well-generalized decision function $g : \mathcal{X} \mapsto \mathbb{R}$ over the underlying distribution $\mathcal{D}$ for the binary classification task.

Denote by $\ell : \mathbb{R} \times \mathcal{Y} \mapsto \mathbb{R}_+$ a non-negative Lipschitz-continuous loss function, whose risk over the underlying true distribution $\mathcal{D}$ is

$$\begin{aligned} R(g) &= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(g(\mathbf{x}), y)] \\ &= \pi_P \mathbb{E}_P[\ell(g(\mathbf{x}), +1)] + \pi_N \mathbb{E}_N[\ell(g(\mathbf{x}), -1)], \end{aligned} \tag{1}$$

where π_P is the class-prior of positive data $\Pr[y = +1]$ and π_N of negative data $\Pr[y = -1]$ with $\pi_P + \pi_N = 1$. Besides, $\mathbb{E}_P$ and $\mathbb{E}_N$ denote the expectation of conditional probability $\Pr[\mathbf{x}|y = +1]$ and $\Pr[\mathbf{x}|y = -1]$, respectively.

As only the sampled data are accessible in practice, we approximate the risk by the empirical one,

$$\widehat{R}(g) = \frac{\pi_P}{n_P} \sum_{i=1}^{n_P} \ell(g(\mathbf{x}_i), +1) + \frac{\pi_N}{n_N} \sum_{j=1}^{n_N} \ell(g(\mathbf{x}_j), -1).$$

Given a family of decision functions $\mathcal{G}$, in which each function $g : \mathcal{X} \mapsto \mathbb{R}$, we denote g^* as the optimal decision function, with $\widehat{g}$ as its empirical version,

$$g^* = \arg\min_{g\in\mathcal{G}} R(g), \qquad \widehat{g} = \arg\min_{g\in\mathcal{G}} \widehat{R}(g).$$

3.2 Learning from Incomplete Supervision

In this part, we consider the scenario of learning from incomplete supervision. It is extremely hard to learn with only unlabeled data on hand, so that we assume that we can obtain some prior knowledge for the learning task.

Learning from Positive-Unlabeled Data. As aforementioned in the introduction, we first consider the scenario where the prior knowledge is a handful of data with their *ground-truth labels from one category*. We assume without loss of generality that there are n_P positive data $\{(\mathbf{x}_i, +1)\}_{i=1,\ldots,n_P}$ and n_U unlabeled data $\{\mathbf{x}_k\}_{k=1,\ldots,n_U}$. Our purpose is still to learn a real-valued function g with small generalization error for the binary classification task.

As the negative data are not available in this scenario, thus the partial risk $\pi_N \mathbb{E}_N[\ell(g(\mathbf{x}), -1)]$ in (1) cannot be directly estimated. Fortunately, based on the seminal work [8], the risk $R(g)$ can be recovered in an unbiased manner by only using the accurate positive (or negative) and unlabeled data. In the following, we suppose that the loss function ℓ satisfies the symmetric condition,

$$\ell(g(\mathbf{x}), +1) + \ell(g(\mathbf{x}), -1) = 1. \tag{2}$$

The symmetric condition is met by using a scaled ramp loss as the surrogate loss, which is classification-calibrated [9]. Based on the symmetric condition, we retrieve the partial risk $\pi_N \mathbb{E}_N[\ell(g(\mathbf{x}), -1)]$ by regarding the unlabeled data as negative data, and write the risk $\mathbb{E}_U[\ell(g(\mathbf{x}), -1)]$ as

$$\begin{aligned} &\pi_P \mathbb{E}_P[\ell(g(\mathbf{x}), -1)] + \pi_N \mathbb{E}_N[\ell(g(\mathbf{x}), -1)] \\ &= \pi_P \mathbb{E}_P[1 - \ell(g(\mathbf{x}), +1)] + \pi_N \mathbb{E}_N[\ell(g(\mathbf{x}), -1)] \\ &= -\pi_P \mathbb{E}_P[\ell(g(\mathbf{x}), +1)] + \pi_N \mathbb{E}_N[\ell(g(\mathbf{x}), -1)] + \pi_P. \end{aligned}$$

Therefore, let ℓ be a non-negative Lipschitz-continuous loss function and satisfies the symmetric condition in (2), then the risk can be rewritten in in unbiased manner as

$$R(g) = 2\pi_P \mathbb{E}_P[\ell(g(\mathbf{x}), +1)] + \mathbb{E}_U[\ell(g(\mathbf{x}), -1)] - \pi_P.$$

Approximating the risk $R(g)$ by empirical data, we obtain,

$$\widehat{R}_{PU}(g) = \frac{2\pi_P}{n_P} \sum_{i=1}^{n_P} \ell(g(\mathbf{x}_i), +1) + \frac{1}{n_U} \sum_{k=1}^{n_U} \ell(g(\mathbf{x}_k), -1). \tag{3}$$

For a given family of decision functions $\mathcal{G}$, we denote $\widehat{g}_{PU}$ as the minimizer of (3), that is,

$$\widehat{g}_{PU} = \arg\min_{g\in\mathcal{G}} \widehat{R}_{PU}(g).$$

Learning from Unlabeled-Unlabeled Data. We then consider another scenario in the incomplete supervision learning, in which only the prior knowledge of *class-priors* is accessible. We suppose that we have two discrepant unlabeled datasets with necessary known class-priors. Let $\Pr_{U_1}(\mathbf{x}, y)$ and $\Pr_{U_2}(\mathbf{x}, y)$ be two marginal densities where these two discrepant unlabeled datasets are generated. We denote by

θ_P and θ'_P ($\neq \theta_P$) the two class-priors for the positive data of two unlabeled datasets, so that

$$\Pr_{U_1}(\mathbf{x}, y) = \theta_P \Pr(\mathbf{x}|y=+1) + \theta_N \Pr(\mathbf{x}|y=-1),$$
$$\Pr_{U_2}(\mathbf{x}, y) = \theta'_P \Pr(\mathbf{x}|y=+1) + \theta'_N \Pr(\mathbf{x}|y=-1),$$

where $\theta_P + \theta_N = 1$ and $\theta'_P + \theta'_N = 1$. Here $\Pr(\mathbf{x}|y=\pm 1)$ are the class conditional densities from which the positive/negative data are generated.

Since both positive and negative data are not available in this scenario, we construct an unbiased estimation of the underlying true risk with discrepant incomplete supervisions. Following the work of [11], we rewrite $R(g)$ in the form that

$$R(g) = \mathbb{E}_{U_1}(\bar{\ell}(g(\mathbf{x}), +1)) + \mathbb{E}_{U_2}(\bar{\ell}(g(\mathbf{x}), -1)),$$

where $\mathbb{E}_{U_1}$ and $\mathbb{E}_{U_2}$ are the exception over the marginal distributions of these two discrepant unlabeled datasets, $\bar{\ell}(\cdot, +1) = a\ell(\cdot, +1) + b\ell(\cdot, -1)$ and $\bar{\ell}(\cdot, -1) = c\ell(\cdot, +1) + d\ell(\cdot, -1)$ are the corrected loss functions for incomplete supervisions, respectively. Then, we can write

$$\begin{aligned}
R(g) &= \mathbb{E}_{U_1}[\bar{\ell}_{U_1}(g(\mathbf{x}), +1)] + \mathbb{E}_{U_2}[\bar{\ell}_{U_2}(g(\mathbf{x}), -1)] \\
&= \theta_P \mathbb{E}_P[a \cdot \ell(g(\mathbf{x}), +1) + b \cdot \ell(g(\mathbf{x}), -1)] \\
&\quad + (1-\theta_P)\mathbb{E}_N[a \cdot \ell(g(\mathbf{x}), +1) + b \cdot \ell(g(\mathbf{x}), -1)] \\
&\quad + \theta'_P \mathbb{E}_P[c \cdot \ell(g(\mathbf{x}), -1) + d \cdot \ell(g(\mathbf{x}), +1)] \\
&\quad + (1-\theta'_P)\mathbb{E}_N[c \cdot \ell(g(\mathbf{x}), -1) + d \cdot \ell(g(\mathbf{x}), +1)] \\
&= (a \cdot \theta_P + d \cdot \theta'_P)\mathbb{E}_P[\ell(g(\mathbf{x}), +1)] \\
&\quad + (b \cdot \theta_P + c \cdot \theta'_P)\mathbb{E}_P[\ell(g(\mathbf{x}), -1)] \\
&\quad + [a \cdot (1-\theta_P) + d \cdot (1-\theta'_P)]\mathbb{E}_N[\ell(g(\mathbf{x}), +1)] \\
&\quad + [b \cdot (1-\theta_P) + c \cdot (1-\theta'_P)]\mathbb{E}_N[\ell(g(\mathbf{x}), -1)].
\end{aligned}$$

By setting the coefficients of terms $\mathbb{E}_P[\ell(g(x), -1)]$ and $\mathbb{E}_N[\ell(g(x), +1)]$ to zero and letting $a \cdot \theta_P + d \cdot \theta'_P = \pi_P$, $b \cdot (1-\theta_P) + c \cdot (1-\theta'_P) = \pi_N$, we immediately retrieve the risk $R(g)$, with four coefficients

$$a = \frac{(1-\theta'_P)\pi_P}{\theta_P - \theta'_P}, \quad b = -\frac{\theta'_P(1-\pi_P)}{\theta_P - \theta'_P},$$
$$c = \frac{\theta_P(1-\pi_P)}{\theta_P - \theta'_P}, \quad d = -\frac{(1-\theta_P)\pi_P}{\theta_P - \theta'_P},$$

and $R(g)$ is rewritten in an unbiased manner as

$$\begin{aligned}
R(g) &= \alpha \cdot \mathbb{E}_{U_1}[\ell(g(\mathbf{x}), +1)] + \alpha' \cdot \mathbb{E}_{U_2}[\ell(g(\mathbf{x}), -1)] \\
&\quad - \frac{\theta'(1-\pi_P) + (1-\theta)\pi_P}{\theta - \theta'},
\end{aligned}$$

where $\alpha = (\theta' + \pi_P - 2\theta'\pi_P)/(\theta - \theta')$ and $\alpha' = (\theta + \pi_P - 2\theta\pi_P)/(\theta - \theta')$.

Given two discrepant unlabeled datasets of size n_{U_1} and n_{U_2}, the empirical estimator can be approximated by

$$\begin{aligned}
\widehat{R}_{UU}(g) &= \frac{1}{n_{U_1}} \sum_{i=1}^{n_{U_1}} \alpha \ell(g(\mathbf{x}_i), +1) + \frac{1}{n_{U_2}} \sum_{j=1}^{n_{U_2}} \alpha' \ell(g(\mathbf{x}_i), -1) \\
&\quad - \frac{\theta'(1-\pi_P) + (1-\theta)\pi_P}{\theta - \theta'}.
\end{aligned} \tag{4}$$

Let $\widehat{g}_{UU} \in \mathcal{G}$ denote the minimizer of the risk estimated by the two unlabeled datasets with known class-priors π_P, θ_P and θ'_P in (4), that is,

$$\widehat{g}_{UU} = \arg\min_{g \in \mathcal{G}} \widehat{R}_{UU}(g).$$

4 Learning from Incomplete and One-sided Inaccurate Supervision

In this section, we present our approach to leverage incomplete supervision to help learning with one-sided inaccurate supervision, in particular, instances with one-sided instance-dependent noisy labels. We demonstrate that the incomplete supervision plays a significant role in learning from the one-sided inaccurate supervision, especially when these noisy labeled data are scarce.

