Front. Comput. Sci.
https://doi.org/10.1007/s11704-019-8452-2

REVIEW ARTICLE

Safe semi-supervised learning: a brief introduction

Yu-Feng LI (✉)[1,2], De-Ming LIANG[1,2]

1 National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
2 Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 210023, China

Abstract Semi-supervised learning constructs the predictive model by learning from a few labeled training examples and a large pool of unlabeled ones. It has a wide range of application scenarios and has attracted much attention in the past decades. However, it is noteworthy that although the learning performance is expected to be improved by exploiting unlabeled data, some empirical studies show that there are situations where the use of unlabeled data may degenerate the performance. Thus, it is advisable to be able to exploit unlabeled data safely. This article reviews some research progress of safe semi-supervised learning, focusing on three types of safeness issue: data quality, where the training data is risky or of low-quality; model uncertainty, where the learning algorithm fails to handle the uncertainty during training; measure diversity, where the safe performance could be adapted to diverse measures.

Keywords machine learning, semi-supervised learning, safe

1 Introduction

Machine learning has achieved great success in numerous tasks, particularly in supervised learning tasks such as classification and regression. Typically, machine learning algorithms learn from a training data set with a certain amount of training examples. Each training example has a feature vector to describe the corresponding instance, and a label indicating the ground-truth output (the class to which the training example belongs in classification or the real-valued response corresponding to the example in regression). The most famous technique in recent years, deep learning [1], makes enormous commercial success in computer vision based on the fact that ground-truth labels for a broad training data set are provided. However, in various tasks such as disease diagnosis, auto-vehicle, physical system, it is hard to attain considerable ground-truth information due to the cost of time or expense. Thus, it is desirable to be able to learn a good model with a few labeled data. Semi-supervised learning (SSL) [2] is a promising direction which attempts to exploit labeled data (often of limited amount) and a huge pool of unlabeled data for optimizing machine learning models. Existing semi-supervised approaches can be roughly grouped into four categories, including generative models [3, 4], semi-supervised SVMs [5, 6], graph-based semi-supervised methods [7–11] and disagreement-based methods [12, 13].

It is generally recognized that by using unlabeled data, semi-supervised learning can help improve the performance, particularly when the number of labeled data is limited. However, it is noteworthy that although the learning performance is expected to be improved by exploiting unlabeled data, some empirical studies [4, 9, 14–20] show that there are cases in which the use of the unlabeled data may degenerate the performance, making it even worse after semi-supervised learning. This phenomenon prevents the deployment of semi-supervised learning in real-world applications because the cost of utilizing unlabeled data may not be rewarded sometimes. It is desirable to have **safe semi-supervised learning** approaches which never reduce learning performance significantly when using unlabeled data. *Safe*, here means that the generalization performance is never statistically significantly worse than methods using only labeled data. It is meaning-

Received December 27, 2018; accepted March 5, 2019

E-mail: liyf@nju.edu.cn

less to talk about a single trial because for a single trial, even exploiting more labeled data might result in a worse performance. The issue of safeness has been raised and studied for many years [21]; however, only recently some solid progress has been reported. In this article, we will discuss some progress in this line of research.

Existing studies try to improve the safeness of algorithms from three aspects: data quality, model uncertainty and measure diversity. As for data quality, the graph used in graph-based SSL and risky unlabeled samples may degenerate the performance [22–24]. In the model part, we now understand that the exploitation of unlabeled data naturally leads to more than one model option, and inadequate choice may lead to poor performance [25]; In practical applications, the performance measures are often diverse, so the safeness should also be considered under different measures [26].

For simplicity, in this article we consider binary classification concerning two exchangeable classes $\boldsymbol{Y}$ and $\boldsymbol{N}$. Formally, let $\mathcal{X}$ be the input space and $\mathcal{Y} = \{\pm\}$ be the label space. Given a set of l labeled instances $\{\boldsymbol{x}_i, y_i\}_{i=1}^{l}$ and u unlabeled instances $\{\boldsymbol{x}_j\}_{j=l+1}^{l+u}$, semi-supervised learning algorithms aim to find a decision function $f : \mathcal{X} \to \{\pm 1\}$ and a label assignment on unlabeled instances $\boldsymbol{y_u} = \{y_{l+1}\ , \ldots, y_{l+u}\} \in \mathcal{B}$. We use identifier $perf(\cdot)$ as the target performance measure. Figure 1 provides an illustration of three aspects of the safeness problem in semi-supervised learning, which we will discuss in this article. Note that although they are discussed separately for clarity, they often occur simultaneously.

2 Data quality

Semi-supervised learning algorithms try to exploit vast unlabeled data to help machine learning models to improve performance. Actually, there are two basic assumptions in semi-supervised learning, i.e., the *cluster assumption* and the *manifold assumption*; both concerns about data distribution. The former assumes that data have inherent cluster structure and instances falling into the same cluster have the same class label. The latter assumes that the data lie on a manifold and nearby instances have similar predictions. Semi-supervised learning helps when the assumption matches the inherent structure of the data. In other words, when it was broken, semi-supervised learning algorithms would not perform as the user wish.

