Learning from Incomplete and Inaccurate Supervision

Zhen-Yu Zhang, Peng Zhao, Yuan Jiang and Zhi-Hua Zhou
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
{zhangzy,zhaop,jiangy,zhouzh}@lamda.nju.edu.cn

ABSTRACT

In plenty of real-life tasks, strongly supervised information is hard to obtain, such that there is not sufficient high-quality supervision to make traditional learning approaches succeed. Therefore, weakly supervised learning has drawn considerable attention recently. In this paper, we consider 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 usually with *one-sided* noise. For instance, considering the bug detection task in the software system, the identified buggy codes are indeed with defects whereas the codes that have been checked many times or newly fixed may still have other flaws due to the complexity of the system. We propose a novel method which is able to effectively alleviate the negative influence of one-sided label noise with the help of a vast number of unlabeled data. Excess risk analysis is provided as theoretical justifications on the usefulness of incomplete and one-sided inaccurate supervision. We conduct experiments on synthetic, benchmark datasets, and real-life tasks to validate the effectiveness of the proposed approach.

CCS CONCEPTS

• **Computing methodologies** → **Semi-supervised learning settings**; **Machine learning**; *Supervised learning*;

KEYWORDS

weakly supervised learning; label noise; unlabeled data

ACM Reference Format:
Zhen-Yu Zhang, Peng Zhao, Yuan Jiang and Zhi-Hua Zhou. 2019. Learning from Incomplete and Inaccurate Supervision. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330902

1 INTRODUCTION

Machine learning has achieved great success in many real-world tasks, especially in supervised learning tasks. These successful techniques, such as deep learning [13], typically require a vast number of training data with accurate labels. However, it is often the case that such strong supervision is not easy to obtain due to the high cost of the labeling process. Therefore, it is desired to facilitate the learning system with the capability of learning from *weak supervision* [33].

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. Meanwhile, the given labels might be inaccurate. This setting is crucial since it occurs in many real-world applications. For instance, considering the annotation of medical images in the hospital, as the number of doctors is limited, there exist amounts of medical images without labels. Even for those labeled images, they can be wrongly annotated due to fatigue or negligence. Similar situations also occur in learning with biology data. Supervised information of each molecule is not always correct due to limitations of the equipment capability, and the number of labeled molecule is also limited as the biological experiment is typically costly and can last several days.

These two issues have been studied separately in the area of *Semi-Supervised Learning* (SSL) [4, 34] and *Noisy Label Learning* (NLL) [19, 20]. For incomplete supervision, SSL approaches leverage unlabeled data and limited labeled data to construct the predictor. However, when labeled data are inaccurate, these noisy labels can seriously deceive the learning system. For inaccurate supervision, NLL approaches manage to learn the predictor with respect to the underlying noise-free distribution in order to resist the noise. Nevertheless, they need sufficient labeled data and cannot utilize numerous unlabeled data. Therefore, it is very much appreciated for approaches which are able to handle the unlabeled and noisy data simultaneously.

There lack relevant studies for the problem of learning with incomplete and inaccurate supervision simultaneously, where there are a vast number of unlabeled data and a limited number of potentially noisy labeled data. This problem turns out quite challenging, and it is non-trivial to combine advantages of SSL and NLL approaches to address this problem. For traditional 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.

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330902

Given only the observed data, the problem of learning with incomplete and inaccurate supervision is probably almost impossible to solve, particularly when the limited labeled data are arbitrarily corrupted. Fortunately, in many real-world applications, we can have side information on the structure of the noise on the labeled data. Specifically, apart from a vast number of unlabeled data, the limited labeled data are with *one-sided instance-dependent* noise, namely, only 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 circumstance is quite common in real-life tasks. For instance, considering the bug detection task, whose purpose is to find out buggy codes in the software system. The codes reported with bug issues by the senior engineers are surely buggy (clean label), nevertheless, even though some codes might have been checked many times or fixed recently, they are not definitely reliable and potentially with bugs (noisy label) due to the complexity of the software system. Meanwhile, plenty of newly submitted codes are not labeled since it is hard to check them one-by-one entirely (unlabeled). Another example emerges the primary screening of cancer in the community hospital. One is believed to be healthy if s/he passed the screening examination. On the contrary, when suspected with cancer, s/he will need further examination as equipment in the community hospital may be not that accurate. Consequently, people who passed the careful screening are deemed as healthy (clean label) whereas those who are detected ill may not be a real patient (noisy label). Meanwhile, there remain a large number of residents to be checked (unlabeled).

In this work, we propose a novel semi-supervised learning method, leveraging the incomplete supervision to alleviate the negative effect of the one-sided inaccurate supervision, and thus step towards learning with 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 positive-unlabeled learning [21], we utilize 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 Learning from Incomplete and one-sided Inaccurate Supervision (LIoIS). Both theoretical justifications and empirical studies show the benefit of unlabeled data and noisy labeled data, leading to the optimal convergence rate and remarkable performance gain.

We summarize the main contributions of this paper as follows:

(1) We study the problem of learning from incomplete and structured inaccurate supervision simultaneously, which occurs in many real-world applications but is rarely studied.
(2) We propose a novel learning method, which is able to 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 studies on both benchmark datasets and real-world applications to show the superiority and robustness of our proposed method.

In the following, we first briefly review related work in Section 2. Then, preliminary is introduced in Section 3. Next, we propose our method in Section 4, following with experimental results on both synthetic and real-world benchmark datasets in Section 5. Finally, we conclude the paper in Section 6.

2 RELATED WORK

Starting from the work [1], many studies on inaccurate supervision have been proposed in the theoretical community, for instance, [2] studied the learnability of noise tolerant learning in finite VC-dimension. Then, various practical approaches are proposed to reduce the effect caused by inaccurate supervision [8, 10]. Following the line of noisy label learning, instance-independent noise has been well studied [17, 20]. They provide guarantees for risk minimization under random classification noise in the general setting of convex surrogates. However, in practice, instance-dependent noise [11] is closer to the practical 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 [19] shows that the Bayes optimal classifiers can be recovered from the noisy distribution under certain assumptions. However, noisy label learning mainly focuses on supervised learning, how to deal with limited labeled data and large amounts of unlabeled data has not yet been well studied.