To deal with the instances with one-sided noisy labels, we first rewrite the risk of the underlying true distribution, in which weights σ_+ and σ_- for each noisy labeled instance play crucial roles. Then, we proceed to estimate these two weights with the help of a vast number of unlabeled data. Finally, we provide our learning algorithm for incomplete and one-sided inaccurate supervision.

4.1 Learning from one-sided Inaccurate Supervision

In the one-sided inaccurate supervision, without loss of generality, we suppose positive data are clean and negative data are with instance-dependent label noise.

Notations and Settings. Suppose that we have $n_{\widetilde{P}}$ clean positive data $\widetilde{P} = \{(\mathbf{x}_i, +1)\}_{i=1,\ldots,n_{\widetilde{P}}}$[1] and $n_{\widetilde{N}}$ noisy negative data $\widetilde{N} = \{(\mathbf{x}_j, -1)\}_{j=1,\ldots,n_{\widetilde{N}}}$. For each instance $\mathbf{x}$, let its true label be y and the observed label be $\widehat{y}$. Evidently, we have $y = \widehat{y}$ for clean data, while it does not hold for the noisy data. Meanwhile, let $\pi_{\widetilde{P}}$ be the class-prior of the observed positive label $\Pr[\widehat{y} = +1]$ and $\pi_{\widetilde{N}}$ be $\Pr[\widehat{y} = -1]$ with $\pi_{\widetilde{P}} + \pi_{\widetilde{N}} = 1$.

We suppose that the observed noisy data are with *instance-dependent* label noise [6], [22]. Specifically, for any (underlying, true) positive example $\mathbf{x}$ (whose true label $y = +1$), it is observed as a negative example ($\widehat{y} = -1$) based on its feature. We define this probability as the *hardness*, formally, $h_P : \mathcal{X} \to [0, 1]$, with,

$$h_P(\mathbf{x}) = \Pr[\widehat{y} = -1|\mathbf{x}, y = +1].$$

As the observed positive data are *always* accurate, we have for any $\mathbf{x} \in \widetilde{P}$,

$$\Pr[y = +1|\mathbf{x}, \widehat{y} = +1] = 1.$$

Now we are ready to retrieve the risk of underlying distribution under one-sided inaccurate supervision.

Rewrite True Risk. In the inaccurate supervision learning scenario, if we simply treat all observed data as accurate ones and directly adopt the risk in (1), both empirical and theoretical performance will suffer from the label noise heavily. In order to obtain the optimal classifier, it is necessary to rewrite the true risk. In the following, we propose the *oInAS risk* for the one-sided InAccurate Supervision, and show that it provably retrieves the true risk.

Definition 1 (Risk for one-sided InAccurate Supervision (oInAS Risk))**.** For any function $g \in \mathcal{G}$, its oInAS risk $R^{IA}_{os}(g)$ is defined as,

$$\begin{aligned}
R^{IA}_{os}(g) &= \pi_{\widetilde{P}} \mathbb{E}_{\widetilde{P}}[\sigma_+(\mathbf{x}) \cdot \ell(g(\mathbf{x}), +1)] \\
&\quad + \pi_{\widetilde{N}} \mathbb{E}_{\widetilde{N}}[\sigma_-(\mathbf{x}) \cdot \ell(g(\mathbf{x}), -1)],
\end{aligned}$$

1. We use $\widetilde{P}$ instead of P since there are some true positive data are not revealed, which are observed as negative data.

where weights $\sigma_+(\mathbf{x})$ and $\sigma_-(\mathbf{x})$ are defined as

$$\begin{aligned}\sigma_+(\mathbf{x}) &= 1/\Pr[\widehat{y}=+1|\mathbf{x}, y=+1],\\ \sigma_-(\mathbf{x}) &= \Pr[y=-1|\mathbf{x}, \widehat{y}=-1].\end{aligned} \tag{5}$$

Then we show that the oInAS risk equals to the true risk over the underlying distribution $\mathcal{D}$.

Theorem 1. *The oInAS risk equals to the true risk* (*the risk over the true data distribution*)*, that is,*

$$R_{os}^{IA}(g) = R(g).$$

Proof. The true risk $R(g)$ is the sum of $\pi_P\mathbb{E}_P[\ell(g(\mathbf{x}),+1)]$ and $\pi_N\mathbb{E}_N[\ell(g(\mathbf{x}),-1)]$. For the expectation over the margin distribution of negative data, we have

$$\begin{aligned}&\mathbb{E}_N[\ell(g(\mathbf{x}),-1)]\\ &= \int \ell(g(\mathbf{x}),-1)\Pr[\mathbf{x}|\widehat{y}=-1]\frac{\Pr[\mathbf{x}|y=-1]}{\Pr[\mathbf{x}|\widehat{y}=-1]}\,\mathrm{d}\mathbf{x}\\ &= \int \ell(g(\mathbf{x}),-1)\Pr[\mathbf{x}|\widehat{y}=-1]\frac{\Pr[\widehat{y}=-1]}{\Pr[y=-1]}\sigma_-(\mathbf{x})\,\mathrm{d}\mathbf{x}\\ &= \frac{\pi_{\widetilde{N}}}{\pi_N}\mathbb{E}_{\widetilde{N}}[\sigma_-(\mathbf{x})\ell(g(\mathbf{x}),-1)].\end{aligned}$$

The second equation holds due to a simple observation that all the true negative data are essentially observed as negative, and all observed positive data are indeed true positive.

Therefore, we have

$$\begin{aligned}&\frac{\Pr[\mathbf{x}|y=-1]}{\Pr[\mathbf{x}|\widehat{y}=-1]} = \frac{\Pr[\widehat{y}=-1]}{\Pr[y=-1]}\cdot\frac{\Pr[\mathbf{x},y=-1]}{\Pr[\mathbf{x},\widehat{y}=-1]}\\ &= \frac{\Pr[\widehat{y}=-1]}{\Pr[y=-1]}\cdot\frac{\Pr[\mathbf{x},y=-1,\widehat{y}=-1]+\overbrace{\Pr[\mathbf{x},y=-1,\widehat{y}=+1]}^{=0}}{\Pr[\mathbf{x},\widehat{y}=-1]}\\ &= \frac{\Pr[\widehat{y}=-1]}{\Pr[y=-1]}\cdot\sigma_-(\mathbf{x}).\end{aligned}$$

A similar result can be obtained for the positive side by an analogous argument. To this end, we complete the proof of Theorem 1. □

Remark 1. Theorem 1 justifies the usefulness of noisy negative data. Instead of discarding noisy data or regarding them as the unlabeled data, a more efficient method should consider the noisy negative data, since they can be used to recover the underlying noise-free distribution, along with clean positive data.

As the underlying distribution of the positive and the noisy negative data is not available, we approximate the risk by the empirical oInAS risk, defined as follows.

Definition 2 (Empirical Risk for one-sided InAccurate Supervision, Empirical oInAS Risk)**.** For any function $g\in\mathcal{G}$, its empirical oInAS risk $\widehat{R}_{os}^{IA}(g)$ is defined as,

$$\begin{aligned}\widehat{R}_{os}^{IA}(g) = &\frac{\pi_{\widetilde{P}}}{n_{\widetilde{P}}}\sum_{i=1}^{n_{\widetilde{P}}}\sigma_+(\mathbf{x}_i)\cdot\ell(g(\mathbf{x}_i),+1)\\ &+\frac{\pi_{\widetilde{N}}}{n_{\widetilde{N}}}\sum_{j=1}^{n_{\widetilde{N}}}\sigma_-(\mathbf{x}_j)\cdot\ell(g(\mathbf{x}_j),-1),\end{aligned}$$

where the weights $\sigma_+(\mathbf{x})$ and $\sigma_-(\mathbf{x})$ are defined in (5).

Denote by $\widehat{g}_{os}^{IA}$ the minimizer of Empirical oInAS Risk, we introduce the following excess risk bound, showing that the risk of $\widehat{g}_{os}^{IA}$ converges to that of the optimal decision function in the function family $\mathcal{G}$.

Theorem 2 (Excess risk of learning from one-sided inaccurate supervision)**.** *Assume that the loss function ℓ is non-negative and L-Lipschitz continuous. With* hardness $h_P(\mathbf{x})\in[0,h]$*, then, for any $\delta>0$, with probability at least $1-\delta$, we have*

$$\begin{aligned}R(\widehat{g}_{os}^{IA})-R(g^*) \le &\frac{4\pi_{\widetilde{P}}L}{1-h_P}\mathfrak{R}_{n_{\widetilde{P}}}(\mathcal{G})+4\pi_{\widetilde{N}}L\mathfrak{R}_{n_{\widetilde{N}}}(\mathcal{G})\\ &+2\pi_{\widetilde{P}}\sqrt{\frac{\ln(4/\delta)}{2n_{\widetilde{P}}}}+2\pi_{\widetilde{N}}\sqrt{\frac{\ln(4/\delta)}{2n_{\widetilde{N}}}},\end{aligned}$$

where $\mathfrak{R}_{n_{\widetilde{P}}}(\mathcal{G})$ is Rademacher complexity of $\mathcal{G}$ for the sample of size $n_{\widetilde{P}}$ from $p_+=\Pr[\mathbf{x}|\widehat{y}=+1]$ and $\mathfrak{R}_{n_{\widetilde{N}}}(\mathcal{G})$ follows a similar definition over the observed negative data. Detailed proofs are presented in the supplemental material.

Remark 2. In Theorem 2, the uniform boundedness of hardness is necessary; otherwise, the excess risk can be unbounded. When the hardness h_P is very close to 1, there exist some instances whose true labels are positive but are regarded as negative with probability close to 1. As only $\widetilde{P}$ instead of the original set P is accessible, we cannot recover the information of those extremely hard examples.

Remark 3. When it degenerates to the *instance-independent* label noise scenario, namely, there exists a constant noise rate $h_P=\Pr[\widehat{y}=-1|\mathbf{x},y=+1]=\Pr[\widehat{y}=-1|y=+1]$, our algorithm recovers the importance reweighting method proposed in [20]. Specifically, we set the instance-dependent label noise as a constant, and thereby recover their method.

4.2 Estimating σ_+ and σ_- via Incomplete Supervision

In the oInAS risk, it is crucial to estimate the weights σ_+ and σ_- defined in (5). A direct weighting technique for inaccurate supervision is also adopted in [20], but their method only handles instance-independent label noise and is not able to utilize unlabeled data. However, in the semi-supervised learning scenario, labeled data are limited, while unlabeled data are comparatively sufficient. It is very desired to exploit unlabeled data when we only have scarce noisy labeled data on hand. In this paragraph, we estimate the weights $\sigma_{+/-}$ with the help of incomplete supervision.

As shown in [9], positive-unlabeled learning is provably better than supervised learning in terms of risk bounds when infinite unlabeled data are available. Therefore, provided with sufficient unlabeled data, the classifier learned from positive and unlabeled data also has a good capability in estimating the underlying noise-free distribution. Consequently, we employ $\widehat{g}_{PU}$, the minimizer of empirical PU risk in (3), to produce pseudo labels for the noisy negative data and unlabeled data, which are used to estimate the weights $\sigma_{+/-}$.