Some progress has been made in the field of graph-based semi-supervised learning (GSSL). GSSL attracts significant attention since it was proposed. Many works have been presented on optimizing the label prediction of the unlabeled data [5,7–9,11,27]. Besides, various kinds of graphs and metrics have been considered [27–29]. It is now widely accepted that the quality of the graph seriously affects the performance of GSSL methods [8,27,29,30], and many empirical studies have shown that an inappropriate graph can even cause a degenerated performance [11,30–32]. At a deep level, different graphs have diverse manifold structures. When the manifold structure of the graph breaks the manifold assumption in one

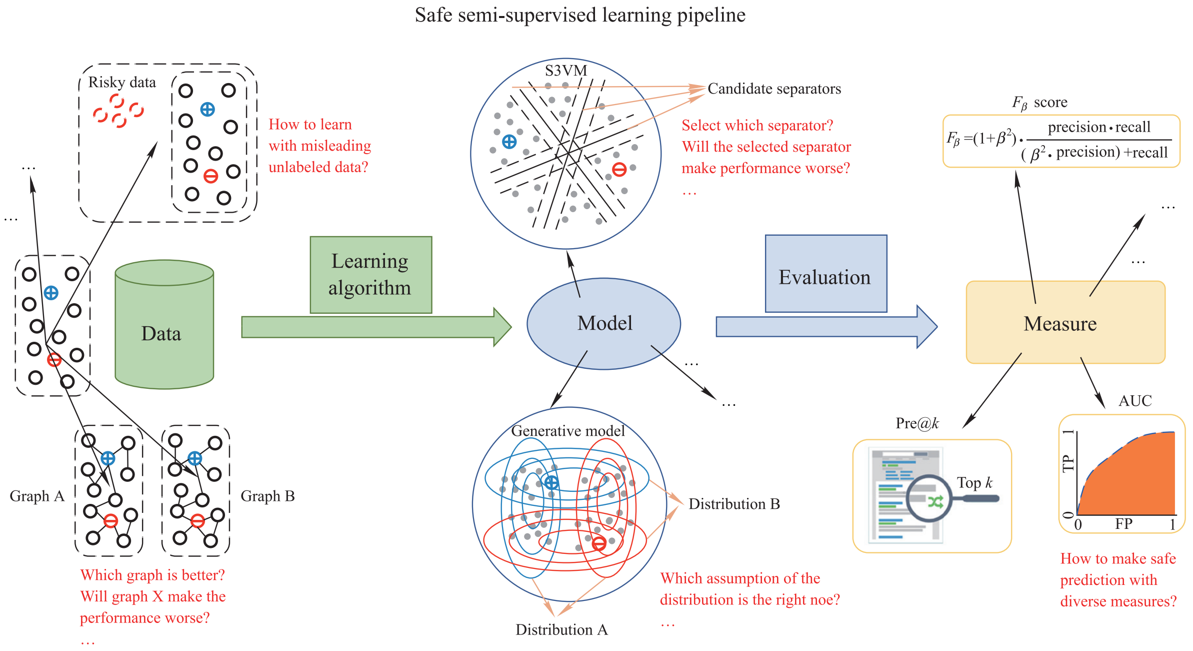


Fig. 1 Semi-supervised learning pipeline that displays the sources of unsafeness: data quality, model uncertainty, and measure diversity

dataset, GSSL will make a poor performance. In order to improve safeness, an important problem is the way to identify the quality of the graph before we get the ultimate performance.

Li et al. [22] firstly study the quality of the graph. They propose a large margin based method for graph-judging. The basic idea is that given a set of candidate graphs, when one graph has a high quality, its predictive results may have a large margin separation as shown in Fig. 2. Therefore, given multiple graphs with unknown quality, one should encourage to use the graphs with a large margin, rather than the graphs with a small margin, and reduce the chances of performance degradation consequently. They propose a stacking method, which first regenerates a new SSL data set with the predictive results of GSSL on candidate graphs, and then formulates safe GSSL as the classical Semi-Supervised SVM optimization on the regenerated dataset. The experimental results demonstrate that a large margin principle is helpful in judging the graph quality and improving the safeness of GSSL.

Later, Liang and Li [33] formulate the safe graph construction as a minimax optimization problem. They optimize the worst-case nearest-neighbor error on candidate graphs, and the optimal solution can be regarded as conventional GSSL methods on a "safe" graph. Guo et al. [34] propose a graph construction method based on the large margin principle. They optimize a margin-type loss to learn a graph that has a large margin on its prediction.

Apart from the large margin principle in graph quality judgment, there are some efforts focusing on data selection. Wang et al. [23] proposed a safe GSSL method named GsslIs concentrating on sample selection. The basic idea is that given a set of unlabeled instances, it is not the best to exploit all the unlabeled instances; Instead, the unlabeled instances which are highly possible to help improve the performance, while do not take the ones with high risk into account. GsslIs constructed multiple GSSL classifiers to identify the risky unlabeled samples and tried to reduce the degeneration probability. Li and Zhou [24] propose S3VM with unlabeled data selection. It uses hierarchical clustering to estimate the reliability of unlabeled instances and then removes the ones with the lowest reliability. Only the unlabeled instances with high confidence by hierarchical clustering are predicted by TSVM, and the rest are predicted by SVM.

