Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)
Early Career Track

Safe Weakly Supervised Learning

Yu-Feng Li
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
liyf@nju.edu.cn

Abstract

Weakly supervised learning (WSL) refers to learning from a large amount of weak supervision data. This includes i) *incomplete* supervision (e.g., semi-supervised learning); ii) *inexact* supervision (e.g., multi-instance learning) and iii) *inaccurate* supervision (e.g., label noise learning). Unlike supervised learning which typically achieves performance improvement with more labeled data, WSL may sometimes even degenerate performance with more weak supervision data. It is thus desired to study *safe* WSL, which could robustly improve performance with weak supervision data. In this article, we share our understanding of the problem from in-distribution data to out-of-distribution data, and discuss possible ways to alleviate it, from the aspects of worst-case analysis, ensemble-learning, and bi-level optimization. We also share some open problems, to inspire future researches.

1 Introduction

Machine learning has achieved great success in numerous tasks, particularly in supervised learning such as classification and regression. But most successful techniques, such as deep learning [LeCun *et al.*, 2015], require ground-truth labels to be given for a big training data set. It is noteworthy that in many tasks, however, it can be difficult to attain strong supervision due to the fact that the hand-labeled data sets are time-consuming and expensive to collect. Thus, it is desirable for machine learning techniques to be able to work well with weakly supervised data [Zhou, 2017].

Compared to the data in traditional supervised learning, weakly supervised data does not have a large amount of precise label information. Specifically, three types of weakly supervised data commonly exist [Zhou, 2017].

- *Incomplete* supervised data, i.e., only a small subset of training data is given with labels whereas the other data remain unlabeled. For example, in image categorization, it is easy to get a huge number of images fro the Internet, whereas only a small subset of images can be annotated due to the annotation cost.Representative techniques for this situation are *semi-supervised learning* [Chapelle *et al.*, 2006], which aims to learn a prediction model by leveraging a number of unlabeled data.
- *Inexact* supervised data, i.e., only coarse-grained labels are given. Reconsider the image categorization task, it is desirable to have every object in the images annotated, however, usually we only have image-level labels rather than object-level labels. One representative technique for this scenario is *multi-instance learning* [Carbonneau *et al.*, 2018], which aims to improve the performance by considering the coarse-grained label information.
- *Inaccurate* supervised data, i.e., the given labels have not always been ground-truth. Such situation occurs in various tasks when the annotator is careless or weary, or the annotator is not an expert. For this type of label information, *label noise learning* techniques are one main paradigm to learn a promising prediction from noisy label [Frénay and Verleysen, 2014].

In traditional machine learning, it is often expected that machine learning techniques, such as supervised learning, with the usage of more data will be able to improve learning performance. Such observation, however, no longer holds for weakly supervised learning. There are many studies [Li and Zhou, 2015; Li *et al.*, 2017; Guo and Li, 2018; Oliver *et al.*, 2018] reporting that the usage of weakly supervised data may sometimes lead to performance degradation, that is, the learning performance is even worse than that of baseline methods without using weakly supervised data. More specifically, semi-supervised learning using unlabeled data may be worse than vanilla supervised learning with only limited labeled data [Li and Zhou, 2015; Li *et al.*, 2016]. Multi-instance learning may be outperformed by the naive learning methods which simply assign the coarse-grained label to a bag of instances [Carbonneau *et al.*, 2018]. Label noise learning may be worse than that of learning from a small amount of high-quality labeled data [Frénay and Verleysen, 2014]. Such phenomena undoubtedly go against of the expectation of WSL and limits its effectiveness in a large number of practical tasks.