Estimating Weights $\sigma_{+/-}$. Firstly, we rewrite the weights $\sigma_+(\mathbf{x})$ and $\sigma_-(\mathbf{x})$ defined in (5) as follows,

$$\begin{aligned}\sigma_+(\mathbf{x}) &= \frac{\Pr[\mathbf{x},y=+1]}{\Pr[\mathbf{x},y=+1,\widehat{y}=+1]+\underbrace{\Pr[\mathbf{x},y=-1,\widehat{y}=+1]}_{=0}}\\ &= \frac{\pi_P\Pr[\mathbf{x}|y=+1]}{\pi_{\widetilde{P}}\Pr[\mathbf{x}|\widehat{y}=+1]} = \frac{\pi_P}{\pi_{\widetilde{P}}}\sigma'_+(\mathbf{x}),\end{aligned}$$

$$\sigma_-(\mathbf{x}) = \frac{\Pr[\mathbf{x}, y=-1, \widehat{y}=-1] + \overbrace{\Pr[\mathbf{x}, y=-1, \widehat{y}=+1]}^{=0}}{\Pr[\mathbf{x}, \widehat{y}=-1]} = \frac{\pi_N \Pr[\mathbf{x}|y=-1]}{\pi_{\widetilde{N}} \Pr[\mathbf{x}|\widehat{y}=-1]} = \frac{\pi_N}{\pi_{\widetilde{N}}}\sigma'_-(\mathbf{x}),$$

where $\pi_{\widetilde{P}/\widetilde{N}}$ and $\pi_{P/N}$ are class-priors of (noisy) positive/negative data and $\sigma'_{+/-}(\mathbf{x})$ denote the remaining density ratio terms, which are defined as

$$\sigma'_+(\mathbf{x}) = \Pr[\mathbf{x}|y=+1]/\Pr[\mathbf{x}|\widehat{y}=+1],$$
$$\sigma'_-(\mathbf{x}) = \Pr[\mathbf{x}|y=-1]/\Pr[\mathbf{x}|\widehat{y}=-1].$$

In the following, we provide the estimation procedure of $\sigma_+(\mathbf{x})$, and the estimator of $\sigma_-(\mathbf{x})$ can be similarly obtained. Based on the law of large numbers, π_P and $\pi_{\widetilde{P}}$ can be estimated by the ratio of the number of samples as

$$\widehat{\pi}_P = \frac{n_{y_{PU}=+1}}{n_{\widetilde{P}} + n_{\widetilde{N}}}, \qquad \widehat{\pi}_{\widetilde{P}} = \frac{n_{\widehat{y}=+1}}{n_{\widetilde{P}} + n_{\widetilde{N}}},$$

in which $n_{y_{PU}=+1}$ denotes the number of positive data estimated by empirical PU classifier $\widehat{g}_{PU}$ while $n_{\widehat{y}=+1}$ denotes the number of observed positive data.

Then, in the incomplete and one-sided inaccurate supervision learning scenario, we estimate ratio σ'_+ with the help of the learned classifier $\widehat{g}_{PU}$ over unlabeled data. We measure the discrepancy between estimated ratio and the true ratio by the Bregman divergence, defined as follows.

Definition 3 (Bregman divergence of ratio models [41])**.** Assume that the function $f : \mathbb{R} \mapsto \mathbb{R}$ is differentiable and strictly convex. Let $\nabla f(\mathbf{x})$ denote the subgradient of $f(\mathbf{x})$, the Bregman divergence associated with f from the true density ratio σ'_+ to the estimated ratio $\widehat{\sigma}'_+$ is defined as,

$$\begin{aligned} B_f(\sigma'_+\|\widehat{\sigma}'_+) = &\int \Pr[\mathbf{x}|\widehat{y}=+1]\nabla f(\widehat{\sigma}'_+(\mathbf{x}))\widehat{\sigma}'_+(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &- \int \Pr[\mathbf{x}|\widehat{y}=+1] f(\widehat{\sigma}'_+(\mathbf{x}))\,\mathrm{d}\mathbf{x} \\ &- \int \Pr[\mathbf{x}|y=+1]\nabla f(\widehat{\sigma}'_+(\mathbf{x}))\,\mathrm{d}\mathbf{x}. \end{aligned}$$

Denote by $P_{PU} = \{(\mathbf{x}_i, \widehat{g}_{PU}(\mathbf{x}_i) = +1)\}_{i=1,\ldots,m}$ the set of instances that are labeled as $+1$ by $\widehat{g}_{PU}$ of size m, which approximates the sampled instances generated from $\Pr[\mathbf{x}|y=1]$. As $\widetilde{P}$ is directly sampled from $\Pr[\mathbf{x}|\widehat{y}=1]$, we estimate the empirical Bregman divergence $\widehat{B}_f^{PU}(\sigma'_+\|\widehat{\sigma}'_+)$ of estimated ratio and the true ratio by

$$\begin{aligned} \widehat{B}_f^{PU}(\sigma'_+\|\widehat{\sigma}'_+) = &\frac{1}{n_{\widetilde{P}}}\sum_{i=1}^{n_{\widetilde{P}}} \nabla f(\widehat{\sigma}'_+(\mathbf{x}_i))\widehat{\sigma}'_+(\mathbf{x}_i) \\ &- \frac{1}{n_{\widetilde{P}}}\sum_{i=1}^{n_{\widetilde{P}}} f(\widehat{\sigma}'_+(\mathbf{x}_i)) - \frac{1}{m}\sum_{j=1}^{m}\nabla f(\widehat{\sigma}'_+(\mathbf{x}_j)). \end{aligned}$$

Therefore, provided with two sets of instances sampled from the observed positive data and the pseudo positive data, namely $\widetilde{P}$ and P_{PU}, we are able to approximate the true density ratio by minimizing the empirical Bregman divergence. We denote by $\widehat{\sigma}'^{PU}_+$ the minimizer of the empirical Bregman divergence of function family $\{\widehat{\sigma}'_+\}$, that is,

$$\widehat{\sigma}'^{PU}_+ = \operatorname*{arg\,min}_{\widehat{\sigma}'_+ \in \{\widehat{\sigma}'_+\}} \widehat{B}_f^{PU}(\sigma'_+(\mathbf{x})\|\widehat{\sigma}'_+(\mathbf{x})).$$

We provide the following bound to show that the estimated ratio converges to the optimal density ratio in the function family $\{\widehat{\sigma}'_+\}$.

Theorem 3. *Assume that $\sigma'_+(\mathbf{x})$ is bounded. Then, for any $\delta > 0$, the following bound holds with probability at least $1-\delta$,*

$$B_f(\sigma'_+\|\widehat{\sigma}'^{PU}_+) \le 2C\mathfrak{R}(\{\widehat{\sigma}'_+\}) + b\sqrt{\frac{\log(4/\delta)}{2n_{\widetilde{P}}}},$$

where $\mathfrak{R}(\{\widehat{\sigma}'_+\})$ is the Rademacher complexity of ratio model set, in the order of $\mathcal{O}(1/\sqrt{n_{\widetilde{P}}})$; C and b are constants. Detailed proofs are provided in the supplemental material.

Theorem 3 guarantees that our estimated weights $\sigma_{+/-}$ converge to the optimal one in the hypothesis space, in the order of $\mathcal{O}(1/\sqrt{n_{\widetilde{P}}})$. This analysis accords to the intuition that the estimator will be more accurate with more clean positive data available.

4.3 Our Approach

In order to learn from incomplete and one-sided inaccurate supervision, we minimize the weighted combination of oInAS risk and PU risk (which we denoted by $R_{os}^{IC}(g)$),

$$\begin{aligned} &\gamma R_{os}^{IA}(g) + (1-\gamma)R_{os}^{IC}(g) \\ =&\gamma\pi_{\widetilde{P}}\mathbb{E}_{\widetilde{P}}[\sigma_+(\mathbf{x})\ell(g(\mathbf{x}),+1)] + \gamma\pi_{\widetilde{N}}\mathbb{E}_{\widetilde{N}}[\sigma_-(\mathbf{x})\ell(g(\mathbf{x}),-1)] \\ &+ 2(1-\gamma)\pi_P\mathbb{E}_P[\ell(g(\mathbf{x}),+1)] + (1-\gamma)\mathbb{E}_U[\ell(g(\mathbf{x}),-1)], \end{aligned}$$

where $\gamma \in [0,1]$ is the trade-off coefficient. As the classifier $\widehat{g}_{PU}$ is required to provide pseudo-labels for negative and unlabeled data, we split the positive data $\widetilde{P}$ into two disjoint sets $\widetilde{P}_1$ and $\widetilde{P}_2$ of size $n_{\widetilde{P}_1}$ and $n_{\widetilde{P}_2}$, which are respectively adopted in the (empirical) oInAS and PU risk,

$$\begin{aligned} \widehat{R}_{os}(g) = &\frac{\gamma\widehat{\pi}_{\widetilde{P}}}{n_{\widetilde{P}_1}}\sum_{i=1}^{n_{\widetilde{P}_1}}\sigma_+(\mathbf{x}_i)\ell(g(\mathbf{x}_i),+1) + \frac{\gamma\widehat{\pi}_{\widetilde{N}}}{n_{\widetilde{N}}}\sum_{j=1}^{n_{\widetilde{N}}}\sigma_-(\mathbf{x}_j)\ell(g(\mathbf{x}_j),-1) \\ &+ \frac{2(1-\gamma)\widehat{\pi}_P}{n_{\widetilde{P}_2}}\sum_{i=1}^{n_{\widetilde{P}_2}}\ell(g(\mathbf{x}_i),+1) + \frac{(1-\gamma)}{n_U}\sum_{k=1}^{n_U}\ell(g(\mathbf{x}_k),-1). \end{aligned} \tag{6}$$

Denote by $\widehat{g}_{os}$ the minimizer of empirical risk of incomplete and one-sided inaccurate supervision in (6). For the learned decision function, we have the following excess risk bound, demonstrating that the risk of $\widehat{g}_{os}$ converges to that of optimal decision function in $\mathcal{G}$.

Corollary 1 (Excess Risk of LIISP(os))**.** *Assume that the loss function ℓ is bounded, non-negative and L-Lipschitz continuous. Suppose the hardness $h_P(\mathbf{x}) \le h$ holds uniformly for each instance, and there is a constant $C_{\mathcal{G}} > 0$ such that $\mathfrak{R}_n(\mathcal{G}) \le C_{\mathcal{G}}/\sqrt{n}$ for positive/noisy negative and unlabeled data (with $n = n_{\widetilde{P}}/n_{\widetilde{N}}/n_U$). Then for any $\delta > 0$, with probability at least $1-\delta$, we have*

$$R(\widehat{g}_{os}) - R(g^*) \le \mathcal{O}(1/\sqrt{n_{\widetilde{P}}} + 1/\sqrt{n_{\widetilde{N}}} + 1/\sqrt{n_U}).$$

Remark 4. Corollary 1 implies the usefulness of leveraging unlabeled data to alleviate the instance-dependent label noise. As we can see, the risk bound is tighter with an increasing number of unlabeled data. The above risk bound is in optimal convergence rate without any additional assumption [42].

5 Learning from Incomplete and Inaccurate Supervision with Class-Priors

In real-world applications, both positive and negative data could be polluted by the instance-dependent label noise. In such a learning scenario, it is extremely hard to retrieve the underly true risk and design an algorithm with theoretical guarantee since there are no reliable label information. Therefore, we proactively collect additional class-prior information of the unlabeled data, namely, we have two discrepant unlabeled datasets with all necessarily known class-priors. We demonstrate that these discrepant incomplete supervisions could help to resist the two-sided label noise.

To deal with the two-sided noisy labels, we also rewrite the expect risk in the importance weighting form, and then estimate these weights with the help of the discrepant unlabeled datasets. Finally, we provide our learning algorithm for incomplete and inaccurate supervision with class-priors.