3 Model uncertainty

From the model view, the uncertainty of the SSL model makes it unstable in different trials and incorrect selection under uncertainty degenerate the performance of SSL. For example, for generative methods, Cozman et al. [21] conjectured that the performance degradation is caused by the use of incorrect model assumption in the generative methods; however, it is almost impossible to find the correct model assumption without adequate domain knowledge. For disagreement-based methods [35], such as co-training [12], Zhou and Li [13] realized that incorrect pseudo-labels used in the SSL process are the root of performance degradation. For semi-supervised SVMs (S3VMs), Li and Zhou [36] showed that the existence of multiple low-density separators and the incorrect selection will lead to performance degradation. Figure 3 visualizes the uncertainty of S3VM: there is more than one large margin separator, and these separators are usually diverse. Conventional S3VMs select one of them as the output, while incorrect selection degenerates the performance. The uncertainty of separator results in unsafeness.

Li et al. [25, 36] firstly develop the safe S3VM (S4VM) method. S4VMs first generate a pool of diverse large margin low-density separators, and then optimizes the label

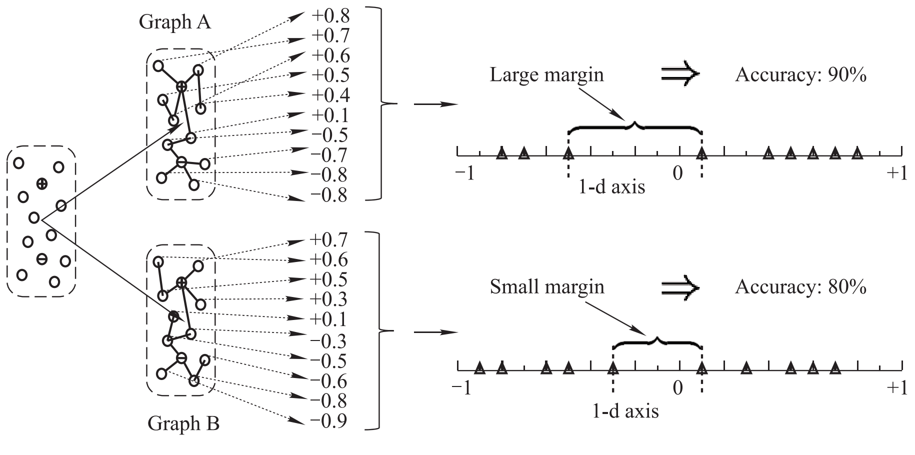


Fig. 2 Large margin principle for graph quality judgement

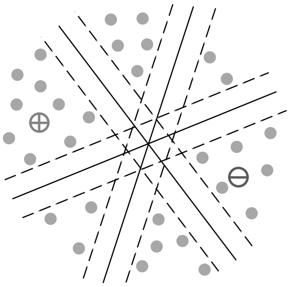


Fig. 3 Example of model uncertainty: there are multiple candidate large margin separators for S3VM

assignment for the unlabeled data in the worse case under the assumption that the ground-truth label assignment can be realized by one of the obtained low-density separators. Formally, let $\boldsymbol{y}_{\boldsymbol{u}}^{*}$ be the ground-truth label assignment and $\hat{\boldsymbol{y}}_{\boldsymbol{u}}^{svm}$ be the predictive labels of inductive SVM on unlabeled instances. A pool of T low-density predictions $\{\hat{\boldsymbol{y}}_{\boldsymbol{u}}^{t}\}_{t=1}^{T}$ is employed by existing S3VMs. For any label assignment of unlabeled instances $\hat{\boldsymbol{y}}_{\boldsymbol{u}}$, denote $gain(\hat{\boldsymbol{y}}_{\boldsymbol{u}}, \boldsymbol{y}_{\boldsymbol{u}}^{*}, \hat{\boldsymbol{y}}_{\boldsymbol{u}}^{svm})$ and $loss(\hat{\boldsymbol{y}}_{\boldsymbol{u}}, \boldsymbol{y}_{\boldsymbol{u}}^{*}, \hat{\boldsymbol{y}}_{\boldsymbol{u}}^{svm})$ as the gained and lost accuracies compared to the inductive SVM. The goal is to learn a label assignment $\hat{\boldsymbol{y}}_{\boldsymbol{u}}$ such that the improved performance against the inductive SVM is maximized,

$$\max_{\hat{\boldsymbol{y}}_{\boldsymbol{u}} \in \{\pm 1\}^{u}} gain(\hat{\boldsymbol{y}}_{\boldsymbol{u}}, \boldsymbol{y}_{\boldsymbol{u}}^{*}, \hat{\boldsymbol{y}}_{\boldsymbol{u}}^{svm}) - loss(\hat{\boldsymbol{y}}_{\boldsymbol{u}}, \boldsymbol{y}_{\boldsymbol{u}}^{*}, \hat{\boldsymbol{y}}_{\boldsymbol{u}}^{svm}), \qquad (1)$$

where λ is a parameter for trading-off how much risk the user would like to undertake. Denote $gain(\hat{\boldsymbol{y}}_{\boldsymbol{u}}, \boldsymbol{y}_{\boldsymbol{u}}^{*}, \hat{\boldsymbol{y}}_{\boldsymbol{u}}^{svm})-$ $loss(\hat{\boldsymbol{y}}_{\boldsymbol{u}}, \boldsymbol{y}_{\boldsymbol{u}}^{*}, \hat{\boldsymbol{y}}_{\boldsymbol{u}}^{svm})$ as $J(\hat{\boldsymbol{y}}_{\boldsymbol{u}}, \boldsymbol{y}_{\boldsymbol{u}}^{*}, \hat{\boldsymbol{y}}_{\boldsymbol{u}}^{svm})$ for simplicity notations.