To make use 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 [3, 22], 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 [34], S3VMs [4], and disagreement-based methods [32]. Some SSL approaches have been extended to deep learning [6, 24]. In traditional SSL, 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) [9, 26]. Different from using unlabeled data as the regularizer of hypothesis space, PU learning assumes the unlabeled data essentially 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 [25]. Another recent work by [14] analyzes the problem of biased negative data. In their formulation, negative data only represent a small portion of the whole negative distribution. Their work is closer to the dataset shift problem whose labeled data generating from one (training) distribution while unlabeled data from the other (test) distribution. 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.

Note that disagreement-based SSL approaches exploit pseudo-labels of unlabeled data [32], and to handle misleading pseudo-labels, some strategies such as data editing [15] or one-sided noisy label learning [31] have been incorporated. These can be seen as early studies considering both incomplete supervision and inaccurate supervision, though the inaccurate supervision was generated during SSL learning procedure, rather than from the initial training

data. Some recent studies about Safe-SSL [16] also have inherent mechanisms to handle the label noise, though these mechanisms are implicit. Later there are some other studies [12, 18, 30] which tried to improve robustness of SSL but were mostly heuristic and did not consider structural property.

3 PRELIMINARY

In this section, we first review some notations for learning from complete and accurate supervision, namely, the conventional supervised learning. Then, we demonstrate the preliminary knowledge for learning from incomplete but accurate supervision, namely, the semi-supervised learning.

3.1 Learning from Complete and Accurate Supervision

In this scenario, we observe the ground-truth accurate label for each instance. Let $\mathcal{D}$ be the underlying true distribution from which all data $(x, y) \in (\mathcal{X} \times \mathcal{Y})$ are independently and identically sampled, where $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} = \{-1, +1\}$. Suppose that we have n_P positive data $\{x_i, +1\}_{i=1,\ldots,n_P}$ and n_N negative data$\{x_j, -1\}_{j=1,\ldots,n_N}$. Our purpose is to learn a real-valued decision function $g : \mathbb{R}^d \to \mathbb{R}$ for the binary classification.

Let $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$ be a non-negative Lipschitz-continuous loss function, whose expected risk is,

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

where θ_P is the class-prior of positive label $\Pr[y = +1]$ and θ_N is $\Pr[y = -1]$ with $\theta_P + \theta_N = 1$. Besides, $\mathbb{E}_P$ and $\mathbb{E}_N$ denote the expectation over $\Pr[x|y = +1]$ and $\Pr[x|y = -1]$, respectively.

As only empirical data are available in practice, we use the empirical risk to approximate the expected one,

$$\widehat{R}(g) = \frac{\theta_P}{n_P} \sum_{i=1}^{n_P} \ell(g(x_i), +1) + \frac{\theta_N}{n_N} \sum_{j=1}^{n_N} \ell(g(x_j), -1). \tag{2}$$

Suppose we are given a family of decision functions $\mathcal{G}$, in which the each function $g : \mathcal{X} \to \mathbb{R}$. Among them, let g^* denote the optimal decision function, with $\widehat{g}$ as its empirical version,

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

3.2 Learning from Incomplete but Accurate Supervision

Then, we consider learning from incomplete but accurate supervision, where we can only observe part of the ground-truth accurate labels for one category, and the rest remain unlabeled. Suppose that we have n_P positive data $\{x_i, +1\}_{i=1,\ldots,n_P}$ and n_U unlabeled data$\{x_k\}_{k=1,\ldots,n_U}$. Our purpose is still learning a real-valued decision function $g : \mathcal{X} \to \mathbb{R}$ for the binary classification.

However, negative data are not available in this scenario and thus $\theta_N \mathbb{E}_N \left[\ell(g(x), -1)\right]$ in (1) cannot be directly estimated. Fortunately, du Plessis *et al.* [9] show that an unbiased estimator of $R(g)$ can be constructed by only using accurate positive and unlabeled data. Suppose the loss function ℓ satisfies the symmetric condition,

$$\ell(g(x), +1) + \ell(g(x), -1) = 1. \tag{3}$$

We first regard the unlabeled data as negative data,

$$\begin{aligned} \mathbb{E}_U \left[\ell(g(x), -1)\right] &= \theta_P \mathbb{E}_P \left[\ell(g(x), -1)\right] + \theta_N \mathbb{E}_N \left[\ell(g(x), -1)\right] \\ &= \theta_P - \theta_P \mathbb{E}_P \left[\ell(g(x), +1)\right] + \theta_N \mathbb{E}_N \left[\ell(g(x), -1)\right]. \end{aligned}$$

Let ℓ be a non-negative Lipschitz-continuous loss function and satisfy (3), then the expected risk can be rewritten as,

$$R(g) = 2\theta_P \mathbb{E}_P \left[\ell(g(x), +1)\right] + \mathbb{E}_U \left[\ell(g(x), -1)\right] - \theta_P. \tag{4}$$

By approximating $R(g)$ with empirical samples, we obtain,

$$\widehat{R}_{PU}(g) = 2\frac{\theta_P}{n_P} \sum_{i=1}^{n_P} \ell(g(x_i), +1) + \frac{1}{n_U} \sum_{k=1}^{n_U} \ell(g(x_k), -1). \tag{5}$$

For a given family of decision functions $\mathcal{G}$, let $\widehat{g}_{PU}$ denote the minimizer of empirical risk under such incomplete but accurate supervision, that is,

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

4 THE PROPOSED APPROACH

In this section, we present our approach to leverage the incomplete supervision to help learning with inaccurate supervision, in particular, the one-sided instance-dependent label noise.

We first present how to rewrite the risk for Learning with one-sided Inaccurate Supervision (LoIS), in which weights σ_+ and σ_- play crucial roles. Then, we propose to estimate these two weights by incorporating the unlabeled data. Finally, we provide our LIoIS method for Learning from Incomplete and one-sided Inaccurate Supervision.