5.1 Learning from Inaccurate Supervision with Class-Priors

Notations and Settings. Throughout this section, we assume that we have $n_{\widetilde{P}}$ noisy positive data which is denoted by $\widetilde{P} = \{(\mathbf{x}_i, +1)\}_{i=1,\dots,n_{\widetilde{P}}}$ and $n_{\widetilde{N}}$ noisy negative data denoted by $\widetilde{N} = \{(\mathbf{x}_j, -1)\}_{j=1,\dots,n_{\widetilde{N}}}$. Let $\pi_{\widetilde{P}}$ be the class-prior of the observed positive labels $\Pr[\widehat{y} = +1]$ and $\pi_{\widetilde{N}}$ be the one of observed negative labels $\Pr[\widehat{y} = -1]$ with $\pi_{\widetilde{P}} + \pi_{\widetilde{N}} = 1$. We denote by π_P and π_N the class-priors of underlying true data, which are known ahead in this scenario.

As previously assumed, we suppose that the observed data are with instance-dependent label noise. Following the definition of h_P, we define the hardness h_N on underlying true negative data, formally, $h_N : \mathcal{X} \mapsto [0, 1]$, with,

$$h_N(\mathbf{x}) = \Pr[\widehat{y} = +1 | \mathbf{x}, y = -1].$$

Rewrite True Risk. When both positive and negative data are with instance-dependent noisy labels, we propose the *InAS Risk* for the InAccurate Supervision, and show that it is provably equal to the underlying true risk.

Definition 4 (Risk for InAccurate Supervision (InAS Risk))**.** For any function $g \in \mathcal{G}$, given the class-priors π_P and π_N, its InAS risk $R_{ts}^{IA}(g)$ is defined as,

$$\begin{aligned} R_{ts}^{IA}(g) = \pi_P \mathbb{E}_{\widetilde{P}} \left[\sigma'_+(\mathbf{x}) \ell(g(\mathbf{x}), +1)\right] \\ + \pi_N \mathbb{E}_{\widetilde{N}} \left[\sigma'_-(\mathbf{x}) \ell(g(\mathbf{x}), -1)\right], \end{aligned}$$

where weights $\sigma'_+(\mathbf{x})$ and $\sigma'_-(\mathbf{x})$ are defined as

$$\begin{aligned} \sigma'_+(\mathbf{x}) &= \Pr[\mathbf{x}|y = +1] / \Pr[\mathbf{x}|\widehat{y} = +1], \\ \sigma'_-(\mathbf{x}) &= \Pr[\mathbf{x}|y = -1] / \Pr[\mathbf{x}|\widehat{y} = -1]. \end{aligned} \tag{7}$$

Then, we demonstrate that the InAS risk equals to the true risk over the underlying distribution $\mathcal{D}$.

Theorem 4. *The InAS risk equals to the true risk (the risk over the true data distribution), that is,*

$$R_{ts}^{IA}(g) = R(g).$$

Proof. For the expectation over the marginal distribution of the clean positive data, we rewrite it as

$$\begin{aligned} &\mathbb{E}_P \left[\ell(g(\mathbf{x}), +1)\right] \\ &= \int \ell(g(\mathbf{x}), +1) \Pr[\mathbf{x}|y = +1] \, \mathrm{d}\mathbf{x} \\ &= \int \ell(g(\mathbf{x}), +1) \Pr[\mathbf{x}|\widehat{y} = +1] \frac{\Pr[\mathbf{x}|y = +1]}{\Pr[\mathbf{x}|\widehat{y} = +1]} \, \mathrm{d}\mathbf{x} \\ &= \mathbb{E}_{\widetilde{P}} \left[\sigma'_+(\mathbf{x}) \ell(g(\mathbf{x}), +1)\right]. \end{aligned}$$

While for the negative data, a similar result can be obtained by an analogous argument. □

Remark 5. Theorem 4 demonstrates that the true risk can be retrieved by assigning proper weights to each noisy instance. Therefore, a well-generalized classifier can be obtained by estimating these weights and then minimizing the weighted empirical risk. However, as defined in (7), we should approximate the underlying true conditional distribution like the $\Pr[\mathbf{x}|y = +1]$. There is no hope to learn a good estimation for these weights when the noisy data are scarce and occur arbitrarily. This observation motivates us to handle this task with the help of unlabeled data, which is often with a large amount and rather easy to obtain.

As we only have the sampled data, we approximate the true risk by the empirical InAS risk, which is defined as

Definition 5 (Empirical Risk for InAccurate Supervision, Empirical InAS Risk)**.** For any function $g \in \mathcal{G}$, its empirical InAS risk $\widehat{R}_{ts}^{IA}(g)$ is defined as

$$\begin{aligned} \widehat{R}_{ts}^{IA}(g) = \frac{\pi_P}{n_{\widetilde{P}}} \sum_{i=1}^{n_{\widetilde{P}}} \sigma'_+(\mathbf{x}_i) \cdot \ell(g(\mathbf{x}_i), +1) \\ + \frac{\pi_N}{n_{\widetilde{N}}} \sum_{j=1}^{n_{\widetilde{N}}} \sigma'_-(\mathbf{x}_j) \cdot \ell(g(\mathbf{x}_j), -1), \end{aligned}$$

where the weights $\sigma'_+(\mathbf{x})$ and $\sigma'_-(\mathbf{x})$ are defined in (7).

Denote by $\widehat{g}_{ts}^{IA}$ the minimizer of above empirical InAS risk, we then show that this obtained classifier enjoys the following excess risk bound.

Theorem 5 (Risk of learning from Inaccurate Supervision)**.** *Assume that the loss function ℓ is non-negative and L-Lipschitz continuous. Suppose that the hardness $h_P(\mathbf{x}) \in [0, h_P]$ and $h_N(\mathbf{x}) \in [0, h_N]$, for any $\delta > 0$, with probability at least $1 - \delta$, we have*

$$\begin{aligned} R(\widehat{g}_{ts}^{IA}) - R(g^*) \leq \frac{4\pi_P L}{1 - h_P} \mathfrak{R}_{n_{\widetilde{P}}}(\mathcal{G}) + \frac{4\pi_N L}{1 - h_N} \mathfrak{R}_{n_{\widetilde{N}}}(\mathcal{G}) \\ + 2\pi_P \sqrt{\frac{\ln(4/\delta)}{2n_{\widetilde{P}}}} + 2\pi_N \sqrt{\frac{\ln(4/\delta)}{2n_{\widetilde{N}}}}, \end{aligned}$$

where $\mathfrak{R}_{n_{\widetilde{P}}}(\mathcal{G})$ is Rademacher complexity of function family $\mathcal{G}$ for the sampling of size $n_{\widetilde{P}}$ from $\Pr[\mathbf{x}|\widehat{y} = +1]$ and $\mathfrak{R}_{n_{\widetilde{N}}}(\mathcal{G})$ follows a similar definition. Detailed proofs are presented in the supplemental material.

5.2 Estimating σ'_+ and σ'_- by Discrepant Incomplete Supervisions

We proceed to estimate the weights σ'_+ and σ'_- defined in (7) with the help of incomplete supervisions. As shown in [11],

the risk of underlying true distribution can be retrieved by discrepant unlabeled datasets, when their corresponding class-priors θ_P and θ'_P and the class-prior π_P of the joint unlabeled datasets are known in advance. Therefore, with sufficient unlabeled data on hand, the classifier learned from the discrepant unlabeled datasets is valuable in estimating the underlying noise-free distribution. Accordingly, we employ $\widehat{g}_{UU}$, the minimizer of empirical UU risk in (4), to produce pseudo labels, which are then used to estimate the weights $\sigma'_{+/-}(\mathbf{x})$. In the following, we propose the estimation of $\sigma'_{+}(\mathbf{x})$, while $\sigma'_{-}(\mathbf{x})$ can be similarly obtained.

Estimating Weights $\sigma'_{+/-}$. Again we adopt the ratio matching method proposed in [41] to estimate weights for observed labeled data. Let $P_{UU} = \{(\mathbf{x}_i, \widehat{g}_{UU}(\mathbf{x}_i) = +1)\}_{i=1,\ldots,n_{U+}}$ be the set of instances labeled as $+1$ by $\widehat{g}_{UU}$ of size n_{U+} in training data, which used to approximate the instances sampled from $\Pr[\mathbf{x}|y=1]$. We approximate the Bregman divergence between estimated ratio and the true ratio by the empirical one, namely,

$$\widehat{B}^{UU}_f(\sigma'_+\|\widehat{\sigma}'_+) = \frac{1}{n_{\widetilde{P}}}\sum_{i=1}^{n_{\widetilde{P}}}\nabla f(\widehat{\sigma}'_+(\mathbf{x}_i))\widehat{\sigma}'_+(\mathbf{x}_i) - \frac{1}{n_{\widetilde{P}}}\sum_{i=1}^{n_{\widetilde{P}}} f(\widehat{\sigma}'_+(\mathbf{x}_i)) - \frac{1}{n_{U_+}}\sum_{j=1}^{n_{U_+}}\nabla f(\widehat{\sigma}'_+(\mathbf{x}_j)).$$

With two sets of instances sampled from the observed positive data and the pseudo positive data, namely $\widetilde{P}$ and P_{UU}, we are able to approximate the true density ratio by minimizing the empirical Bregman divergence. Let $\widehat{\sigma}'^{UU}_+$ be the minimizer of the above empirical Bregman divergence, we provide the following bound to show that the estimated ratio converges to the optimal one in $\{\widehat{\sigma}'_+\}$.

Theorem 6. *Assume that $\sigma'_+(\mathbf{x})$ is bounded. Let $n = \min\{n_{\widetilde{P}}, n_{\widetilde{N}}, n_{U_1}, n_{U_2}\}$, for any $\delta > 0$, the following bound holds with probability at least $1-\delta$,*

$$B_f(\sigma'_+\|\widehat{\sigma}'^{UU}_+) \leq 2C\mathfrak{R}(\{\widehat{\sigma}'_+\}) + b\sqrt{\frac{\log(4/\delta)}{2n}},$$

where $\mathfrak{R}(\{\widehat{\sigma}'_+\})$ is the Rademacher complexity of ratio model set, in the order of $\mathcal{O}(1/\sqrt{n})$; C and b are constants. Detailed proofs are provided in the supplemental material.

Theorem 6 guarantees that our estimated weights $\sigma'_{+/-}$ converge to the optimal one in the hypothesis space, in the order of $\mathcal{O}(1/\sqrt{n})$. This analysis accords to the intuition that the estimator will be more accurate with more data available.

5.3 Our Approach

To exploit noisy data and unlabeled data simultaneously, we introduce the weighted combination of the InAS risk and the UU risk (which we denoted by $R^{IC}_{ts}(g)$), to learn from the incomplete and inaccurate supervision with class-priors, namely,

$$\begin{aligned}&\gamma R^{IA}_{ts}(g) + (1-\gamma)R^{IC}_{ts}(g)\\ &= \gamma(\pi_P\mathbb{E}_{\widetilde{P}}[\sigma'_+(\mathbf{x})\ell(g(\mathbf{x}),+1)] + \pi_N\mathbb{E}_{\widetilde{N}}[\sigma'_-(\mathbf{x})\ell(g(\mathbf{x}),-1)])\\ &\quad + (1-\gamma)(\mathbb{E}_{U_1}[\bar{\ell}_{U_1}(g(\mathbf{x}),+1)] + \mathbb{E}_{U_2}[\bar{\ell}_{U_2}(g(\mathbf{x}),-1)]),\end{aligned}$$

where $\gamma \in [0,1]$ is the trade-off coefficient. The empirical version $\widehat{R}_{cp}(g)$ is defined as

$$\begin{aligned}\widehat{R}_{cp}(g) &= \gamma\widehat{R}^{IA}_{ts}(g) + (1-\gamma)\widehat{R}_{UU}(g)\\ &= \frac{\gamma\pi_P}{n_{\widetilde{P}}}\sum_{i=1}^{n_{\widetilde{P}}}\sigma'_+(\mathbf{x}_i)\ell(g(\mathbf{x}_i),+1) + \frac{\gamma\pi_N}{n_{\widetilde{N}}}\sum_{j=1}^{n_{\widetilde{N}}}\sigma'_-(\mathbf{x}_j)\ell(g(\mathbf{x}_j),-1)\\ &\quad + \frac{1-\gamma}{n_{U_1}}\sum_{i=1}^{n_{U_1}}\alpha\ell(g(\mathbf{x}_i),+1) + \frac{1-\gamma}{n_{U_2}}\sum_{j=1}^{n_{U_2}}\alpha'\ell(g(\mathbf{x}_i),-1),\end{aligned}$$

where $\alpha = (\theta' + \pi_P - 2\theta'\pi_P)/(\theta - \theta')$ and $\alpha' = (\theta + \pi_P - 2\theta\pi_P)/(\theta - \theta')$.