Because the ground-truth $\boldsymbol{y}_{\boldsymbol{u}}^{*}$ is unknown, S4VM assumes that the ground-truth $\boldsymbol{y}_{\boldsymbol{u}}^{*}$ is realized by a low-density separator, i.e., $\boldsymbol{y}_{\boldsymbol{u}}^{*} \in \mathcal{M} \triangleq \{\hat{\boldsymbol{y}}_{\boldsymbol{u}}^{t}\}_{t=1}^{T}$. Without further domain knowledge in distinguishing these separators, the *worst−case* improvement over inductive SVM is maximized and the optimal solution

$$\overline{\boldsymbol{y}}_{\boldsymbol{u}} = \arg \max_{\boldsymbol{y}_{\boldsymbol{u}} \in \{\pm 1\}^{u}} \min_{\hat{\boldsymbol{y}}_{\boldsymbol{u}} \in \mathcal{M}} J(\boldsymbol{y}_{\boldsymbol{u}}, \hat{\boldsymbol{y}}_{\boldsymbol{u}}, \hat{\boldsymbol{y}}_{\boldsymbol{u}}^{svm}). \qquad (2)$$

Considering that cluster assumption cannot reflect the real-data distribution adequately in some case, Wang et al. [37] suggested a modified cluster assumption that similar instance should share similar class memberships rather than a crisp class label and accordingly developed a new Semi-Supervised Classification method based on class memberships (SSCCM). In SSCCM, each instance can belong to multiple classes with the corresponding class membership, and each instance and its local weight mean share the same label membership vector. However, SSCCM also yields worse performance than its supervised baseline. Later, they improve SSCCM with a safety-control mechanism for safe semi-supervised classification by adaptively controlling the trade-off between semi-supervised and supervised classification in terms of the existing unlabeled data [38]. They expect the final prediction to approach to that of the supervised counterparts (LS-SVM) when given unreliable unlabeled data, and approach to that of the semi-supervised method (SSCCM) when given reliable unlabeled data.

There are some other related studies trying to improve the stability of semi-supervised learning: Balsubramani and Freund [39] propose a new method to utilize unlabeled examples by combining an ensemble of classifiers. They formulate the task as a game played over a set of unlabeled data, and the supervised information is encoded as the constraint on the game. The minimax solution of the game is assured to be significantly better than any single base classifier. Niu et al. [40] give a theoretical study about when positive unlabeled learning outperforms positive negative learning. Kawakita and Takeuchi [41] propose a safe semi-supervised learning method based on the weighted likelihood which was expected to be safe in any situation. They prove that their proposal is asymptotically safe under certain conditions.

The works mentioned above focus on semi-supervised classification. For semi-supervised regression (SSR), Li et al. [42] focus on performance degradation when using unlabeled data. They try to learn a safe prediction given a set of SSR predictions obtained in various ways. They cast the safe semi-supervised regression problem as a geometric projection issue. When the ground-truth label assignment is realized by a convex linear combination of base regressors, the proposal is probably safe and achieve the maximal worst-case performance gain. Figure 4 illustrates the intuition of the proposed method via the viewpoint of geometric projection. $\mathbf{f_0}$ is certain direct supervised regression prediction, $\{\mathbf{f}_i\}_{i=1}^{b}$ are multiple SSR predictions, $\mathbf{f}^{*} \in \Omega$ is the ground-truth label assignments, where $\Omega = \{\mathbf{f} | \sum_{i=1}^{b} \alpha_i \mathbf{f}_i, \boldsymbol{\alpha} \in \mathcal{M}\}$, $\mathcal{M} = \{\boldsymbol{\alpha} | \sum_{i=1}^{b} \alpha_i = 1, \alpha_i \geqslant 0\}$, and $\overline{\mathbf{f}}$ is the target prediction of the proposal. $\|\overline{\mathbf{f}} - \mathbf{f}^{*}\|^2 \leqslant \|\mathbf{f}_0 - \mathbf{f}^{*}\|$ is true, which reveals that the proposal is probably safe [42].

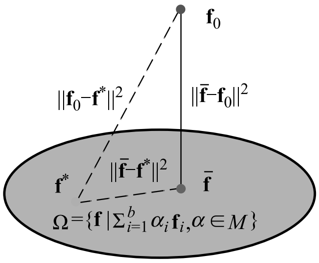

$\mathbf{f}_0$
$\|\mathbf{f}_0-\mathbf{f}^{*}\|^2$
$\|\overline{\mathbf{f}}-\mathbf{f}_0\|^2$
$\mathbf{f}^{*}$
$\|\overline{\mathbf{f}}-\mathbf{f}^{*}\|^2$
$\overline{\mathbf{f}}$
$\Omega=\{\mathbf{f}|\Sigma_{i=1}^{b}\alpha_i\mathbf{f}_i,\alpha\in M\}$

Fig. 4 The projection viewpoint of safe SSR: $\overline{\mathbf{f}}$, the projection of $\mathbf{f}_0$ onto the convex feasible set $\boldsymbol{\Omega}$, is closer to $\mathbf{f}^{*}$ than $\mathbf{f}_0$ is

4 Measure diversity

In the classification and regression task, square loss is usually employed. But in many practical applications, the performance measures are usually diverse. For example, in text categorization, the F1-score and precision-recall are typically used; in information retrieval, precision and recall are more preferred; in ranking applications, the area under the ROC curve (AUC) and Top-k precision are more popular. One needs to develop a safe SSL method which can work with such a diversity of performance measures.