Building a safe WSL, that is to say, WSL using extra weakly supervised data will not be inferior to a simple supervised learning model, it the Holy Grail of WSL [Chapelle *et al.*, 2006; Li and Zhou, 2015; Zhou, 2017]. Since the problem was pointed out in [Cozman *et al.*, 2003], there are many

attempts trying to solve this important and challenging problem. In this article, we will review the recent developments on safe WSL, and share our contributions on two aspects of safe WSL:

- For WSL with in-distribution data, we proposed a general ensemble scheme that maximize the performance gain in the worse-case to improve the safeness of WSL.
- For WSL with out-of-distribution (OOD) data, we give a particular focus on SSL with unseen class unlabeled data and propose a bi-level optimization based framework to alleviate the potential performance hurt caused by OOD unlabeled examples.

We will also discuss some open challenges in real-world applications that may have been less noticed and desire more attentions.

2 Safe WSL with In-Distribution Data

WSL with in-distribution data, i.e., all supervised data and weakly supervised data are drawn from a same distribution, is the most natural situation. [Cozman *et al.*, 2003] pointed out that WSL could suffer performance degradation problem with in-distribution data. There are multiple reasons, for example, the adopted assumption of WSL algorithm is not suitable for the data distribution [Chapelle *et al.*, 2006]; there are many candidate large-margin decision boundaries existing in semi-supervised support vector machine (SVM) and prior knowledge is insufficient to help choose the best one [Li and Zhou, 2015], and so on.

Some attempts have been devoted to this problem [Li and Zhou, 2015; Loog, 2015; Li *et al.*, 2017; Krijthe and Loog, 2017]. For example, [Li and Zhou, 2015] builds safe semi-supervised SVMs through optimizing the worst-case performance gain given a set of candidate low-density separators. [Loog, 2015] proposes to maximize the likelihood gain over a supervised model in the worst-case for generative models. [Balsubramani and Freund, 2015] proposes to learn a robust prediction given that the ground-truth label assignment is restricted to a specific candidate set. [Wei *et al.*, 2018] study safe multi-label learning of weakly labeled data. They optimize multi-label evaluation metrics (F1 score and Top-k precision) given that the ground-truth label assignment is realized by a convex combination of base multi-label learners. More introductions can be found in our recent summary [Li and Liang, 2019].

To address this problem, we propose a general ensemble learning scheme, SAFEW (SAFE Weakly supervised learning) [Li *et al.*, 2021], which learning prediction by integrating multiple weakly supervised learners. Specifically, we propose a maximin framework, which maximize the performance gain in the worse case. Suppose we have obtained b predictions $\{\mathbf{f}_1, \cdots, \mathbf{f}_b\}$ generated by base weakly supervised learners, let $\mathbf{f}_0$ denote the prediction of baseline approaches, i.e., directly supervised learning with only limited labeled data. Our ultimate goal is here to derive a safe prediction $\mathbf{f} = g(\{\mathbf{f}_1, \cdots, \mathbf{f}_b\}, \mathbf{f}_0)$, which often outperforms the baseline $\mathbf{f}_0$, meanwhile it would not be worse than $\mathbf{f}_0$. In other words, we wold like to maximize the performance gain between our prediction and the baseline prediction.

By assuming that the ground-truth label assignment $\mathbf{f}^*$ can be realized as a convex combination of base learners, specifically, $\mathbf{f}^* = \sum_{i=1}^{b} \alpha_i \mathbf{f}_i$ where $\alpha = [\alpha_1; \alpha_2; \cdots; \alpha_b] \geq 0$ be the weight of base learners and $\sum_{i=1}^{b} \alpha_i = 1$, then we have the following objective function:

$$\max_{\mathbf{f}} \ell(\mathbf{f}_0, \sum_{i=1}^{b} \alpha_i \mathbf{f}_i) - \ell(\mathbf{f}, \sum_{i=1}^{b} \alpha_i \mathbf{f}_i) \tag{1}$$

This is in line with our goal that is to find a prediction f that maximizes the performance gain against the baseline $\mathbf{f}_0$.