Let $\widehat{g}_{cp}$ be the minimizer of the weighted combination risk $\widehat{R}_{cp}(g)$ in the function family $\mathcal{G}$, we have the following excess risk bound, demonstrating that the risk of $\widehat{g}_{cp}$ converges to that of optimal decision function in $\mathcal{G}$.

Corollary 2 (Excess Risk of LIISP(cp))**.** *Assume that the bounded loss function ℓ is non-negative and L-Lipschitz continuous. Suppose the hardness $h_P(\mathbf{x}), h_N(\mathbf{x}) \leq h$ holds uniformly for each instance, and there is a constant $C_{\mathcal{G}} > 0$ such that $\mathfrak{R}_n(\mathcal{G}) \leq C_{\mathcal{G}}/\sqrt{n}$ for positive/noisy negative and unlabeled data (with $n = n_{\widetilde{P}}/n_{\widetilde{N}}/n_U$). Then for any $\delta > 0$, with probability at least $1-\delta$, we have*

$$R(\widehat{g}_{cp}) - R(g^*) \leq \mathcal{O}(1/\sqrt{n_{\widetilde{P}}} + 1/\sqrt{n_{\widetilde{N}}} + 1/\sqrt{n_U}).$$

6 Experiment

In this section, we examine the performance of the proposed LIISP(os) and LIISP(cp) algorithms on both benchmark datasets and real-world tasks. Specifically, we evaluate our algorithms in the following three aspects:

(i) **Comparisons on Synthetic Datasets:** we provide intuitive illustrations on the advantage of our approaches against traditional algorithms designed for only incomplete or only inaccurate supervision;

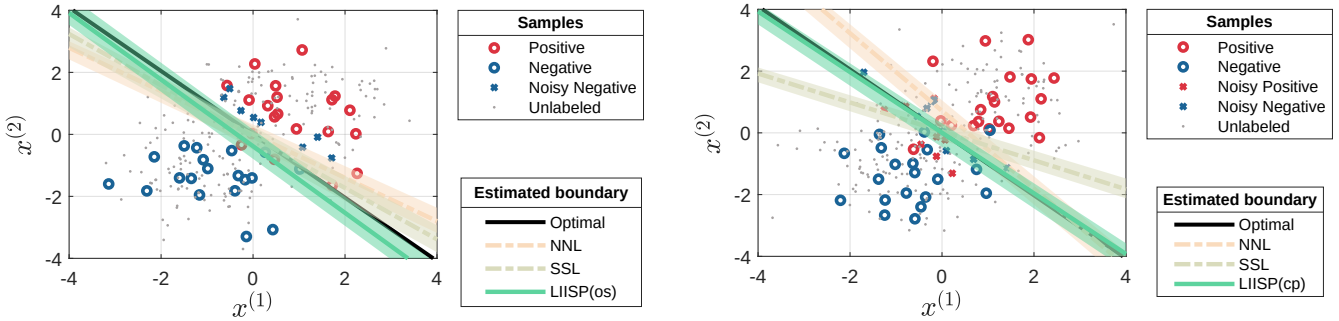
(ii) **Comparisons on Benchmark Datasets:** we compare the LIISP algorithms with robust SSL methods on benchmark datasets, to demonstrate the superiority of the LIISP algorithms in exploiting incomplete and inaccurate supervision, and the usefulness of noisy labeled data and the unlabeled data;

(iii) **Bug Detection Task:** we validate the effectiveness of the LIISP(os) algorithm on the bug detection task, which aims at detecting defects in software systems.

In order to simulate the *instance-dependent* label noise, we first pre-train a SVM classifier on clean data and flip 20% positive data into negative according to their confidence. For the LIISP(cp) algorithm, we additionally flip 20% negative data into positive to imitate the two-sided label noise. We perform experiments 10 times on various splits of datasets, and present the average as well as the standard deviation of the results. We also conduct 10-fold cross validation to choose a proper trade-off coefficient γ.

6.1 Comparisons on Synthetic Datasets

We first numerically illustrate the effectiveness of the LIISP algorithms under incomplete and inaccurate supervision. We



(a) Decision boundaries and error bounds of NLL, SSL and LIISP(os) (b) Decision boundaries and error bounds of NLL, SSL and LIISP(cp)

Fig. 1: Comparisons with Noisy Label Learning (NLL) algorithm and Semi-Supervised Learning (SSL) algorithm on the synthetic dataset generated by two-dimensional Normal distributions with instance-dependent label noise.

first generate a synthetic dataset as underlying true distribution from two class-conditional distributions, with each instance $(\mathbf{x}, y)$ generated from standard two-dimensional Normal distribution $\mathcal{N}_{\mathbf{x}}$ according to

$$\Pr[\mathbf{x}|y=-1] = \mathcal{N}_{\mathbf{x}}([-1,-1]), \Pr[\mathbf{x}|y=1] = \mathcal{N}_{\mathbf{x}}([1,1]).$$

Then we generate the instance-dependent label noise according to their confidence assigned by a pre-trained SVM model. In the synthetic comparisons, we generate 200 noisy labeled data. Apart from the noisy data, we also provide 1000 unlabeled data as incomplete supervision. The optimal boundary is shown in the solid black line. The NLL method denotes the one that learning only with noisy labeled data, and here we apply a robust SVM [43]. Similarly, the SSL approach denotes the method that learning with unlabeled data, and we apply the PNU algorithm [10] for comparison.

We report the boundaries as well as the error bounds returned by the LIISP, NLL, SSL algorithms in Figure 1. Both the NLL and SSL approaches suffer from the scare noisy labeled data. The green area in Figure 1(a) denotes the boundary and the error bound of the LIISP(os) algorithm, which is closest to the optimal boundary. A similar result can be obtained for the LIISP(cp) algorithm, which is shown in Figure 1(b). To conclude, our proposed LIISP(os/cp) algorithms could empirically approximate the optimal boundaries under incomplete and inaccurate supervision.

6.2 Comparisons on Benchmark Datasets

In this part, we examine the performance of the LIISP(os/cp) algorithms on benchmark datasets and test the usefulness of both noisy labeled data and unlabeled data. We notice that the LIISP(os) algorithm and LIISP(cp) algorithm deal with different learning scenarios, thus they are not directly comparable. We conduct the benchmark comparisons on the UCI datasets[2] and the LIBSVM datasets[3], including diabetes, breastw, wdbc, house, letter7vs9, ionosphere, australian, isolet, german, a5a, clean1, mnist7vs9, autavn and rcv1 from various fields. The number of data items varies from 232 to 20,242 and their dimension varies from 8 to 47,236. We summarize their brief statistical information in Table 1.

2. The UCI datasets can be downloaded from https://archive.ics.uci.edu/ml/datasets.php with detailed description for each dataset.

3. The LIBSVM datasets can be downloaded from https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/ with detailed description for each dataset.

TABLE 1: Brief statistics of benchmark datasets

Dataset	# Instance	# Dim	Dataset	# Instance	# Dim
diabetes	768	8	isolet	600	51
breastw	683	9	german	1,000	59
wdbc	569	14	a5a	6,414	122
house	232	16	clean1	476	166
letter7vs9	1,528	16	mnist7vs9	14,251	784
ionosphere	351	33	autavn	7,118	20,707
australian	690	42	rcv1	20,242	47,236

We compare the proposed LIISP(os/cp) algorithms with six contenders, including two supervised learning methods and four semi-supervised learning algorithms. The two supervised learning baselines:

- LIBSVM [44] is an SVM baseline.
- IW [20] is a noisy label learning approach which resist the label noise by importance reweighting technique.

There are other four robust semi-supervised learning algorithms, which consider the noisy labels in SSL,

- LSSC [38] is a sparse coding based SSL method. It gives a L_1-norm formulation of Laplacian regularization based on the manifold structure of the data.
- ROSSEL [39] generates a set of pseudo labels for unlabeled data, and approximates the ground-truth labels by multiple label kernel learning.
- SIIS [40] is a graph-based SSL algorithm. It emphasizes the leading eigenvectors of the Laplacian matrix associated with small eigenvalues, such that this method constructs a label noise robust graph and propagates labels on this graph.
- SAFEW [45] builds the final prediction results by integrating several weakly supervised learners on noisy labeled data and makes it never worse than a simple supervised learning baseline.

In addition, we also include the following two methods into comparisons, which are direct combinations of the NLL and SSL approaches, in order to demonstrate the superiority of our algorithm to these naive combinations.

- PUIW is a direct combination of PU learning and IW, which first adopts PU learning to generate pseudo labels for unlabeled data, and then applies IW on labeled data to alleviate the effect of noisy labels.
- UUIW is a straightforward combination of UU learning and IW, whose procedures are similar to PUIW

TABLE 2: Performance comparisons on benchmark datasets. On each dataset, 10 test runs were conducted and the average accuracy as well as standard deviation are presented, with the best one emphasized in bold. Besides, • (◦) indicates our approach is significantly better (worse) than the compared method (paired t-tests at 95% significance level).