Li et al. [26] propose a method named UMVP which integrates multiple semi-supervised learners and maximizes the worst-case performance gain to derive the final prediction, where the performance gain (with respect to various multivariate performance criteria) is maximized relative to the baseline supervised model in the worst-case scenario. Then, a tight minimax convex relaxation technique is adopted to solve the optimization problem. They show that when the performance measure is the Top-*k* Precision, F_β score or AUC, this convex relaxation problem can be solved in closed-form solutions and small linear programs.

Specially, the UMVP method uses b semi-supervised learners to construct safe predictions (they can be obtained by running different SSL algorithms with different parameters) and assume that the ground-truth label assignment can be realized by a convex linear combination of base regressors. Let $perf$ be the performances measure (e.g., Top-*k* precision, F_β, AUC). Without loss of generality, assume that the larger the $perf$ value, the better the performance. The goal is to find a prediction $\hat{\boldsymbol{y}}_{\boldsymbol{u}}$ which maximally aligns with predictions from the b semi-supervised learners, and also performs better than a given baseline learner whose prediction is $\hat{\boldsymbol{y}}^0_{\boldsymbol{u}}$. The goal can be formulated as the following optimization problem

$$\max_{\hat{\boldsymbol{y}}_{\boldsymbol{u}} \in \mathcal{Y}} \sum_{i=1}^{b} \alpha_i \left(perf(\hat{\boldsymbol{y}}_{\boldsymbol{u}}, \boldsymbol{y}^i_{\boldsymbol{u}}) - perf(\hat{\boldsymbol{y}}^0_{\boldsymbol{u}}, \boldsymbol{y}^i_{\boldsymbol{u}}) \right), \tag{3}$$

where $\{\boldsymbol{y}^1_{\boldsymbol{u}}, \ldots, \boldsymbol{y}^b_{\boldsymbol{u}}\}$ are predictions of the b semi-supervised learners on the unlabeled instances, $\boldsymbol{y}^i_{\boldsymbol{u}} = \left[y^i_{l+1}, \ldots, y^i_{l+u}\right] \in \{0,1\}^u$. If $\boldsymbol{y}^i$ is the ground-truth label assignment, then $\left(perf(\hat{\boldsymbol{y}}_{\boldsymbol{u}}, \boldsymbol{y}^i_{\boldsymbol{u}}) - perf(\hat{\boldsymbol{y}}^0_{\boldsymbol{u}}, \boldsymbol{y}^i_{\boldsymbol{u}}) \right)$ is the performance gain of $\hat{\boldsymbol{y}}$ relative to $\boldsymbol{y}^0$. $\boldsymbol{\alpha}$ captures the relative importance of the b learners and it is in the simplex $\mathcal{M} = \{\boldsymbol{\alpha} | \sum_{i=1}^{b} \alpha_i = 1, \alpha_i \geqslant 0\}$.

The relative importance of base learners is unknown. To address the problem, UMVP considers the worst-case, adversarial setting of $\boldsymbol{\alpha}$ and get the following maximin optimization problem:

$$\max_{\hat{\boldsymbol{y}}_{\boldsymbol{u}} \in \mathcal{Y}} \min_{\boldsymbol{\alpha} \in \mathcal{M}} \sum_{i=1}^{b} \alpha_i \left(perf(\hat{\boldsymbol{y}}_{\boldsymbol{u}}, \boldsymbol{y}^i_{\boldsymbol{u}}) - perf(\hat{\boldsymbol{y}}^0_{\boldsymbol{u}}, \boldsymbol{y}^i_{\boldsymbol{u}}) \right). \tag{4}$$

When one of the semi-supervised learners realizes the ground-truth label assignment, the UMVP solution is safe. It is challenging to see which assumption is more suitable for a particular dataset when doing SSL. Equation (4) shows that as long as one of these assumptions realizes a perfect solution, a safe SSL method can be obtained.

Briefly, after convex relaxation and using the cutting-plane algorithm, we need to solve the following optimization problem given the current $\boldsymbol{\alpha}$,

$$\arg\max_{\hat{\boldsymbol{y}}_{\boldsymbol{u}} \in \mathcal{Y}} \sum_{i=1}^{b} \alpha_i \left(perf(\hat{\boldsymbol{y}}_{\boldsymbol{u}}, \boldsymbol{y}^i_{\boldsymbol{u}}) - perf(\hat{\boldsymbol{y}}^0_{\boldsymbol{u}}, \boldsymbol{y}^i_{\boldsymbol{u}}) \right), \tag{5}$$

then the problem can be finally reduced to the simpler one:

$$\max_{\hat{\boldsymbol{y}}_{\boldsymbol{u}} \in \mathcal{Y}} \sum_{i=1}^{b} \alpha_i \; perf(\hat{\boldsymbol{y}}_{\boldsymbol{u}}, \boldsymbol{y}^i). \tag{6}$$

When the performance measure is either the Top-*k* Precision, F_β score or AUC, Eq. (6) has the closed-form optimal solutions displayed in Table 1 (refer to [26] for more details), where π^s is the ranking vector given prediction $\boldsymbol{s}$, constant $c = \hat{\boldsymbol{y}}^{i\prime}_{\boldsymbol{u}}\mathbf{1}$, $P_{\boldsymbol{y}_{\boldsymbol{u}}}$ and $N_{\boldsymbol{y}_{\boldsymbol{u}}}$ are the numbers of positive and negative labels in $\boldsymbol{y}_{\boldsymbol{u}}$, respectively.