4.1 Learning from one-side Inaccurate Supervision (LoIS)

In the setting of one-sided inaccurate supervision, without loss of generality, we suppose the 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 $\widetilde{P} = \{x_i, +1\}_{i=1,\ldots,n_{\widetilde{P}}}$[1] and $n_{\widetilde{N}}$ noisy negative data $\widetilde{N} = \{x_j, -1\}_{j=1,\ldots,n_{\widetilde{N}}}$. For each data item x, we denote its true label as y and the observed label as $\widehat{y}$. Evidently, we have $y = \widehat{y}$ for clean data, while it does not hold for noisy data. Meanwhile, let $\theta_{\widetilde{P}}$ be the class-prior of positive label $\Pr[\widehat{y} = +1]$ and $\theta_{\widetilde{N}}$ be $\Pr[\widehat{y} = -1]$ with $\theta_{\widetilde{P}} + \theta_{\widetilde{N}} = 1$. As there are positive and noisy negative data (the empirical $\widetilde{P}$ and $\widetilde{N}$ data) on hand, throughout the paper, we assume the class-prior is known in advance. Actually, in practice it can be estimated by the empirical positive and unlabeled data [23].

We assume that the noisy negative data have *instance-dependent* label noise [11, 19]. Specifically, for any (underlying, real) positive example x (whose true label $y = +1$), it is observed as a negative example ($\widehat{y} = -1$). We define this probability as the *hardness*, formally, $h : \mathcal{X} \to [0, 1]$, with,

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

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

We also suppose observed positive data are *always* accurate, which means for any $x \in \widetilde{P}$

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

Rewrite the True Risk. When only one-sided accurate supervision is available, 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 noise heavily. To cope with the inaccurate supervision, it is necessary to rewrite the true risk. In the following, we propose the oIS risk for the one-sided Inaccurate Supervision, and show that it is provably equal to the true risk.

Definition 1 (Risk for one-sided Inaccurate Supervision (oIS Risk))**.** For any function $g \in \mathcal{G}$, its oIS risk $R_{oIS}(g)$ is defined as,

$$R_{oIS}(g) = \theta_{\widetilde{P}}\mathbb{E}_{\widetilde{P}}[\sigma_+(x)\ell(g(x), +1)] + \theta_{\widetilde{N}}\mathbb{E}_{\widetilde{N}}\left[\sigma_-(x)\ell(g(x), -1)\right],$$

where $\sigma_+(x)$ and $\sigma_-(x)$ are defined as

$$\begin{aligned} \sigma_+(x) &= 1/\Pr[\widehat{y} = +1|x, y = +1], \\ \sigma_-(x) &= \Pr[y = -1|x, \widehat{y} = -1], \end{aligned} \tag{6}$$

which are the weights for positive and negative data.

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

$$R_{oIS}(g) = R(g).$$

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

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

The second equation holds due to the 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

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

A similar result can be obtained for the positive side by an analogous argument. This completes 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 take the noisy negative data into consideration, 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 accessible, we approximate the risk by the empirical oIS risk, defined as follows.

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

$$\widehat{R}_{oIS}(g) = \frac{\theta_{\widetilde{P}}}{n_{\widetilde{P}}}\sum_{i=1}^{n_{\widetilde{P}}}\sigma_+(x_i)\ell(g(x_i), +1) + \frac{\theta_{\widetilde{N}}}{n_{\widetilde{N}}}\sum_{j=1}^{n_{\widetilde{N}}}\sigma_-(x_j)\ell(g(x_j), -1), \tag{7}$$

where the weights $\sigma_+(x)$ and $\sigma_-(x)$ are defined in (6).

Therefore, provided with one-sided inaccurate supervision, we are able to learn the decision function according to the *empirical oIS risk minimization.* Let $\widehat{g}_{oIS}$ be the minimizer of the empirical oIS risk in the function family $\mathcal{G}$, that is,

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

We have the following excess risk bound, showing that the risk of $\widehat{g}_{oIS}$ converges to that of the optimal decision function in the function family $\mathcal{G}$.

Theorem 2 (Excess Risk of learning from oIS). *Assume that the loss function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$ is non-negative and L-Lipschitz continuous. Given that* hardness *$h(x) \in [0, h]$, then, for any $\delta > 0$, with probability at least $1 - \delta$, we have*

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

where $\mathfrak{R}_{n_{\widetilde{P}}}(\mathcal{G})$ is Rademacher complexity of the function family $\mathcal{G}$ for the sampling of size $n_{\widetilde{P}}$ from $p_+ = \Pr[x|\widehat{y} = +1]$ and $\mathfrak{R}_{n_{\widetilde{N}}}(\mathcal{G})$ follows a similar definition. Details will be provided in the longer version.

Remark 2. In Theorem 2, the uniform boundedness of hardness is necessary, otherwise, the excess risk can be unbounded. When the hardness h 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 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 setting, namely, there exists constant $h = \Pr[\widehat{y} = -1|x, y = +1] = \Pr[\widehat{y} = -1|y = +1]$, our method recovers the importance reweighting method proposed in [17]. Specifically, we can set the noise rate $\rho_{-1} = \Pr[\widehat{y} = +1|y = -1]$ defined in their method as 0 and $\rho_{+1} = \Pr[\widehat{y} = -1|y = +1] = h$.

4.2 Estimating σ_+ and σ_- by Incomplete Supervision

The key point of learning from oIS is to estimate the weights σ_+ and σ_- defined in (6). A similar weighting strategy for inaccurate supervision is also adopted in [17], but they can only deal with instance-independent label noise and are not able to utilize the unlabeled data, which is not desired particularly when the labeled data are scarce. However, it becomes a severe issue when we only have limited noisy labeled data on hand. Fortunately, with the help

of unlabeled data, such estimation can be fulfilled. In this paragraph, we consider estimating $\sigma_{+/-}$ from incomplete supervision.