(a) Learning from incomplete and one-sided inaccurate supervision

Dataset	LIBSVM	IW	LSSC	ROSSEL	SIIS	SAFEW	PUIW	LIISP(os)
diabetes	74.91 ± 1.50 •	60.79 ± 8.81 •	68.45 ± 2.35 •	75.69 ± 2.48	67.92 ± 1.37 •	67.69 ± 0.90 •	75.93 ± 2.35	**76.26 ± 1.03**
breastw	93.65 ± 1.98	94.71 ± 1.23	96.21 ± 1.29	**96.53 ± 0.83**	96.49 ± 0.68	96.31 ± 0.49	95.53 ± 1.52	94.85 ± 4.32
wdbc	89.65 ± 2.75 •	77.47 ± 19.3 •	91.97 ± 2.01 •	90.87 ± 2.07 •	92.95 ± 1.32 •	86.46 ± 1.14 •	77.64 ± 12.1 •	**95.52 ± 1.08**
house	91.90 ± 1.72 •	**96.29 ± 1.27**	93.49 ± 2.17 •	93.26 ± 1.88 •	88.84 ± 2.70 •	95.47 ± 1.59	94.53 ± 2.80	96.03 ± 0.97
letter7vs9	90.04 ± 3.88 •	95.21 ± 1.72 •	94.23 ± 0.88 •	94.94 ± 1.43 •	78.47 ± 1.39 •	93.97 ± 0.90 •	95.04 ± 1.34 •	**98.82 ± 0.95**
ionosphere	81.54 ± 3.19 •	83.28 ± 6.51 •	79.26 ± 6.88 •	88.23 ± 4.64	72.11 ± 16.6 •	81.31 ± 3.92 •	85.69 ± 2.56	**90.23 ± 7.43**
australian	80.20 ± 3.24 •	80.65 ± 12.9	81.87 ± 2.81 •	79.89 ± 6.92 •	72.96 ± 3.50 •	71.77 ± 7.89 •	84.47 ± 3.90	**86.19 ± 1.05**
isolet	86.50 ± 2.16 •	91.63 ± 2.44 •	96.48 ± 1.25	80.13 ± 2.06 •	**98.82 ± 0.48**	96.16 ± 2.00	91.74 ± 2.38 •	98.61 ± 1.21
german	64.52 ± 3.89 •	67.45 ± 4.81 •	62.53 ± 1.86 •	73.03 ± 0.96	72.24 ± 1.19	72.73 ± 0.68	68.65 ± 2.21 •	**74.37 ± 2.58**
a5a	70.91 ± 2.42 •	73.82 ± 4.35 •	68.45 ± 1.69 •	79.36 ± 1.66 •	76.36 ± 0.82 •	78.20 ± 1.58	74.13 ± 2.47 •	**83.29 ± 0.47**
clean1	72.84 ± 3.81 •	64.52 ± 4.15 •	61.03 ± 1.00 •	77.40 ± 2.37 •	60.89 ± 5.98 •	69.11 ± 1.38 •	71.07 ± 3.92 •	**86.16 ± 3.42**
mnist7vs9	85.63 ± 2.29 •	90.18 ± 1.62 •	88.76 ± 1.43 •	81.41 ± 1.18 •	92.46 ± 1.73	85.10 ± 7.08 •	91.82 ± 1.53 •	**96.19 ± 0.33**
autavn	65.46 ± 0.82 •	65.59 ± 6.33 •	76.54 ± 7.68	69.54 ± 0.82 •	65.75 ± 2.20 •	72.76 ± 0.21	68.90 ± 1.57 •	**77.72 ± 4.56**
rcv1	62.57 ± 3.79 •	61.52 ± 1.20 •	**70.92 ± 5.55**	68.72 ± 3.59	62.42 ± 0.64 •	70.15 ± 0.88	67.33 ± 2.51	69.93 ± 2.52
LIISP(os) w/ t/ l	13/ 1/ 0	12/ 1/ 1	11/ 1/ 2	10/ 3/ 1	11/ 1/ 2	11/ 1/ 2	10/ 3/ 1	rank first 10/ 14

(b) Learning from incomplete and inaccurate supervision with class-priors

Dataset	LIBSVM	IW	LSSC	ROSSEL	SIIS	SAFEW	UUIW	LIISP(cp)
diabetes	70.55 ± 2.17	60.21 ± 12.7 •	67.37 ± 2.88 •	70.58 ± 1.89	60.64 ± 10.9 •	69.26 ± 6.25 •	66.17 ± 12.8 •	**72.70 ± 4.20**
breastw	94.93 ± 0.77	88.45 ± 12.1 •	95.21 ± 1.04	96.62 ± 0.69	**96.80 ± 4.60** ◦	96.29 ± 0.52	93.22 ± 3.09	95.18 ± 0.57
wdbc	88.03 ± 4.45 •	71.37 ± 14.0 •	90.51 ± 3.51	93.41 ± 0.81	**94.44 ± 1.05** ◦	87.95 ± 4.27	69.69 ± 14.9 •	91.93 ± 1.47
house	92.84 ± 3.91	89.25 ± 15.0 •	90.73 ± 1.47 •	93.26 ± 2.62	89.65 ± 2.48 •	92.56 ± 2.86	90.93 ± 3.33	**94.49 ± 1.99**
letter7vs9	90.72 ± 3.91 •	81.78 ± 20.3 •	92.31 ± 2.40 •	95.16 ± 1.87	77.45 ± 1.68 •	94.40 ± 1.09	85.98 ± 10.7 •	**95.18 ± 3.60**
ionosphere	78.01 ± 3.19 •	74.39 ± 9.34 •	79.49 ± 4.65 •	88.94 ± 2.81	81.82 ± 7.17 •	77.40 ± 7.90 •	72.98 ± 10.2 •	**90.31 ± 4.17**
australian	76.38 ± 4.21 •	66.22 ± 16.1 •	79.45 ± 2.81 •	81.53 ± 2.62 •	73.96 ± 5.18 •	83.43 ± 3.08	72.41 ± 9.91 •	**83.99 ± 1.85**
isolet	73.93 ± 0.59 •	85.83 ± 2.33 •	94.82 ± 0.73	88.47 ± 2.24 •	**98.84 ± 0.33** ◦	97.37 ± 1.03	73.92 ± 18.7 •	95.35 ± 1.76
german	58.33 ± 2.30 •	59.83 ± 8.74 •	61.31 ± 2.28 •	60.24 ± 3.58 •	60.06 ± 10.8 •	63.43 ± 4.27 •	67.77 ± 13.2 •	**72.86 ± 3.38**
a5a	61.36 ± 3.81 •	66.34 ± 8.74 •	67.72 ± 1.86 •	75.25 ± 3.03 •	69.27 ± 15.9 •	71.81 ± 3.93 •	81.71 ± 4.09	**83.13 ± 0.57**
clean1	65.67 ± 2.94 •	59.35 ± 6.64 •	60.65 ± 4.16 •	70.89 ± 3.27	54.56 ± 3.29 •	53.95 ± 4.21 •	61.25 ± 11.7 •	**75.15 ± 6.94**
mnist7vs9	75.84 ± 4.02 •	89.13 ± 1.51 •	87.53 ± 2.18 •	77.47 ± 4.83 •	89.46 ± 1.01	85.39 ± 3.74 •	78.89 ± 20.8 •	**93.29 ± 0.46**
autavn	70.06 ± 9.32 •	65.49 ± 6.04 •	73.76 ± 5.02 •	73.65 ± 4.29 •	66.87 ± 1.16 •	77.46 ± 9.35	67.09 ± 8.45 •	**78.92 ± 8.18**
rcv1	69.65 ± 5.26 •	67.60 ± 1.83 •	74.48 ± 8.12	63.35 ± 3.34 •	60.01 ± 5.67 •	68.56 ± 4.54 •	70.64 ± 2.68 •	**75.80 ± 3.52**
LIISP(cp) w/ t/ l	12/ 2/ 0	14/ 0/ 0	11/ 2/ 1	10/ 2/ 2	11/ 0/ 3	10/ 2/ 2	12/ 2/ 0	rank first 11/ 14

by first using UU learning to generate pseudo labels for unlabeled data and then applying IW method.

Table 2 shows that the LIISP(os/cp) algorithms solve the LIISP tasks and outperform other contenders. In the one-sided label noise scenario, LIISP(os) ranks first in 10 out of 14 datasets in terms of the average accuracy. Overall, the LIISP(os) algorithm outperforms both supervised baselines and robust SSL methods. Compared to the conventional noisy label learning methods (LIBSVM and IW), LIISP(os) achieves higher accuracy and better stability. As shown in Table 2(a), although the IW reweights the noisy label, it is not always better than the LIBSVM baseline. Besides its decline in the average accuracy, the large variance makes it hard to be practical. Such a phenomenon indicates the instability caused by the limited noisy labeled data, and thereby it is essential to utilize the unlabeled data.

Compared with the robust semi-supervised learning approaches, the LIISP(os) algorithm achieves a very promising performance, as it explores the structure of noisy labeled data. Notably, the LIISP(os) algorithm outperforms the ROSSEL approaches, which heavily rely on the performance of the weak learner(s) generated from noisy labeled data. This phenomenon also validates the effectiveness of our approach in utilizing unlabeled data. Compared with the naive combination of PU and IW approaches (PUIW), the LIISP(os) algorithm attains higher accuracy on almost all datasets, which demonstrates that a direct combination of NLL and SSL approaches is not applicable in practice.

In the two-sided label noise scenario, the LIISP(cp) algorithm also outperforms both the supervised baselines and robust SSL methods, which is shown in Table 2(b). The LIISP(cp) algorithm achieves higher accuracy and lower variance compared with two supervised learning baselines on 14 datasets, which demonstrates the benefit of utilizing unlabeled data. We additionally exploit the statistical information (class priors) of the discrepant unlabeled data, so that we can solve the LIISP task with theoretical guarantees. Therefore, it is not surprising that the LIISP(cp) algorithm outperforms the other four robust SSL approaches. Furthermore, the LIISP(cp) also shows superiority over the direct combination of the NLL and SSL methods, namely the UUIW, which verifies the advantage of the proposed LIISP(cp) algorithm.

Comparison with Incomplete Supervision Algorithms. We then implement the LIISP(os/cp) algorithms with deep neural networks and demonstrate their effectiveness. In this part, we focus on validating the usefulness of noisy labeled data. In the following, we empirically compare our proposed algorithms with deep semi-supervised learning approaches (for incomplete supervision) on two real-world datasets.

- MNIST [46] is a large handwritten digits dataset with a training set of 60,000 examples, and a test set of 10,000 examples. These handwritten digits vary from 0 to 9 and we set the even and odd digits as the positive class and negative class, thus construct the binary classification task.
- SVHN [47] is a house number dataset obtained from

TABLE 3: Specification of the benchmark datasets, deep neural network models, and optimization algorithms

Dataset	# Train	# Test	# Feature	Model	Optimizer
MNIST	60,000	10,000	784	MLP	Adam
SVHN	73,257	26,032	150,528	ResNet	Adam

Google Street View images. It contains 10 classes from 0 to 9 and we also set the even and odd digits as the positive class and negative class, respectively.

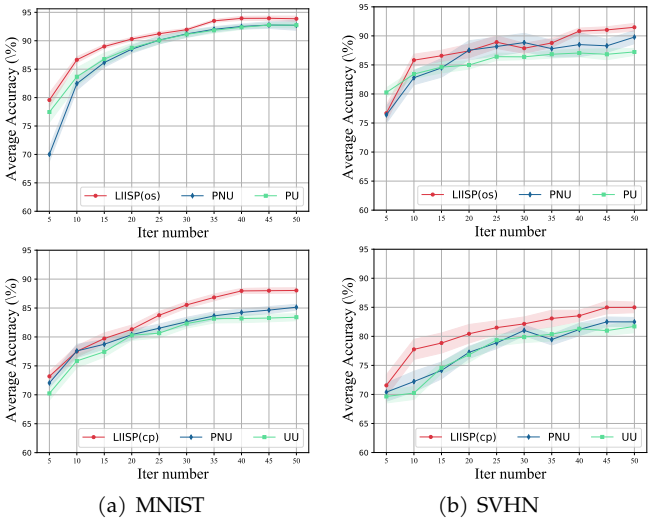
For the MNIST dataset, we apply a 6-layer multilayer perceptron (MLP) with ReLU [48] as activation functions (more specifically, d-300-300-300-300-1, where d is the dimension of training data, 300 is width of the layer). Batch normalization [49] is applied after each fully connected layer. For the SVHN dataset, we apply an pre-trained deep residual network ResNet-18 [50] to generate features and then use a fully connected layer to make the prediction.[4] The architecture of ResNet-18 is: $(224\times224\times3)$-$C(7\times7,64)$-max pool-$[C(3\times3,64)]\times2$-$[C(3\times3,128)]\times2$-$[C(3\times3,256)]\times2$-$[C(3\times3,512)]\times2$-average pool-1000-1, where $(224\times224\times3)$ is the input RGB data with 3 channels, $C(3\times3,64)$ means 64 channels of 3×3 convolutions followed by ReLU, $[\cdot]\times2$ means there are two such layers, etc. The down sampling of the volume though the network is achieved by increasing the stride from 1 to 2. These two models are trained by Adam [51] with a learning rate of 10^{-4} and regularization weight decay 0.005. Table 3 summarizes brief statistics of datasets and used models.