It is worthy to note that existing solutions all focus on making a safe prediction under a single measure. A common safe semi-supervised algorithm whose performance is probably safe under different measures remains an open question now.

5 Related fields

In this section, we discuss some related fields with safe semi-supervised learning.

Table 1 Summaries of safe optimal solutions with different measures

Inputs	Optimal safe output $\hat{\boldsymbol{y}}^*_{\boldsymbol{u}}$ with different measure		
	Top-*k* Precision	F_β Score	AUC
Predictions of base learners $\{\boldsymbol{y}^1_{\boldsymbol{u}}, \cdots, \boldsymbol{y}^b_{\boldsymbol{u}}\}$	any $\hat{\boldsymbol{y}}_{\boldsymbol{u}}$ ranks the unlabeled instances as $\boldsymbol{s} = \sum_{i=1}^{b} \alpha_i \boldsymbol{y}^i_{\boldsymbol{u}}$	$\hat{\boldsymbol{y}}^*_{\boldsymbol{u}} = [\hat{y}^*_j], \hat{y}^*_j = 1$, if $\pi^s_j > u - c$, $\boldsymbol{s} = \sum_{i=1}^{b} \frac{\alpha_i(1+\beta^2)}{c+\beta^2 P_{\boldsymbol{y}^i_{\boldsymbol{u}}}} \boldsymbol{y}^i_{\boldsymbol{u}}$	any $\hat{\boldsymbol{y}}_{\boldsymbol{u}}$ that ranks the unlabeled instances as $\boldsymbol{s} = \sum_{i=1}^{b} \frac{\alpha_i}{P_{\boldsymbol{y}^i_{\boldsymbol{u}}} N_{\boldsymbol{y}^i_{\boldsymbol{u}}}} \boldsymbol{y}^i_{\boldsymbol{u}}$

Weakly supervised learning is closely related to semi-supervised learning. Semi-supervised learning can be included in weakly supervised learning as a type of machine learning under incomplete supervision [43], i.e., only a subset of training data is given with labels. Other types of weak supervision include inexact supervision, i.e., only coarse-grained labels are given, and inaccurate supervision, i.e., the given labels are not always ground-truth. Weakly supervised learning covers a variety of studies which attempt to build predictive models under weak supervision. Weakly supervised data are widespread in real-world applications apart from semi-supervised learning, such as label noise learning [44], domain adaptation [45] and so on. Similarly, learning with the help of weakly supervised data may also cause performance degradation. For example, label noise learning may be worse than learning from only a small number of noise-free data [44]; domain adaptation methods may have the phenomenon of negative transfer [45]. The way to *safely* exploiting weakly supervised data thus is an important problem.

There are some studies on safe weakly supervised learning in recent years. Existing work focuses on integrating base learners to improve safeness. Guo and Li [46] present a general scheme that builds the final prediction results by integrating several weakly supervised learners. The proposed formulation provides safeness guarantees for commonly used loss functions (i.e., square loss, hinge loss) supposing that the ground-truth can be constructed by the base learners. Wei et al. [47] 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. In the scenarios of large-scale multi-label learning, labels have different impacts on the commonly used performance metric, and there are a few studies utilizing this feature to reduce the model size [48,49]. However, the impact of distinct labels on safeness when exploiting weakly supervised information remains uncertain.

Clustering analysis plays an important role in unsupervised learning and aims to discover a group structure of the dataset, which can be helpful when learning with few labels. Similarly, without prior-knowledge about a dataset, it is hard to judge the quality of results generated by different clustering methods, and there is still much room for improving the clustering performance for categorical data. Some effort has been paid in this line of research [50,51].

Automated machine learning (AutoML) [52, 53], 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 SSL problem. Efforts on automated SSL remain limited right now. Automated SSL introduces some new challenges, i.e., various meta-features extracted from a number of labeled examples are no longer available and suitable [53]; the use of auxiliary unlabeled instances may sometimes even be outperformed by direct supervised learning. Therefore, safeness is one of the crucial aspects of AutoSSL, since it is not desirable to have an automated yet performance degenerated SSL system. Li et al. [54] 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 hyperparameters as well as alleviate performance deterioration.

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 [55–57] when learning in data streams, however, the issue of safeness remains an open problem for semi-supervised learning in dynamic environments, i.e., an interesting problem is when the unlabeled data are useful in online learning.

6 Conclusion

Semi-supervised learning has a wide range of application scenarios and has attracted considerable attention in the past decades. However, phenomena of performance degeneration hinder the deployment of semi-supervised learning in real applications. It is desirable to be able to exploit unlabeled data safely.

This article focuses on three viewpoints to tackle the issue of safeness in semi-supervised learning: data quality, model uncertainty, and measure diversity. Although we consider them separately, in practice they often occur simultaneously. Note that due to the space limit, this article actually serves more like a literature index rather than a comprehensive review. Readers interested in some details are encouraged to read the corresponding references. In the future, it will be worth understanding the boundary of safe SSL and making valuable decisions for SSL data.