As shown in [21], positive-unlabeled learning is provably better than supervised learning in terms of risk bounds when infinite unlabeled data are available. Therefore, with sufficient unlabeled data on hand, 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 (5), to produce pseudo labels for the noisy negative data and unlabeled data, which are then used to estimate the weights $\sigma_{+/-}$.

Estimate weights $\sigma_{+/-}$. We adopt ratio matching method proposed in [27] to estimate weights, and we only present the estimation of $\sigma_+(x)$, while the estimator for $\sigma_-(x)$ can be similarly obtained. Firstly, we rewrite $\sigma_+(x)$ as

$$\sigma_+(x) = \frac{\Pr[x, y=+1]}{\Pr[x, \widehat{y}=+1]} = \frac{\theta_P \Pr[x|y=+1]}{\theta_{\widetilde{P}} \Pr[x|\widehat{y}=+1]} = \frac{\theta_P}{\theta_{\widetilde{P}}} \sigma_{+_r}(x),$$

where θ_P and $\theta_{\widetilde{P}}$ are class-prior and $\sigma_{+_r}(x)$ denotes the remaining density ratio term. Based on the law of large numbers, θ_P and $\theta_{\widetilde{P}}$ can be estimated by the ratio of the number of samples as

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

in which $n_{y_{PU}=+1}$ and $n_{\widehat{y}=+1}$ denote the number of positive data estimated by $\widehat{g}_{PU}$ and the number of original observed positive data, respectively.

Then, we proceed to estimate $\sigma_{+_r}(x)$ by ratio matching method. Let $P_{PU} = \{x_i, \widehat{g}_{PU}(x_i) = +1\}_{i=1,\ldots,m}$ be the set of instances which are labeled as +1 by $\widehat{g}_{PU}$ of size m in training data, which approximates the instances sampled from $\Pr[x|y=1]$. As $\widetilde{P}$ is directly sampled from $\Pr[x|\widehat{y}=1]$, we can empirically approximate the discrepancy between estimated ratio and the true ratio by the Bregman divergence, defined as follows.

Definition 3 (Bregman divergence of ratio models [27])**.** For any differentiable and strictly convex function $f : \mathbb{R} \to \mathbb{R}$, let $\nabla f(x)$ denote the subgradient of $f(x)$. The Bregman divergence associated with f from the true density ratio σ_{+_r} to the estimated density ratio $\widehat{\sigma}_{+_r}$ is defined as,

$$B_f(\sigma_{+_r}||\widehat{\sigma}_{+_r}) = \int \Pr[x|\widehat{y}=+1] \nabla f(\widehat{\sigma}_{+_r}(x)) \widehat{\sigma}_{+_r}(x)\,\mathrm{d}x - \int \Pr[x|\widehat{y}=+1] f(\widehat{\sigma}_{+_r}(x))\,\mathrm{d}x - \int \Pr[x|y=+1] \nabla f(\widehat{\sigma}_{+_r}(x))\,\mathrm{d}x.$$

Thus, the empirical version $\widehat{B}_f(\sigma_{+_r}||\widehat{\sigma}_{+_r})$ is defined as,

$$\widehat{B}_f(\sigma_{+_r}||\widehat{\sigma}_{+_r}) = \frac{1}{n_{\widetilde{P}}} \sum_{i=1}^{n_{\widetilde{P}}} \nabla f(\widehat{\sigma}_{+_r}(x_i)) \widehat{\sigma}_{+_r}(x_i) - \frac{1}{n_{\widetilde{P}}} \sum_{i=1}^{n_{\widetilde{P}}} f(\widehat{\sigma}_{+_r}(x_i)) - \frac{1}{m} \sum_{j=1}^{m} \nabla f(\widehat{\sigma}_{+_r}(x_j)). \tag{8}$$

Therefore, provided with the set of sampled instances, namely $\widetilde{P}$ and P_{PU}, we are able to estimate the true density ratio by minimizing the empirical Bregman divergence. Let $\widehat{\sigma}^*_{+_r}$ be the minimizer of the empirical Bregman divergence in function family $\{\widehat{\sigma}_{+_r}\}$, that is,

$$\widehat{\sigma}^*_{+_r} = \operatorname*{arg\,min}_{\widehat{\sigma}_{+_r} \in \{\widehat{\sigma}_{+_r}\}} \widehat{B}_f(\sigma_{+_r}||\widehat{\sigma}_{+_r}).$$

For the density ratio estimated by the empirical Bregman divergence minimization, we have the following bound, showing that the estimated ratio $\widehat{\sigma}^*_{+_r}$ converges to the optimal density ratio in the function family $\{\widehat{\sigma}_{+_r}\}$.

Theorem 3. *Assuming $\sigma_{+_r}(x)$ is bounded, and function family $\{\widehat{\sigma}_{+_r}\}$ contains σ_{+_r}. For any $\delta > 0$, with probability at least $1-\delta$*

$$B_f(\sigma_{+_r}||\widehat{\sigma}^*_{+_r}) \le 2C\Re(\{\widehat{\sigma}_{+_r}\}) + b\sqrt{\frac{\log(4/\delta)}{2n_{\widetilde{P}}}},$$

where $\Re(\{\widehat{\sigma}_{+_r}\})$ is the Rademacher complexity of ratio model set, in the order of $O(1/\sqrt{n_{\widetilde{P}}})$. Meanwhile, C and b are constants. Detailed proofs will be presented in the longer version.

Theorem 3 guarantees that our estimated weight converges the optimal one in the hypothesis space, in the order of $O(1/\sqrt{n_{\widetilde{P}}})$, which enjoys a radical dependence of the number of instances. The analysis accords to the intuition as the estimator will be more accurate with more positive data available.