In practice, however, on may be hard to know about the precise weight of base learners. We further assume that α is from a convex set $\mathcal{M}$ to make the proposal more practical, where $\mathcal{M}$ captures the priors knowledge about the importance of base learners. Without any further information to locate the weight of base learners, to guarantee the safeness, we aim to optimize the worse-case performance gain, since, intuitively, the algorithm would be robust as long as the good performance is guaranteed in the worst case. Then we obtain a general formulation for weakly supervised data as,

$$\max_{\mathbf{f}} \min_{\alpha \in \mathcal{M}} \ell(\mathbf{f}_0, \sum_{i=1}^{b} \alpha_i \mathbf{f}_i) - \ell(\mathbf{f}, \sum_{i=1}^{b} \alpha_i \mathbf{f}_i) \tag{2}$$

We have the following theorem to guarantees the safeness of our proposal for commonly used convex loss functions in both classification and regression tasks, e.g., hinge loss, cross-entropy loss, mean-square loss, etc.

Theorem 1. *Suppose the ground-truth $\mathbf{f}^*$ can be constructed by base learners, i.e., $\mathbf{f}^* \in \{\mathbf{f} | \sum_{i=1}^{b} \alpha_i \mathbf{f}_i, \alpha \in \mathcal{M}\}$. Let $\hat{\mathbf{f}}$ and $\hat{\alpha}$ be the optimal solution to Eq.(2). We have $\ell(\hat{\mathbf{f}}, \mathbf{f}^*) \leq \ell(\mathbf{f}_0, \mathbf{f}^*)$ and $\hat{\mathbf{f}}$ has already achieved the maximal performance gain against $\mathbf{f}_0$.*

Theorem 1 show that Eq.(2) is a reasonable formulation for our purpose, that is, the derived optimal solution $\hat{\mathbf{f}}$ from Eq.(2) often outperforms $\mathbf{f}_0$ and it would not get any worse than $\mathbf{f}_0$.

The objective formulation can be globally and efficiently addressed via a simple convex quadratic program or linear program. For example, with mean square loss, the objective can be equivalently written as

$$\min_{\boldsymbol{\alpha} \in \mathcal{M}} \quad \boldsymbol{\alpha}^\top \mathbf{F} \boldsymbol{\alpha} - \mathbf{v}^\top \boldsymbol{\alpha} \tag{3}$$

where $\mathbf{F} \in \mathbb{R}^{b \times b}$ is a linear kernel matrix of $\mathbf{f}_i$'s, i.e., $F_{ij} = \mathbf{f}_i^\top \mathbf{f}_j, \forall 1 \leq i, j \leq b$ and $\mathbf{v} = [2\mathbf{f}_1^\top \mathbf{f}_0; \ldots; 2\mathbf{f}_b^\top \mathbf{f}_0]$. Since $\mathbf{F}$ is positive semi-definite, Eq.(3) is convex and can be efficiently solved. After solving the optimal solution $\boldsymbol{\alpha}^*$, the optimal $\bar{\mathbf{f}} = \sum_{i=1}^{b} \alpha_i^* \mathbf{f}_i$ can be obtained.

Moreover, the optimization can be written as a geometric projection problem. Specifically, let $\Omega = \{\mathbf{f} | \sum_{i=1}^{b} \alpha_i \mathbf{f}_i, \boldsymbol{\alpha} \in \mathcal{M}\}$, $\bar{\mathbf{f}}$ can be rewritten as,

$$\bar{\mathbf{f}} = \arg\min_{\mathbf{f} \in \Omega} \|\mathbf{f} - \mathbf{f}_0\|^2, \tag{4}$$

which learns a projection of $\mathbf{f}_0$ onto the convex set Ω. Figure 1 illustrates the intuition of our proposed method via the viewpoint of geometric projection.

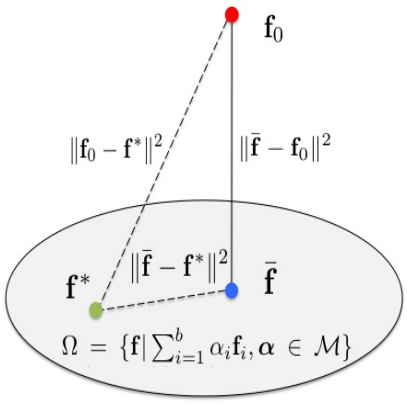

Figure 1: Illustration of the intuition of our proposal via the projection viewpoint. Intuitively, the proposal learns a projection of $\mathbf{f}_0$ onto a convex feasible set Ω.