Acknowledgements This research was supported by the National Key R&D Program of China (2017YFB1001903), the National Natural Science Foundation of China (Grant No. 61772262) and the Fundamental Research Funds for the Central Universities.

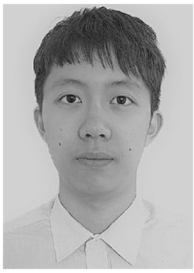
References

1. Goodfellow I, Bengio Y, Courville A. Deep Learning. MA: MIT Press, 2016
2. Chapelle O, Schölkopf B, Zien A. Semi-supervised Learning. MA: MIT Press, 2006
3. Miller D J, Uyar H S. A mixture of experts classifier with learning based on both labelled and unlabelled data. In: Proceedings of the 10th Annual Conference on Neural Information Processing Systems. 1996, 571–577
4. Nigam K, McCallum A, Thrun S, Mitchell T M. Text classification from labeled and unlabeled documents using EM. Machine Learning, 2000, 39(2–3): 103–134
5. Joachims T. Transductive inference for text classification using support vector machines. In: Proceedings of the 16th International Conference on Machine Learning. 1999, 200–209
6. Bennett K P, Demiriz A. Semi-supervised support vector machines. In: Proceedings of the 11th International Conference on Neural Information Processing Systems. 1998, 368–374
7. Zhu X J, Ghahramani Z, Lafferty J D. Semi-supervised learning using gaussian fields and harmonic functions. In: Proceedings of the 20th International Conference on Machine Learning. 2003, 912–919
8. Belkin M, Niyogi P, Sindhwani V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research, 2006, 7(Nov): 2399–2434
9. Blum A, Chawla S. Learning from labeled and unlabeled data using graph mincuts. In: Proceedings of the 18th International Conference on Machine Learning. 2001, 19–26
10. Liu W, Wang J, Chang S F. Robust and scalable graph-based semi-supervised learning. Proceedings of the IEEE, 2012, 100(9): 2624–2638
11. Zhou D, Bousquet O, Lal T N, Weston J, Schölkopf B. Learning with local and global consistency. Advances in Neural Information Processing Systems, 2004, 16: 321–328
12. Blum A, Mitchell T M. Combining labeled and unlabeled data with co-training. In: Proceedings of the 11th Annual Conference on Computational Learning Theory. 1998, 92–100
13. Zhou Z H, Li M. Tri-training: exploiting unlabeled data using three classifiers. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(11): 1529–1541
14. Singh A, Nowak R D, Zhu X. Unlabeled data: now it helps, now it doesn't. In: Proceedings of the 21st International Conference on Neural Information Processing Systems. 2008, 1513–1520
15. Yang T, Priebe C E. The effect of model misspecification on semi-supervised classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(10): 2093–2103
16. Chapelle O, Sindhwani V, Keerthi S S. Optimization techniques for semi-supervised support vector machines. Journal of Machine Learning Research, 2008, 9: 203–233
17. Chawla N V, Karakoulas G I. Learning from labeled and unlabeled data: an empirical study across techniques and domains. Journal of Artificial Intelligence Research, 2005, 23: 331–366
18. Chen K, Wang S. Semi-supervised learning via regularized boosting working on multiple semi-supervised assumptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(1): 129–143
19. Cozman F G, Cohen I, Cirelo M C. Semi-supervised learning of mixture models. In: Proceedings of the 20th International Conference on Machine Learning. 2003, 99–106
20. Grandvalet Y. Semi-supervised learning by entropy minimization. Advances in Neural Information Processing Systems, 2005, 17: 529–536
21. Cozman F G, Cohen I, Cirelo M. Unlabeled data can degrade classification performance of generative classifiers. In: Proceedings of the 15th International Florida Artificial Intelligence Research Society Conference. 2002, 327–331
22. Li Y F, Wang S B, Zhou Z H. Graph quality judgement: a large margin expedition. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence. 2016, 1725–1731
23. Wang H, Wang S B, Li Y F. Instance selection method for improving graph-based semi-supervised learning. Frontiers of Computer Science, 2018, 12(4): 725–735
24. Li Y F, Zhou Z H. Improving semi-supervised support vector machines through unlabeled instances selection. In: Proceedings of the 25th AAAI Conference on Artificial Intelligence. 2011, 386–391
25. Li Y F, Zhou Z H. Towards making unlabeled data never hurt. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(1): 175–188
26. Li Y F, Kwok J T, Zhou Z H. Towards safe semi-supervised learning for multivariate performance measures. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence. 2016, 1816–1822
27. Jebara T, Wang J, Chang S F. Graph construction and *b*-matching for semi-supervised learning. In: Proceedings of the 26th Annual International Conference on Machine Learning. 2009, 441–448
28. Carreira-Perpi nán M Á, Zemel R S. Proximity graphs for clustering and manifold learning. In: Proceedings of the 17th International Conference on Neural Information Processing Systems. 2004, 225–232
29. Zhu X. Semi-supervised learning literature survey. Computer Science, University of Wisconsin-Madison, 2006, 2(3): 4
30. Wang F, Zhang C. Label propagation through linear neighborhoods. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(1): 55–67
31. Belkin M, Niyogi P. Semi-supervised learning on riemannian manifolds. Machine Learning, 2004, 56(1-3): 209–239
32. Karlen M, Weston J, Erkan A, Collobert R. Large scale manifold transduction. In: Proceedings of the 25th International Conference on Machine Learning. 2008, 448–455
33. Liang D M, Li Y F. Learning safe graph construction from multiple graphs. In: Proceedings of the International CCF Conference on Artificial Intelligence. 2018, 41–54
34. Guo L Z, Wang S B, Li Y F. Large margin graph construction for semi-supervised learning. In: Proceedings of the International Workshop on Large Scale Graph Representation Learning and Applications. 2018, 1030–1033
35. Zhou Z H, Li M. Semi-supervised learning by disagreement. Knowl-