4.3 Learning from Incomplete and one-sided Inaccurate Supervision (LIoIS)

In order to learn from incomplete and one-sided inaccurate supervision, we propose to minimize the LIoIS risk, which is essentially a weighted combination of oIS risk and PU risk,

$$\begin{aligned} R_{LIoIS}(g) &= (1-\gamma) R_{oIS}(g) + \gamma R_{PU}(g) \\ &= (1-\gamma)\left\{\theta_{\widetilde{P}} \mathbb{E}_{\widetilde{P}}[\sigma_+(x)\ell(g(x),+1)] + \theta_{\widetilde{N}} \mathbb{E}_{\widetilde{N}}[\sigma_-(x)\ell(g(x),-1)]\right\} \\ &\quad + \gamma\left\{2\theta_P \mathbb{E}_P[\ell(g(x),+1)] + \mathbb{E}_U[\ell(g(x),-1)]\right\}, \end{aligned}$$

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 parts $\widetilde{P}_1$ and $\widetilde{P}_2$ of size $n_{\widetilde{P}_1}$ and $n_{\widetilde{P}_2}$, which are respectively adopted in the (empirical) oIS and PU risk,

$$\widehat{R}_{LIoIS}(g) = \gamma\left\{\frac{2\widehat{\theta}_P}{n_{\widetilde{P}_2}} \sum_{i=1}^{n_{\widetilde{P}_2}} \ell(g(x_i),+1) + \frac{1}{n_U} \sum_{k=1}^{n_U} \ell(g(x_k),-1)\right\} (1-\gamma)\left\{\frac{\widehat{\theta}_{\widetilde{P}}}{n_{\widetilde{P}_1}} \sum_{i=1}^{n_{\widetilde{P}_1}} \sigma_+(x_i)\ell(g(x_i),+1) + \frac{1-\widehat{\theta}_{\widetilde{P}}}{n_{\widetilde{N}}} \sum_{j=1}^{n_{\widetilde{N}}} \sigma_-(x_j)\ell(g(x_j),-1)\right\}. \tag{9}$$

Therefore, we can learn $\widehat{g}_{LIoIS}$, the minimizer of the weighted combination risk in the function family $\mathcal{G}$,

$$\widehat{g}_{LIoIS} = \operatorname*{arg\,min}_{g \in \mathcal{G}} \widehat{R}_{LIoIS}(g).$$

For the learned decision function, we have the following excess risk bound, showing that the risk of $\widehat{g}_{LIoIS}$ converges to that of optimal decision function in $\mathcal{G}$.

Theorem 4 (Excess Risk of LIoIS). *Assume that the loss function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$ is bounded, non-negative and L-Lipschitz continuous. Suppose the* hardness *$h(x) \le h$ holds uniformly for each instance,*

and there is a constant $C_{\mathcal{G}} > 0$ such that $\Re_n(\mathcal{G}) \leq C_{\mathcal{G}}/\sqrt{n}$ 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

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

Remark 4. Theorem 4 implies the usefulness of leveraging unlabeled data to alleviate label noise. As we can see, the risk bound is tighter with the number of unlabeled data increasing. Note that the risk bound is in optimal convergence rate (radical dependence) without any additional assumption [29].

5 EXPERIMENT

In this section, we examine the performance of the proposed LIoIS algorithm with applications on benchmark datasets and real-world tasks. Specifically, we evaluate our LIoIS algorithm in the following three aspects:

(i) **Comparisons on the synthetic dataset:** we provide intuitive illustrations on the advantage of our approach against traditional algorithms designed for only incomplete or only inaccurate supervision;
(ii) **Comparisons on benchmark datasets:** we compare LIoIS with robust SSL algorithms in benchmark datasets, to demonstrate the superiority of LIoIS in handling incomplete and structured inaccurate supervision.
(iii) **Bug Detection Task:** we validate the effectiveness of the proposed approach on the bug detection task, which aims at detecting defects in software systems.

For all the experiments, we randomly choose 50 positive and 50 negative examples as labeled data and set the rest as unlabeled. For benchmark datasets, we train an SVM and then flip 20% positive data into negative according to their confidence to simulate the instance-dependent label noise. We adopt Gaussian kernel and perform experiments 10 times on various splits of datasets, and present the average and standard deviation of the results. We conduct 10-fold cross validation to choose a proper trade off coefficient γ in LIoIS.

5.1 Comparisons on the synthetic dataset

We first numerically illustrate the performance of LIoIS under incomplete and structured inaccurate supervision. We generate a synthetic dataset from two class-conditional distributions, with each data item (x, y) generated from standard two-dimensional Normal distribution $\mathcal{N}_x$ according to

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

Apart from the noisy labels generation mentioned above, we provide 800 unlabeled data as incomplete supervision. The optimal boundary is shown in red solid line. Inaccurate Supervision algorithms (InaS) denote methods learning only with noisy labeled data and here we apply robust SVM [7]. Similarly, Incomplete Supervision algorithms (IncS) denote methods learning with unlabeled data, and we use PNU [25] for illustration. As shown in Figure 1, they both suffer from the structured inaccurate labels. The orange solid line denotes the boundary learned by LIoIS algorithm (9), which is closest to the optimal boundary. To conclude, our proposed LIoIS can empirically approximate the optimal boundary.

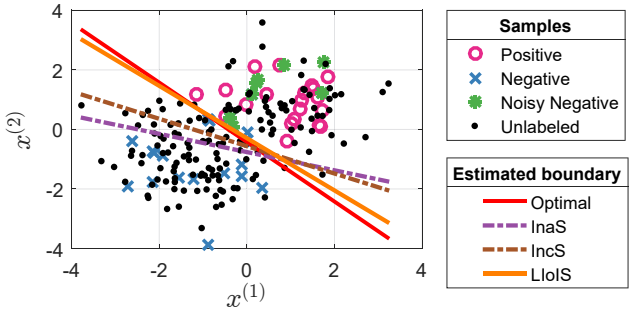


Figure 1: Decision boundaries of InaS, IncS and LIoIS.

Table 1: Brief statistics of benchmark datasets

Dataset	# instance	# dim	Dataset	# instance	# dim
house	232	16	australian	690	42
ionosphere	351	33	diabetes	768	8
clean1	476	166	german	1000	59
wdbc	569	14	letter7vs9	1528	16
isolet	600	51	a5a	6414	122
breastw	683	9	mnist7vs9	14251	600

5.2 Comparisons on benchmark datasets

We examine the performance of LIoIS algorithm on benchmark datasets, whose brief statistics are shown in Table 1. We compare LIoIS with seven methods, including two supervised learning baselines and five semi-supervised learning algorithms. The two supervised learning baselines are:

- LIBSVM [5] is an SVM baseline.
- IW [17] is a supervised method to alternate noise by utilizing importance reweighting.