It is noteworthy that, compared with previous studies in [Li and Zhou, 2015; Balsubramani and Freund, 2015; Li *et al.*, 2017], the SAFEW framework brings multiple advantages to safe WSL. i) It can be shown that the proposal is probably safe as long as the ground-truth label assignment can be expressed as a convex combination of base learners. In contrast to [Li and Zhou, 2015] which requires that the ground-truth is one of the base learners, the condition in Theorem 1 is looser and more practical. ii) Prior knowledge related to the weight of base learners can be easily embedded in this framework. iii) The framework is readily applicable for many loss functions in both classification and regression, which is more general in contrast to [Li *et al.*, 2017] that focuses on regression. iv) The proposed formulation can be globally and efficiently addressed and have intuitive geometry interpretation.

3 Safe WSL with OOD Data

Previous WSL studies are based on a basic assumption that labeled data and weakly supervised data come from the same distribution. Such an assumption is difficult to hold in many practical applications, among which one common case is that OOD unlabeled data that contains classes that are not seen in the labeled set occurs for SSL. For example, in medical diagnosis, unlabeled medical images often contain different foci from the diseases to be diagnosed. Faced with the OOD weakly supervised data, WSL no longer works well and may even be accompanied by severe performance degradation [Oliver *et al.*, 2018].

Efforts on safe WSL with OOD data remains to be limited. We have made particularly efforts to safe SSL problem and proposes a simple and effective safe deep SSL framework DS3L (Deep Safe Semi-Supervised Learning) [Guo *et al.*, 2020].

Specifically, in SSL scenarios, we are given a set of training data from an unknown distribution, which includes n labeled instances $\mathcal{D}_l = \{(\mathbf{x}_1, \mathbf{y}_1), \cdots, (\mathbf{x}_n, \mathbf{y}_n)\}$ and m unlabeled instances $\mathcal{D}_u = \{\mathbf{x}_{n+1}, \cdots, \mathbf{x}_{n+m}\}$. $\mathbf{x} \in \mathcal{X} \in \mathbb{R}^D$, $\mathbf{y} \in \mathcal{Y} = \{1, \cdots, C\}$ where D is the number of input dimension and C is the number of output class in labeled data.

The goal of SSL is to learn a model $h(\mathbf{x}; \theta) : \{\mathcal{X}; \Theta\} \to \mathcal{Y}$ parameterized by $\theta \in \Theta$ from training data to minimize the generalization risk $R(h) = \mathbb{E}_{(X,Y)}[\ell(h(X; \theta), Y)]$, where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ refers to certain loss function, e.g., mean squared error or cross entropy loss.

The way SSL utilizes unlabeled data structures is usually through the introduction of regularization. The objective of SSL is typically formulated as following:

$$\min_{\theta \in \Theta} \sum_{i=1}^{n} \ell(h(\mathbf{x}_i; \theta), \mathbf{y}_i) + \Omega(\mathbf{x}; \theta) \ \text{ s.t. } \mathbf{x} \in \mathcal{D}_l \cup \mathcal{D}_u. \quad (5)$$

where $\Omega(\mathbf{x}; \theta)$ refers to the regularization term, e.g., entropy-minimization regularization [Grandvalet and Bengio, 2005], consistency regularization [Sohn *et al.*, 2020].

Different from the existing SSL techniques which use all unlabeled data, DS3L uses it selectively and keeps tracking the effect of the supervised learning model to prevent performance hazards. Meanwhile, DS3L uses beneficial unlabeled data as much as possible to improve performance, preventing performance gains from being too conservative.