edge and Information Systems, 2010, 24(3): 415–439

36. Li Y F, Zhou Z H. Towards making unlabeled data never hurt. In: Proceedings of the 28th International Conference on Machine Learning. 2011, 1081–1088
37. Wang Y, Chen S, Zhou Z H. New semi-supervised classification method based on modified cluster assumption. IEEE Transactions on Neural Networks and Learning Systems, 2012, 23(5): 689–702
38. Wang Y, Meng Y, Fu Z, Xue H. Towards safe semi-supervised classification: adjusted cluster assumption via clustering. Neural Processing Letters, 2017, 46(3): 1031–1042
39. Balsubramani A, Freund Y. Optimally combining classifiers using unlabeled data. In: Proceedings of the 28th Conference on Learning Theory. 2015, 211–225
40. Niu G, Plessis d M C, Sakai T, Ma Y, Sugiyama M. Theoretical comparisons of positive-unlabeled learning against positive-negative learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016, 1207–1215
41. Kawakita M, Takeuchi J. Safe semi-supervised learning based on weighted likelihood. Neural Networks, 2014, 53: 146–164
42. Li Y F, Zha H W, Zhou Z H. Learning safe prediction for semi-supervised regression. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence. 2017, 2217–2223
43. Zhou Z H. A brief introduction to weakly supervised learning. National Science Review, 2017, 5(1): 44–53
44. Frénay B, Verleysen M. Classification in the presence of label noise: a survey. IEEE Transactions on Neural Networks and Learning Systems, 2014, 25(5): 845–869
45. Pan S J, Yang Q. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345–1359
46. Guo L Z, Li Y F. A general formulation for safely exploiting weakly supervised data. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. 2018, 3126–3133
47. Wei T, Guo L Z, Li Y F, Gao W. Learning safe multi-label prediction for weakly labeled data. Machine Learning, 2018, 107(4): 703–725
48. Wei T, Li Y F. Does tail label help for large-scale multi-label learning. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. 2018, 2847–2853
49. Wei T, Li Y F. Learning from semi-supervised weak-label data. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence. 2019
50. Li F, Qian Y, Wang J, Dang C, Liu B. Cluster's quality evaluation and selective clustering ensemble. ACM Transactions on Knowledge Discovery from Data, 2018, 12(5): 60
51. Qian Y, Li F, Liang J, Liu B, Dang C. Space structure and clustering of categorical data. IEEE Transactions on Neural Networks and Learning Systems, 2016, 27(10): 2047–2059
52. Yao Q, Wang M, Chen Y, Dai W, Hu Y Q, Li Y F, Tu W W, Yang Q, Yu Y. Taking human out of learning applications: a survey on automated machine learning. 2018, ArXiv preprint ArXiv: 1810.13306
53. Feurer M, Klein A, Eggensperger K, Springenberg J T, Blum M, Hutter F. Efficient and robust automated machine learning. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. 2015, 2755–2763
54. Li Y F, Wang H, Wei T, Tu W W. Towards automated semi-supervised learning. In: Proceedings of the 33rd Conference on Artificial Intelligence. 2019
55. Da Q, Yu Y, Zhou Z H. Learning with augmented class by exploiting unlabeled data. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence. 2014, 1760–1766
56. Zhu Y, Ting K M, Zhou Z H. New class adaptation via instance generation in one-pass class incremental learning. In: Proceedings of the IEEE International Conference on Data Mining. 2017, 1207–1212
57. Zhu Y, Ting K M, Zhou Z H. Multi-label learning with emerging new labels. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(10): 1901–1914

Yu-Feng Li received the BSc and PhD degrees in computer science from Nanjing University, China in 2006 and 2013, respectively. He joined the Department of Computer Science & Technology at Nanjing University as an Assistant Researcher in 2013, and is currently associate professor of the National Key Laboratory for Novel Software Technology, China. He is a member of the LAMDA group. His research interests are mainly in machine learning. Particularly, he is interested in semi-supervised learning, statistical learning and optimization. He has published over 30 papers in top-tier journal and conferences such as JMLR, TPAMI, AIJ, ICML, NIPS, AAAI, etc. He is the senior program committee member of top-tier AI conferences such as IJCAI15, IJCAI17, AAAI19, and an editorial board member of machine learning journal special issues. He has received outstanding doctoral dissertation award from China Computer Federation (CCF), outstanding doctoral dissertation award from Jiangsu Province and Microsoft Fellowship Award.



De-Ming Liang received the BSc degree in 2017. He is currently a master student in the Department of Computer Science and Technology at Nanjing University, China. His research interests are mainly in machine learning. Particularly, he is interested in weakly supervised learning.