There are other four robust semi-supervised learning algorithms, which take the noisy nature into consideration in SSL,

- S4VM [16] uses SVMs on labeled data to train a baseline and force final result not worse than supervised baseline.
- LSSC [18] is an SSL method based on sparse coding. It gives a L_1-norm formulation of Laplacian regularization based on the manifold structure of the data.
- ROSSEL [30] uses a set of weak annotators learned from noisy labeled data to generate pseudo labels for unlabeled data, and combines them to approximate the ground-truth labels by multiple label kernel learning.
- SIIS [12] is a graph based SSL algorithm. It emphasizes the leading eigenvectors of the Laplacian matrix associated with small eigenvalues to construct a robust graph and propagates labels on this robust graph.

We also perform a direct combination of the semi-supervised learning algorithm and noisy label learning algorithm,

- PUIW is a direct combination of PU learning and IW. We use PU learning to generate pseudo labels for unlabeled data, and then apply IW on labeled data.

From the results in Table 2, LIoIS ranks first in 9 out of 12 datasets in terms of the average accuracy. Overall, our proposed

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, and the best one is emphasized in bold. Besides, • indicates that LIoIS is significantly better than the compared method (paired t-tests at 95% significance level) and – indicates numerical limits or errors.

Dataset	LIBSVM	IW	S4VM	LSSC	ROSSEL	SIIS	PUIW	LIoIS (ours)
house	91.90 ± 1.72 •	**96.29 ± 1.27**	93.33 ± 1.37 •	93.49 ± 2.17 •	93.26 ± 1.88 •	88.84 ± 2.70 •	94.53 ± 2.80	96.03 ± 0.97
ionosphere	81.54 ± 3.19 •	83.28 ± 6.51	79.34 ± 7.07 •	79.26 ± 6.88 •	88.23 ± 4.64	72.11 ± 16.6 •	85.69 ± 2.56	**90.23 ± 7.43**
clean1	72.84 ± 3.81 •	64.52 ± 4.15 •	78.01 ± 3.62 •	61.03 ± 1.00 •	77.40 ± 2.37 •	60.89 ± 5.98 •	71.07 ± 3.92 •	**86.16 ± 3.42**
wdbc	89.65 ± 2.75 •	77.47 ± 19.3 •	80.76 ± 7.28 •	91.97 ± 2.01 •	90.87 ± 2.07 •	92.95 ± 1.32 •	77.64 ± 12.1 •	**95.52 ± 1.08**
isolet	86.50 ± 2.16 •	91.63 ± 2.44 •	87.34 ± 2.99 •	96.48 ± 1.25 •	80.13 ± 2.06 •	**98.82 ± 0.48**	91.74 ± 2.38 •	98.61 ± 1.21
breastw	93.65 ± 1.98	94.71 ± 1.23	91.54 ± 1.69	96.21 ± 1.29	**96.53 ± 0.83**	96.49 ± 0.68	95.53 ± 1.52	94.85 ± 4.32
australian	80.20 ± 3.24 •	80.65 ± 12.9	82.81 ± 3.19 •	81.87 ± 2.81 •	79.89 ± 6.92 •	72.96 ± 3.50 •	84.47 ± 3.90	**86.19 ± 1.05**
diabetes	74.91 ± 1.50 •	60.79 ± 14.1 •	69.69 ± 3.71 •	68.45 ± 2.35 •	75.69 ± 2.48	67.92 ± 1.37 •	75.93 ± 2.35	**76.26 ± 1.03**
german	64.52 ± 3.89 •	67.45 ± 4.81 •	65.81 ± 2.30 •	62.53 ± 1.86 •	73.03 ± 0.96	72.24 ± 1.19	68.65 ± 2.21 •	**74.37 ± 2.58**
letter7vs9	90.04 ± 3.88 •	95.21 ± 1.72 •	92.45 ± 4.65 •	94.23 ± 0.88 •	94.94 ± 1.43 •	78.47 ± 1.39 •	95.04 ± 1.34 •	**98.82 ± 0.95**
a5a	70.91 ± 2.42 •	73.82 ± 4.35 •	72.29 ± 2.71 •	68.45 ± 1.69 •	79.36 ± 1.66 •	76.36 ± 0.82 •	74.13 ± 2.47 •	**83.29 ± 0.47**
mnist7vs9	85.63 ± 2.29 •	90.18 ± 1.62 •	86.69 ± 1.43 •	88.76 ± 1.43 •	81.41 ± 1.18 •	– ± –	91.82 ± 1.53 •	**96.19 ± 0.33**
LIoIS w/ t/ l	11/ 1/ 0	10/ 1/ 1	12/ 0/ 0	10/ 1/ 1	9/ 2/ 1	9/ 1/ 2	8/ 3/ 1	rank first 9/ 12

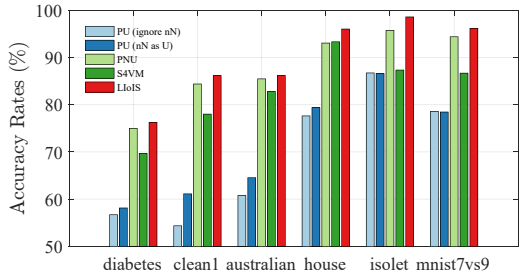
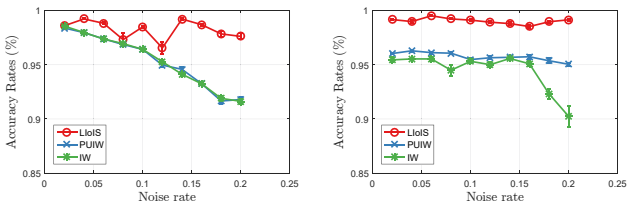


Figure 2: Performance comparisons on benchmark datasets.