On one hand, DS3L uses the unlabeled selectively. The main methodology is to design a weighting function $w : \mathbb{R}^D \to \mathbb{R}$ parameterized by $\alpha \in \mathbb{B}^d$ that maps an instance to a weight. Then, DS3L tries to find the optimal $\hat{\theta}(\alpha)$ that minimizes the corresponding weighted empirical risk,

$$\hat{\theta}(\alpha) = \min_{\theta \in \Theta} \sum_{i=1}^{n} \ell(h(\mathbf{x}_i; \theta), \mathbf{y}_i) + \sum_{i=n+1}^{n+m} w(\mathbf{x}_i; \alpha) \Omega(\mathbf{x}_i; \theta) \quad (6)$$

where $\hat{\theta}(\alpha)$ is denoted as the model trained with the weight function paramaterized by α.

On the other hand, DS3L keeps tracking supervised performance to prevent performance degradation. Specifically, DS3L requires that the model returned by the weighted empirical risk process should maximize the generalization performance, i.e.,

$$\alpha^* = \operatorname*{argmin}_{\alpha \in \mathbb{B}^d} \mathbb{E}_{(X,Y)}[\ell(h(X; \hat{\theta}(\alpha)), Y)] \quad (7)$$

In real practice, the distribution is unknown, similar to the empirical risk minimization, DS3L tries to find the optimal parameters $\hat{\alpha}$ such that the model returned by optimizing the weighted instance loss, should also have a good performance on the labeled data which acts as an unbiased and reliable estimation of the underlying distribution, i.e.,

$$\hat{\alpha} = \operatorname*{argmin}_{\alpha \in \mathbb{B}^d} \sum_{i=1}^{n} \ell(h(\mathbf{x}_i; \hat{\theta}(\alpha)), \mathbf{y}_i) \quad (8)$$

To simplify the notation, we denote $\hat{\theta}(\alpha)$ as $\hat{\theta}$. Taking both the Eq.(6) and Eq.(8) into consideration, the objective of our framework can be formulated as the following bi-level optimization problem,

$$\min_{\alpha \in \mathbb{B}^d} \sum_{i=1}^{n} \ell(h(\mathbf{x}_i; \hat{\theta}), \mathbf{y}_i) \quad (9)$$

s.t.

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \sum_{i=1}^{n} \ell(h(\mathbf{x}_i; \theta), \mathbf{y}_i) + \sum_{i=n+1}^{n+m} w(\mathbf{x}_i; \alpha) \Omega(\mathbf{x}_i; \theta)$$

Eq.(9) can be understood by two stages: first, DS3L seeks the optimal model parameter $\hat{\theta}$ via the weighted empirical risk minimization, then evaluates it on n labeled instances and optimizes the weight function parameter α to make the learned $\hat{\theta}$ to achieve better reliable performance. Moreover, the bi-level optimization can be efficiently solved via stochastic gradient descent methods [Ren *et al.*, 2018].

In order to show the safeness of DS3L, we analyze the empirical risk of DS3L compared with the simple supervised method and obtain the following theorem,

Theorem 2. *Let θ^{SL} be the supervised model, i.e., $\theta^{SL} = \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} \ell(h(\mathbf{x}_i;\theta), \mathbf{y}_i)$. Define the empirical risk as:*

$$\hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^{n}[\ell(h(\mathbf{x}_i;\theta), \mathbf{y}_i)]$$

Then we have the empirical risk of $\hat{\theta}$ returned by DS3L *to be never worse than θ^{SL} that is learned from merely labeled data, i.e., $\hat{R}(\hat{\theta}) \leq \hat{R}(\theta^{SL})$.*

Theorem 2 reveals that compared with previous SSL methods, DS3L can achieve safeness in terms of empirical risk, i.e., the performance is not worse than its supervised counterpart, with the learned α.