We compare the LIISP(os/cp) algorithms with deep semi-supervised learning approaches. For the one-sided label noise, since noisy data only appear in one category, Positive-Unlabeled (PU) learning [52] can be directly applied by discarding the noisy data. While for the two-sided label noise, as we additionally collect the class-prior information of the discrepant unlabeled datasets, we could use the Unlabeled-Unlabeled (UU) learning [11] to obtain the classifier. We also list the performance of the PNU [10] method, which regards all the noisy labels as true ones.

We report the average accuracy and standard deviation on MNIST and SVHN in Figure 2. Among these three algorithms, the LIISP algorithms converge to the highest accuracy. Notice that PNU is comparable or more accurate than the PU/UU methods, although they directly treat the noisy labels as the correct ones. This phenomenon indicates that it is necessary to consider the noisy labeled data in the LIISP scenario, although they are limited.

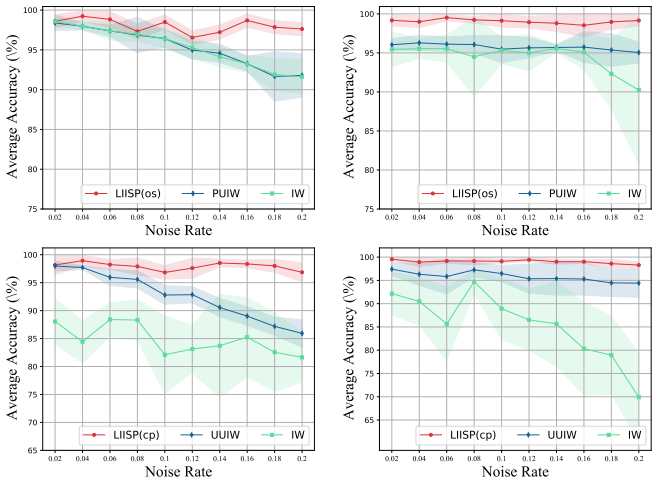
Comparison with Inaccurate Supervision Algorithms. We then compare the LIISP algorithms with noisy label learning approaches (for inaccurate supervision) to validate the usefulness of unlabeled data. Figure 3 reports the average accuracy and standard deviation of the LIISP(os/cp), PUIW/UUIW, and IW algorithms with increasing noise rate. In general, the proposed LIISP(os/cp) algorithms achieve the highest accuracy and drop more slowly than PUIW/UUIW and IW methods, as the proposed algorithms consider the structure of label noise and the statistical information (class-priors) of the discrepant unlabeled data. Additionally, LIISP(os/cp) are always more accurate and stable than the IW under fixed noise rate, indicating the robustness of the proposed LIISP(os/cp) algorithms and usefulness of unlabeled data on alleviating label noise, particularly when there are abundant unlabeled data available.

4. https://pytorch.org/hub/pytorch_vision_resnet/



(a) MNIST (b) SVHN

Fig. 2: Performance curve (in average accuracy and standard deviation) of the deep neural network implementation of the LIISP(os/cp) algorithms with other deep semi-supervised learning methods on the MNIST and SVHN datasets.

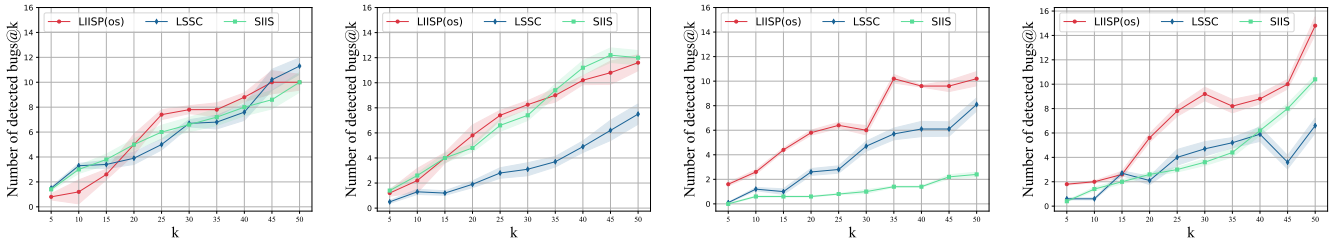


(a) isolet (b) letter7vs9

Fig. 3: Performance curve (in average accuracy and standard deviation) of the LIISP(os/cp) algorithms and the contenders with respect to an increasing noise rate.

6.3 Bug Detection Task

In this part, we examine the LIISP(os) algorithm in the real-world application, the bug detection task, where we aim to predict whether a source code is clean or buggy. Apart from those surely buggy codes reported by senior engineers (clean labeled data), those codes checked many times or newly fixed also potentially conceal bugs (noisy labeled data). Moreover, there exist a number of source codes that are never checked (unlabeled data). Therefore, this real-world application accords to a typical scenario of learning from incomplete and one-sided inaccurate supervision.



(a) OrientDB (File) (b) OrientDB (class) (c) Elasticsearch (File) (d) Elasticsearch (class)

Fig. 4: Performance curve w.r.t. an increasing number of predicted potential bugs. The performance is measured by the number of true bugs in top k predicted data, $|(\mathcal{S}@\mathrm{k}) \cap \mathcal{B}|$. The superior the algorithm is, the larger the quantity $|(\mathcal{S}@\mathrm{k}) \cap \mathcal{B}|$ will be.

TABLE 4: Descriptions of datasets for the bug detection task.

Dataset	Positive (Buggy)	Negative (Clean)	Total	# Dim
OrientDB (File)	270	1233	1503	7
OrientDB (Class)	208	1567	1847	102
Elasticsearch (File)	487	2548	3035	7
Elasticsearch (Class)	678	5230	5908	102

Source codes in a software project are usually modified by engineers. To define the positive and negative data in a bug detection task, we list the following three versions of each source code file according to whether it has related issues committed by engineers during the developing process:

(a) version before the issue is committed;
(b) version after the issue is reported but not yet fixed;
(c) version after the issue is fixed and closed.

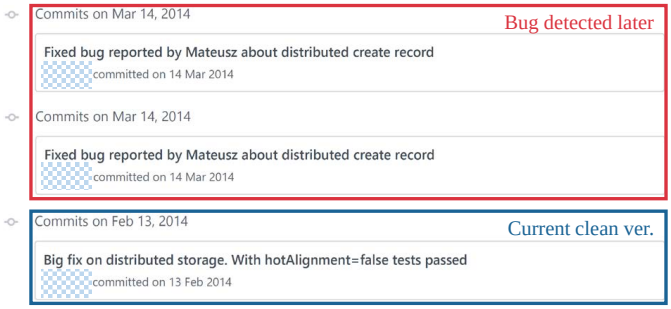
After identifying the relations between source codes and issues, we treat different versions of code files as individual instances. Specifically, we mark code files in version (a) and (b) as buggy instances (clean positive), while treat a code file in version (c) as a clean instance (noisy negative). Meanwhile, those source codes that are unrelated to any issue are marked as the unlabeled instances in our scenario.

Experiment Settings. We choose two public bug detection datasets of Java projects from GitHub [53]: (i) OrientDB[5] and (ii) Elasticsearch[6], where the former one is a database engine project and the latter one is a search engine project. We choose version 2013.12.10 for OrientDB and version 2014.02.03 for Elasticsearch. For each dataset, the feature of each instance is extracted either from the whole code file or from the class modules, and thus there are four datasets in total. Details of these four datasets are listed in Table 4.

We randomly take 50 buggy instances and 50 clean ones as the labeled data and set the rest code instances as unlabeled. For all experiments, we perform 10 tests on various splits of the whole dataset. We use the number of detected true positive bugs to characterize the performance, namely, the number of bugs identified by the algorithm are indeed buggy. More specifically, we denote the set of top k bugs detected for dataset $\mathcal{S}$ by the algorithm as $\mathcal{S}@\mathrm{k}$, and the collection of underlying ground-truth bugs in the test set as $\mathcal{B}$. Then we define the number of true bugs in the detected k instances as $|(\mathcal{S}@\mathrm{k}) \cap \mathcal{B}|$. Evidently, the better the performance of the algorithm is, the larger the quantity $|(\mathcal{S}@\mathrm{k}) \cap \mathcal{B}|$ will be.

Result Analyses. We then study performance of the LIISP(os) algorithm on these four bug detection tasks. We report the average number of detected bugs and their standard deviation in Figure 4. To better present the results, we only choose LSSC and SIIS as comparative methods, as they are the most competitive ones and outperform other baselines on benchmark datasets. Figure 4 demonstrates that the LIISP(os) algorithm has a very promising performance compared with other two methods, especially on the Elasticsearch dataset, see Figure 4(c) and 4(d). SIIS also shows a rather comparable result on the OrientDB dataset but behaves poorly on the Elasticsearch dataset. The reason may be that SIIS is not suitable for a relatively large dataset (like Elasticsearch), as it requires to perform the singular value decomposition on the Laplacian matrix, which is in the cubic dependence of the size of the training set.

Furthermore, we demonstrate the ability of the LIISP(os) algorithm to find out potentially buggy codes in Figure 5. In the following, we take the results on the OrientDB dataset as an example. As highlighted in the blue frame, this code file was fixed and labeled as clean in the current version (Oct 2, 2013). However, the code file is scored high by



(a) OrientDB Dataset

(b) Elasticsearch Dataset

Fig. 5: Potential bugs detected by the LIISP(os) algorithm on the OrientDB and Elasticsearch datasets. The username information is mosaiced in order to protect the privacy.

5. https://github.com/orientechnologies/orientdb
6. https://github.com/elasticsearch/elasticsearch

the LIISP(os) algorithm, which is suspected to be buggy with high probability. After checking their later commit records, which is highlighted in the orange frame, we find that this code file is indeed buggy and fixed after three months, although this concealed bug is not detected in the 2013 version. This strongly supports the effectiveness of our proposed algorithm.

Overall, these phenomena validate the effectiveness of the LIISP(os) algorithm, which not only achieves promising results in benchmark datasets but also succeeds in the real-world application for the bug detection task.

7 Conclusion

In this paper, we study the problem of Learning from Incomplete and Inaccurate SuPervision (LIISP). We observe that in many real-world applications, the label noise usually occurs in a one-side manner, which enables us to exploit the one-sided accurate label and sufficient unlabeled data to alleviate the noisy labeled data via the importance weighting technique. Furthermore, when the noisy labels exist in both positive and negative data, we additionally exploit the class-prior information for the discrepant unlabeled data to resist the label noise. Our proposed approaches are equipped with nice theoretical guarantees: by excess risk analysis, we theoretically justify the usefulness of unlabeled data in defensing instance-dependent label noise. We conduct extensive experiments on benchmark datasets as well as the bug detection task, demonstrating the superiority and robustness of our methods compared with contenders from other categories: semi-supervised learning, noisy label learning, and robust semi-supervised learning.

In the future, we will consider other weakly supervised learning tasks in open and dynamic environments [54], [55]. Moreover, in addition to incomplete and inaccurate supervision, we will further consider other weak supervision, such as the supervised information from knowledge reasoning [56].