(a) isolet (b) letter7vs9

Figure 3: Performance curve (in accuracy) of proposed approach with an increasing noise rate of negative data.

LIoIS method outperforms both supervised baselines and robust SSL algorithms. For supervised learning methods, notice that IW is not always better than LIBSVM baseline, and the large variance even makes the performance worse, which indicates the instability caused by limited noisy labeled data. While compared with the traditional noisy label learning methods (LIBSVM and IW), LIoIS achieves higher accuracy and better stability, showing the usefulness of unlabeled data. For robust semi-supervised learning algorithms, LIoIS achieves a very promising performance over the other four methods, as the prior knowledge of noise structure is provided. Particularly, LIoIS outperforms S4VM and ROSSEL, which heavily depend on the performance of weak learner(s) generated from noisy labeled data, which also shows the usefulness of unlabeled data. For the naive combination of PU and IW, LIoIS attains higher accuracy than PUIW over almost all datasets, which means a direct combination of IncS and InaS is not proper for incomplete and inaccurate supervision.

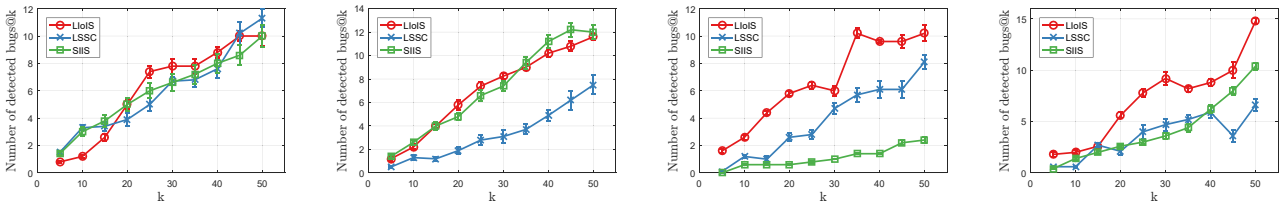
Comparison with Incomplete Supervision algorithms. We compare LIoIS with incomplete supervision algorithms. As noisy labeled data are structured, which only appear in one category, positive-unlabeled learning can be directly applied by discarding the noisy data. We denote them as *PU (w/o nN)* and *PU (nN as U)*, which drops the noisy labeled data or considers them as unlabeled, respectively. We also list the performance of *PNU* [25] and *S4VM* [16], in which both of them regard the noisy labels as true ones.

We report the mean accuracy on six datasets in Figure 2. Among these five algorithms, LIoIS achieves the highest accuracy. Notice that PNU and S4VM are comparable or more accurate than two kinds of PU algorithms, although they directly treat the noisy labels as the correct ones. This indicates that it is necessary to consider the noisy negative data when only limited positive data are provided.

Comparison with Inaccurate Supervision algorithms. We compare LIoIS with inaccurate supervision algorithms which are directly applied on noisy labeled data. Figure 3 reports the mean accuracy and standard deviation of LIoIS, PUIW, and IW with increasing noise rate. In general, LIoIS achieves the highest accuracy and drops more slowly than PUIW and IW, as the proposed LIoIS algorithm considers the instance-dependent label noise. Additionally, LIoIS is always more accurate and stable than IW under fixed noise rate, indicating the robustness of our proposed approach and usefulness of unlabeled data on alleviating label noise particularly when the number labeled data is limited.

5.3 Bug Detection Task

We examine LIoIS on the bug detection task, where the main purpose is to predict whether source code is clean or potentially buggy. Apart from those surely buggy codes reported by the senior engineers, the codes checked many times or newly fixed can also



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

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

(a) `OrientDB` (b) `Elasticsearch`

Figure 5: Potential bugs detected by LIoIS in OrientDB and Elasticsearch datasets.

have potential bugs. Moreover, a number of source codes are unlabeled. This accords to the setting of learning from incomplete and one-sided inaccurate supervision.

Since the codes in the project are usually modified by the engineers, to identify the positive and negative data in the bug detection task, we list the following three versions of each code file according to whether it has relative issues committed:

(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 detecting relations between codes and issues, we treat different versions of code files as instances. Specifically, we mark a code file in version (a) and (b) as two buggy instances (positive), while treat code file in version (c) as a clean instance (negative). Meanwhile, other source codes in the original version which are unrelated to any issue are marked as the clean instances (negative).

Experiment Settings. We choose two public bug detection datasets of Java projects from GitHub [28]: (i) OrientDB[2] and (ii) Elasticsearch[3], 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, and thus there are four datasets in total. Details of these two datasets are listed in Table 3.

We directly perform LIoIS algorithm on these two bug detection datasets. To simulate the incomplete supervision, we randomly take 50 buggy instances and 50 clean ones, and set the rest code instances as unlabeled. Notably, we choose the newly fixed version of code or code checked many times as clean training instances. For all experiments, we perform them 10 times on various splits of the labeled, unlabeled, and test sets. To measure the performance of algorithms, we use the number of detected true positive bugs to characterize the effectiveness, namely, bugs identified by the algorithm are indeed buggy. More specifically, we denote the set of top k bugs detected for dataset S by algorithms as $S@k$, and the set of underlying ground-truth bugs in the test set as $\mathcal{B}$. Then we define the number of detected true positive top k bugs as $|(S@k) \cap \mathcal{B}|$, evidently, the better the performance of the algorithm is, the larger the quantity $|(S@k) \cap \mathcal{B}|$ will be.

Table 3: Descriptions of datasets for the bug detection task.

Name	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

Result Analyses. We first report the average and standard deviation of the number of detected bugs on four bug datasets in Figure 4. To better present the results, we only choose LSSC and SIIS as comparative methods, as they are the most competitive and outperform other baselines in benchmark datasets. Figure 4 shows that our approach LIoIS has a promising performance compared to the other two comparative methods, especially on the Elasticsearch dataset, see Figure 4(c) and 4(d). Meanwhile, SIIS shows a comparable result

[2]https://github.com/orientechnologies/orientdb
[3]https://github.com/elasticsearch/elasticsearch

in OrientDB dataset but behaves poorly in Elasticsearch. The reason is that SIIS is not suitable for the relatively large datasets (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. Therefore, these phenomena validate the effectiveness of our proposed LIoIS, which not only achieves promising results in benchmark datasets but also succeeds in the real-world application on the bug detection task.