We further analyze the generalization risk of DS3L to better understand the effect of the parameter dimension and the size of labeled data to α and drive the following theorem,

Theorem 3. *Assume that the loss function is λ-Lipschitz continuous w.r.t. α. Let $\alpha \in \mathbb{B}^d$ be the parameter of example weighting function w in a d-dimensional unit ball. Let n be the labeled data size. Define the generalization risk as:*

$$R(\theta) = \mathbb{E}_{(X,Y)}[\ell(h(X;\theta), Y)]$$

Let $\alpha^* = \arg\max_{\alpha\in\mathbb{B}^d} R(\hat{\theta}(\alpha))$ be the optimal parameter in the unit ball, and $\hat{\alpha} = \arg\max_{\alpha\in\mathcal{A}} \hat{R}(\hat{\theta}(\alpha))$ be the empirically optima among a candidate set $\mathcal{A}$. With probability at least $1-\delta$ we have,

$$R(\hat{\theta}(\alpha^*)) \leq R(\hat{\theta}(\hat{\alpha})) + \frac{(3\lambda + \sqrt{4d\ln(n) + 8\ln(2/\delta)})}{\sqrt{n}}$$

Theorem 3 establishes that DS3L approaches the optimal weight in the order $O(\sqrt{d\ln(n)/n})$. Based on theorem 2 and theorem 3, from both the safeness and generalization, it is reasonable to expect that DS3L can achieve better generalization performance compared with baseline supervised learning methods.

4 Open Problems

Although significant progress has been made in safe WSL with in-distribution data and out-of-distribution data, there still remain many open problems in this area.

- Safe WSL in dynamic environments. Learning in dynamic environments is far more difficult than in static ones. The challenges come from distribution drift, new class emerging, feature space change, and so on. There are some studies trying to tackle these problems [Da *et al.*, 2014], however, the issue of safeness remains an open problem for weakly-supervised learning in dynamic environments, e.g., an interesting problem is when the unlabeled data are useful in online learning.

- Automated safe WSL [Feurer *et al.*, 2015]. AutoML, which seeks to build an appropriate machine learning model for an unseen dataset in an automatic manner (without human intervention), has received increasing attention recently. However, existing AutoML systems focus on supervised learning, and existing AutoML techniques could not directly be used for the automated WSL problem. Efforts on automated WSL, remain limited right now. Automated WSL introduces some new challenges, e.g., various meta-features extracted from limited number of supervised data are no longer available and suitable; the use of auxiliary weakly supervised examples may sometimes even be outperformed by direct supervised learning. Therefore, safeness if one of the crucial aspects of AutoWSL, since it is not desirable to have an automated yet performance degenerated WSL system. [Li *et al.*, 2019] first present an automated learning system for SSL. They incorporate meta-learning with enhanced meta-features to help searching well-perform instantiations, and a large margin separation method to fine-tune the hyper-parameters as well as alleviate performance deterioration. More efforts are expected to be devoted to this direction.

- Safe deep WSL. Although we have introduced SAFEW and other related methods that aim to solve the safe WSL problem with in-distribution data. However, current safe WSL studies typically work on shallow models such as support vector machines, logistic regression, linear regression, etc. Applying WSL techniques to deep neural networks has attracted much attention in recent years for the promising results achieved by deep models. However, studies of safe WSL with deep neural networks remain to be limited. It is expected to design an efficient scheme for safe deep WSL.

- Safe imbalanced WSL. Previous WSL studies typically assume a balanced class distribution in both labeled sets and unlabeled sets. However, it is well-known that real-world dataset is often imbalanced or long-tailed. The performance of previous WSL studies is seriously decreased when the class distribution is imbalanced since their predictions are biased toward majority classes and result in low recall on minority classes [Kim *et al.*, 2020]. There are some efforts begin to address the imbalanced SSL problem [Kim *et al.*, 2020]. But how to achieve safe performance for imbalanced WSL is still under study and remains an open problem.

Acknowledgments

This work was partially supported by NSFC (61772262), and the Collaborative Innovation Center of Novel Software Technology and Industrialization. The author would like to thanks Lan-Zhe Guo for improving the paper.