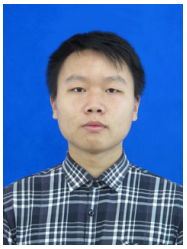
Acknowledgments

This research was supported by National Science Foundation of China (61921006) and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
[2] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 2017.
[3] K. P. Bennett and A. Demiriz, "Semi-supervised support vector machines," in *Advances in Neural Information Processing Systems 11* (*NIPS*), 1998, pp. 368–374.
[4] X. Zhu, Z. Ghahramani, and J. D. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *Proceedings of the 20th International Conference on Machine Learning* (*ICML*), 2003, pp. 912–919.
[5] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, "Learning with noisy labels," in *Advances in Neural Information Processing Systems 26* (*NIPS*), 2013, pp. 1196–1204.
[6] A. K. Menon, B. van Rooyen, and N. Natarajan, "Learning from binary labels with instance-dependent noise," *Machine Learning*, vol. 107, no. 8-10, pp. 1561–1595, 2018.
[7] Z.-Y. Zhang, P. Zhao, Y. Jiang, and Z.-H. Zhou, "Learning from incomplete and inaccurate supervision," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 1017–1025.
[8] M. C. du Plessis, G. Niu, and M. Sugiyama, "Analysis of learning from positive and unlabeled data," in *Advances in Neural Information Processing Systems 27* (*NIPS*), 2014, pp. 703–711.
[9] G. Niu, M. C. du Plessis, T. Sakai, Y. Ma, and M. Sugiyama, "Theoretical comparisons of positive-unlabeled learning against positive-negative learning," in *Advances in Neural Information Processing Systems 29* (*NIPS*), 2016, pp. 1199–1207.
[10] T. Sakai, M. C. Plessis, G. Niu, and M. Sugiyama, "Semi-supervised classification based on classification from positive and unlabeled data," in *Proceedings of the 34th International Conference on Machine Learning* (*ICML*), 2017, pp. 2998–3006.
[11] N. Lu, G. Niu, A. K. Menon, and M. Sugiyama, "On the minimal supervision for training any binary classifier from only unlabeled data," in *Proceedings of the 7th International Conference on Learning Representations* (*ICLR*), 2019.
[12] N. Lu, T. Zhang, G. Niu, and M. Sugiyama, "Mitigating overfitting in supervised classification from two unlabeled datasets: A consistent risk correction approach," in *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics* (*AISTATS*), 2020, pp. 1115–1125.
[13] D. Angluin and P. Laird, "Learning from noisy examples," *Machine Learning*, vol. 2, no. 4, pp. 343–370, 1988.
[14] J. A. Aslam and S. E. Decatur, "On the sample complexity of noise-tolerant learning." *Information Processing Letters*, vol. 57, no. 4, pp. 189–195, 1996.
[15] A. B. Novikoff, "On convergence proofs for perceptrons," Tech. Rep., 1963.
[16] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *Journal of Machine Learning Research*, vol. 7, no. Mar, pp. 551–585, 2006.
[17] L. Xu, K. Crammer, and D. Schuurmans, "Robust support vector machine training via convex outlier ablation," in *Proceedings of the 21st AAAI Conference on Artificial Intelligence* (*AAAI*), 2006, pp. 536–542.
[18] V. S. Denchev, N. Ding, S. Vishwanathan, and H. Neven, "Robust classification with adiabatic quantum optimization," in *Proceedings of the 29th International Conference on Machine Learning* (*ICML*), 2012, pp. 1003–1010.
[19] W. Gao, L. Wang, Z.-H. Zhou *et al.*, "Risk minimization in the presence of label noise," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence* (*AAAI*), 2016.
[20] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 3, pp. 447–461, 2016.
[21] R. Wang, T. Liu, and D. Tao, "Multiclass learning with partially corrupted labels," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2568–2580, 2017.
[22] A. Ghosh, N. Manwani, and P. Sastry, "Making risk minimization tolerant to label noise," *Neurocomputing*, vol. 160, pp. 93–107, 2015.
[23] P. Awasthi, M.-F. Balcan, N. Haghtalab, and R. Urner, "Efficient learning of linear separators under bounded noise," in *Proceedings of the 28th Annual Conference on Learning Theory* (*COLT*), 2015, pp. 167–190.
[24] M.-F. Balcan and A. Blum, "A pac-style model for learning from labeled and unlabeled data," in *Proceedings of the 18th Annual Conference on Learning Theory* (*COLT*), 2005, pp. 111–126.
[25] P. Niyogi, "Manifold regularization and semi-supervised learning: Some theoretical analyses," *Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1229–1250, 2013.
[26] M. Culp and G. Michailidis, "Graph-based semisupervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 1, pp. 174–179, 2007.
[27] W. Wang and Z.-H. Zhou, "A new analysis of co-training," in *Proceedings of the 27th International Conference on Machine Learning* (*ICML*), 2010, pp. 1135–1142.
[28] Z.-H. Zhou and M. Li, "Semi-supervised learning by disagreement," *Knowledge and Information Systems*, vol. 24, no. 3, pp. 415–439, 2010.
[29] Z.-H. Zhou, "Why over-parameterization of deep neural networks does not overfit?" *Science China Information Sciences*, vol. 64, no. 1, p. 116101, 2021.
[30] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Advances in Neural Information Processing Systems 28* (*NIPS*), 2015, pp. 3546–3554.
[31] D.-D. Chen, W. Wang, W. Gao, and Z.-H. Zhou, "Tri-net for semi-supervised deep learning," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (*IJCAI*), 2018, pp. 2014–2020.

[32] E. Sansone, F. G. De Natale, and Z.-H. Zhou, "Efficient training for positive unlabeled learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 11, pp. 2584–2598, 2019.
[33] M. Li and Z.-H. Zhou, "Setred: Self-training with editing," in *Proceedings of the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, 2005, pp. 611–621.
[34] Z.-H. Zhou and M. Li, "Tri-training: Exploiting unlabeled data using three classifiers," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 11, pp. 1529–1541, 2005.
[35] Y.-F. Li and Z.-H. Zhou, "Towards making unlabeled data never hurt," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp. 175–188, 2015.
[36] Y.-F. Li, L.-Z. Guo, and Z.-H. Zhou, "Towards safe weakly supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 334–346, 2021.
[37] Y.-F. Li and D.-M. Liang, "Safe semi-supervised learning: a brief introduction," *Frontiers of Computer Science*, vol. 13, no. 4, pp. 669–676, 2019.
[38] Z. Lu, X. Gao, L. Wang, J.-R. Wen, and S. Huang, "Noise-robust semi-supervised learning by large-scale sparse coding," in *Proceedings of the 29th AAAI Conference on Artificial Intelligence* (*AAAI*), 2015, pp. 2828–2834.
[39] Y. Yan, Z. Xu, I. W. Tsang, G. Long, and Y. Yang, "Robust semi-supervised learning through label aggregation," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence* (*AAAI*), 2016, pp. 2244–2250.
[40] C. Gong, H. Zhang, J. Yang, and D. Tao, "Learning with inadequate and incorrect supervision," in *Proceedings of the 17th International Conference on Data Mining* (*ICDM*), 2017, pp. 889–894.
[41] M. Sugiyama, T. Suzuki, and T. Kanamori, "Density ratio estimation: A comprehensive review," 2010.
[42] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.
[43] R. Collobert, F. Sinz, J. Weston, and L. Bottou, "Trading convexity for scalability," in *Proceedings of the 23rd International Conference on Machine Learning* (*ICML*), 2006, pp. 201–208.
[44] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, 2011.
[45] L.-Z. Guo and Y.-F. Li, "A general formulation for safely exploiting weakly supervised data." in *Proceedings of the 32rd AAAI Conference on Artificial Intelligence* (*AAAI*), 2018, pp. 3126–3133.
[46] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[47] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.
[48] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics* (*AISTATS*), 2011, pp. 315–323.
[49] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32rd International Conference on Machine Learning* (*ICML*), 2015, pp. 448–456.
[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (*CVPR*), 2016, pp. 770–778.
[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2015.
[52] R. Kiryo, G. Niu, M. C. du Plessis, and M. Sugiyama, "Positive-unlabeled learning with non-negative risk estimator," in *Advances in Neural Information Processing Systems 30* (*NIPS*), 2017, pp. 1675–1685.
[53] Z. Tóth, P. Gyimesi, and R. Ferenc, "A public bug database of github projects and its application in bug prediction," in *Proceedings of the 16th International Conference on Computational Science and Its Applications*, 2016, pp. 625–638.
[54] Z.-H. Zhou, "Learnware: on the future of machine learning," *Frontiers of Computer Science*, vol. 10, no. 4, pp. 589–590, 2016.
[55] M. Sugiyama and M. Kawanabe, *Machine Learning in Non-stationary Environments: Introduction to Covariate Shift Adaptation*. MIT press, 2012.
[56] Z.-H. Zhou, "Abductive learning: Towards bridging machine learning and logical reasoning," *Science China Information Sciences*, vol. 62, no. 7, p. 76101, 2019.



Zhen-Yu Zhang received his B.Sc. degree from University of Electronic Science and Technology of China, in 2017. Currently, he is working toward the PhD degree with the National Key Lab for Novel Software Technology in Nanjing University, supervised by Prof. Yuan Jiang and Prof. Zhi-Hua Zhou. His research interest is mainly on machine learning and data mining.

Peng Zhao received his B.Sc. degree from Tongji University, Shanghai, China, in 2016. Currently, he is working toward the PhD degree with the National Key Lab for Novel Software Technology in Nanjing University, supervised by Prof. Zhi-Hua Zhou. His research interest is mainly on machine learning and data mining. He is currently working on robust learning in non-stationary environments.

Yuan Jiang received the PhD degree in computer science from Nanjing University, China, in 2004. At the same year, she became a faculty member in the Department of Computer Science & Technology at Nanjing University, China and currently is a Professor. She was selected in the Program for New Century Excellent talents in University, Ministry of Education in 2009. Her research interests are mainly in artificial intelligence, machine learning, and data mining. She has published over 50 papers in leading international/national journals and conferences.

Zhi-Hua Zhou (S'00-M'01-SM'06-F'13) received the BSc, MSc and PhD degrees in computer science from Nanjing University, China, in 1996, 1998 and 2000, respectively, all with the highest honors. He joined the Department of Computer Science & Technology at Nanjing University as an Assistant Professor in 2001, and is currently Professor, Head of the Department of Computer Science and Technology, and Dean of the School of Artificial Intelligence; he is also the Founding Director of the LAMDA group. His research interests are mainly in artificial intelligence, machine learning and data mining. He has authored the books *Ensemble Methods: Foundations and Algorithms* (2012), *Evolutionary Learning: Advances in Theories and Algorithms* (2019), *Machine Learning* (2016, in Chinese), and published more than 150 papers in top-tier international journals or conference proceedings. He has received various awards/honors including the National Natural Science Award of China, the IEEE Computer Society Edward J. McCluskey Technical Achievement Award, the PAKDD Distinguished Contribution Award, the IEEE ICDM Outstanding Service Award, the Microsoft Professorship Award, etc. He also holds 24 patents. He is the Editor-in-Chief of the *Frontiers of Computer Science*, Associate Editor-in-Chief of the *Science China Information Sciences*, Action or Associate Editor of the *Machine Learning*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *ACM Transactions on Knowledge Discovery from Data*, etc. He served as Associate Editor-in-Chief for *Chinese Science Bulletin* (2008-2014), Associate Editor for *IEEE Transactions on Knowledge and Data Engineering* (2008-2012), *IEEE Transactions on Neural Networks and Learning Systems* (2014-2017), *ACM Transactions on Intelligent Systems and Technology* (2009-2017), *Neural Networks* (2014-2016), etc. He founded ACML (Asian Conference on Machine Learning), served as Advisory Committee member for IJCAI (2015-2016), Steering Committee member for ICDM, PAKDD and PRICAI, and Chair of various conferences such as General co-chair of ICDM 2016 and PAKDD 2014, Program co-chair of AAAI 2019 and SDM 2013, and Area chair of NeurIPS, ICML, AAAI, IJCAI, KDD, etc. He was the Chair of the IEEE CIS Data Mining Technical Committee (2015-2016), the Chair of the CCF-AI (2012-2019), and the Chair of the CAAI Machine Learning Technical Committee (2006-2015). He is a foreign member of the Academy of Europe, and a Fellow of the ACM, AAAI, AAAS, IEEE, IAPR, IET/IEE, CCF, and CAAI.