Furthermore, Figure 5 reports two potentially buggy codes in current version detected by LIoIS in OrientDB and Elasticsearch datasets. Take results in OrientDB as an example. As highlighted in the blue frame, this code file was fixed and labeled as clean in the version (Oct 2, 2013). However, the code file is scored high by our approach LIoIS, which is suspected to be buggy with high probability. After checking their later commit records, highlighted in the orange frame, we find that the code file is indeed buggy and fixed after three months, although the bugs are not detected in the 2013 version. This strongly supports the effectiveness of our proposed approach.

6 CONCLUSION

In this paper, we study the problem of learning from incomplete and inaccurate supervision, which accommodates many real-world applications. We observe that the label noise usually occurs in a one-side manner, and thus are able to utilize the one-sided accurate label and sufficient unlabeled data to alleviate the noisy labeled data via the importance weighting technique. The proposed approach is with nice theoretical guarantees, justifying 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, showing the superiority and robustness of our proposed method compared with contenders from other categories: semi-supervised learning, noisy label learning, and robust semi-supervised learning.

In the future, it is desired to design the approach that can deal with the circumstance that random instance-dependent label noise occurs in both positive and negative data.

7 ACKNOWLEDGMENTS

This research was supported by NSFC (61673201), the National Key R&D Program of China (2018YFB1004300), and Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.
[2] Javed A Aslam and Scott E Decatur. On the sample complexity of noise-tolerant learning. *Information Processing Letters*, 57(4):189–195, 1996.
[3] Maria-Florina Balcan and Avrim Blum. A pac-style model for learning from labeled and unlabeled data. In *Proceedings of the 18th Annual Conference on Learning Theory (COLT)*, pages 111–126, 2005.
[4] Kristin P Bennett and Ayhan Demiriz. Semi-supervised support vector machines. In *Advances in Neural Information Processing Systems 11 (NIPS)*, pages 368–374, 1998.
[5] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.
[6] Dong-Dong Chen, Wei Wang, Wei Gao, and Zhi-Hua Zhou. Tri-net for semi-supervised deep learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2014–2020, 2018.
[7] Ronan Collobert, Fabian Sinz, Jason Weston, and Léon Bottou. Trading convexity for scalability. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 201–208, 2006.
[8] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7(Mar):551–585, 2006.
[9] Marthinus C du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pages 703–711, 2014.
[10] Wei Gao, Bin-Bin Yang, and Zhi-Hua Zhou. On the robustness of nearest neighbor with noisy data. *arXiv:1607.07526*, 2016.
[11] Aritra Ghosh, Naresh Manwani, and PS Sastry. Making risk minimization tolerant to label noise. *Neurocomputing*, 160:93–107, 2015.
[12] Chen Gong, Hengmin Zhang, Jian Yang, and Dacheng Tao. Learning with inadequate and incorrect supervision. In *Proceedings of the 17th International Conference on Data Mining (ICDM)*, pages 889–894, 2017.
[13] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.
[14] Yu-Guan Hsieh, Gang Niu, and Masashi Sugiyama. Classification from positive, unlabeled and biased negative data. *arXiv:1810.00846*, 2018.
[15] Ming Li and Zhi-Hua Zhou. Setred: Self-training with editing. In *Proceedings of 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pages 611–621, 2005.
[16] Yu-Feng Li and Zhi-Hua Zhou. Towards making unlabeled data never hurt. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):175–188, 2015.
[17] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, 2016.
[18] Zhiwu Lu, Xin Gao, Liwei Wang, Ji-Rong Wen, and Songfang Huang. Noise-robust semi-supervised learning by large-scale sparse coding. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI)*, pages 2828–2834, 2015.
[19] Aditya Krishna Menon, Brendan Van Rooyen, and Nagarajan Natarajan. Learning from binary labels with instance-dependent corruption. *arXiv:1605.00751*, 2016.
[20] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 1196–1204, 2013.
[21] Gang Niu, Marthinus Christoffel du Plessis, Tomoya Sakai, Yao Ma, and Masashi Sugiyama. Theoretical comparisons of positive-unlabeled learning against positive-negative learning. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pages 1199–1207, 2016.
[22] Partha Niyogi. Manifold regularization and semi-supervised learning: Some theoretical analyses. *Journal of Machine Learning Research*, 14(1):1229–1250, 2013.
[23] Harish Ramaswamy, Clayton Scott, and Ambuj Tewari. Mixture proportion estimation via kernel embeddings of distributions. In *Proceedings of the 33nd International Conference on Machine Learning (ICML)*, pages 2052–2060, 2016.
[24] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 3546–3554, 2015.
[25] Tomoya Sakai, Marthinus Christoffel Plessis, Gang Niu, and Masashi Sugiyama. Semi-supervised classification based on classification from positive and unlabeled data. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 2998–3006, 2017.
[26] Emanuele Sansone, Francesco GB De Natale, and Zhi-Hua Zhou. Efficient training for positive unlabeled learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.
[27] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density ratio estimation: A comprehensive review. 2010.
[28] Zoltán Tóth, Péter Gyimesi, and Rudolf 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*, pages 625–638, 2016.
[29] Vladimir Naumovich Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999.
[30] Yan Yan, Zhongwen Xu, Ivor W Tsang, Guodong Long, and Yi Yang. Robust semi-supervised learning through label aggregation. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI)*, pages 2244–2250, 2016.
[31] Zhi-Hua Zhou and Ming Li. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge & Data Engineering*, (11):1529–1541, 2005.
[32] Zhi-Hua Zhou and Ming Li. Semi-supervised learning by disagreement. *Knowledge and Information Systems*, 24(3):415–439, 2010.
[33] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2017.
[34] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pages 912–919, 2003.