References

[Balsubramani and Freund, 2015] Akshay Balsubramani and Yoav Freund. Optimally combining classifiers using unlabeled data. In *Proceedings of the 28th Conference on Learning Theory*, pages 211–225, 2015.

[Carbonneau *et al.*, 2018] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018.

[Chapelle *et al.*, 2006] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-Supervised Learning*. The MIT Press, 2006.

[Cozman *et al.*, 2003] Fábio Gagliardi Cozman, Ira Cohen, and Marcelo Cesar Cirelo. Semi-supervised learning of mixture models. In *Proceedings of the 20th International Conference on Machine Learning*, pages 99–106, Washington, DC, 2003.

[Da *et al.*, 2014] Qing Da, Yang Yu, and Zhi-Hua Zhou. Learning with augmented class by exploiting unlabeled data. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pages 1760–1766, 2014.

[Feurer *et al.*, 2015] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Tobias Springenberg, Manuel Blum, and Frank Hutter. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems*, pages 2962–2970, 2015.

[Frénay and Verleysen, 2014] Benoˆit Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014.

[Grandvalet and Bengio, 2005] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*, pages 529–536, 2005.

[Guo and Li, 2018] Lan-Zhe Guo and Yu-Feng Li. A general formulation for safely exploiting weakly supervised data. In *Proceedings of the AAAI conference on Artificial Intelligence*, pages 3126–3133, New Orleans, LA, 2018.

[Guo *et al.*, 2020] Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. In *Proceedings of The 37th International Conference on Machine Learning*, pages 3897–3906, 2020.

[Kim *et al.*, 2020] Jaehyung Kim, Youngbum Hur, Sejun Park, Eunho Yang, Sung Ju Hwang, and Jinwoo Shin. Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 14567–14579, 2020.

[Krijthe and Loog, 2017] Jesse H Krijthe and Marco Loog. Projected estimators for robust semi-supervised classification. *Machine Learning*, 106(7):993–1008, 2017.

[LeCun *et al.*, 2015] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[Li and Liang, 2019] Yu-Feng Li and De-Ming Liang. Safe semi-supervised learning: a brief introduction. *Frontiers Computer Science*, 13(4):669–676, 2019.

[Li and Zhou, 2015] 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.

[Li *et al.*, 2016] Yu-Feng Li, James T Kwok, and Zhi-Hua Zhou. Towards safe semi-supervised learning for multivariate performance measures. In *Proceedings of the AAAI conference on Artificial Intelligence*, pages 1816–1822, Phoenix, AZ, 2016.

[Li *et al.*, 2017] Yu-Feng Li, Han-Wen Zha, and Zhi-Hua Zhou. Learning safe prediction for semi-supervised regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2217–2223, San Francisco, CA, 2017.

[Li *et al.*, 2019] Yu-Feng Li, Hai Wang, Tong Wei, and Wei-Wei Tu. Towards automated semi-supervised learning. In *Proceedings of The Thirty-Third AAAI Conference on Artificial Intelligence*, pages 4237–4244, 2019.

[Li *et al.*, 2021] Yu-Feng Li, Lan-Zhe Guo, and Zhi-Hua Zhou. Towards safe weakly supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):334–346, 2021.

[Loog, 2015] Marco Loog. Contrastive pessimistic likelihood estimation for semi-supervised classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):462–475, 2015.

[Oliver *et al.*, 2018] Avital Oliver, Augustus Odena, Colin Raffel, Ekin Dogus Cubuk, and Ian J. Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, pages 3239–3250, 2018.

[Ren *et al.*, 2018] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4331–4340, 2018.

[Sohn *et al.*, 2020] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, pages 596–608, 2020.

[Wei *et al.*, 2018] Tong Wei, Lan-Zhe Guo, Yu-Feng Li, and Wei Gao. Learning safe multi-label prediction for weakly labeled data. *Machine Learning*, 107(4):703–725, 2018.

[Zhou, 2017] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2017.