Counterfactual Propagation for Semi-supervised Individual Treatment Effect Estimation

Shonosuke Harada[1(✉)] and Hisashi Kashima[1,2]

[1] Kyoto University, Kyoto, Japan
sh1108@ml.ist.i.kyoto-u.ac.jp, kashima@i.kyoto-u.ac.jp
[2] RIKEN AIP, Tokyo, Japan

Abstract. Individual treatment effect (ITE) represents the expected improvement in the outcome of taking a particular action to a particular target, and plays important roles in decision making in various domains. However, its estimation problem is difficult because intervention studies to collect information regarding the applied treatments (i.e., actions) and their outcomes are often quite expensive in terms of time and monetary costs. In this study, we consider a semi-supervised ITE estimation problem that exploits more easily-available unlabeled instances to improve the performance of ITE estimation using small labeled data. We combine two ideas from causal inference and semi-supervised learning, namely, matching and label propagation, respectively, to propose *counterfactual propagation,* which is the first semi-supervised ITE estimation method. Experiments using semi-real datasets demonstrate that the proposed method can successfully mitigate the data scarcity problem in ITE estimation.

Keywords: Causal inference · Treatment effect estimation · Semi-supervised learning

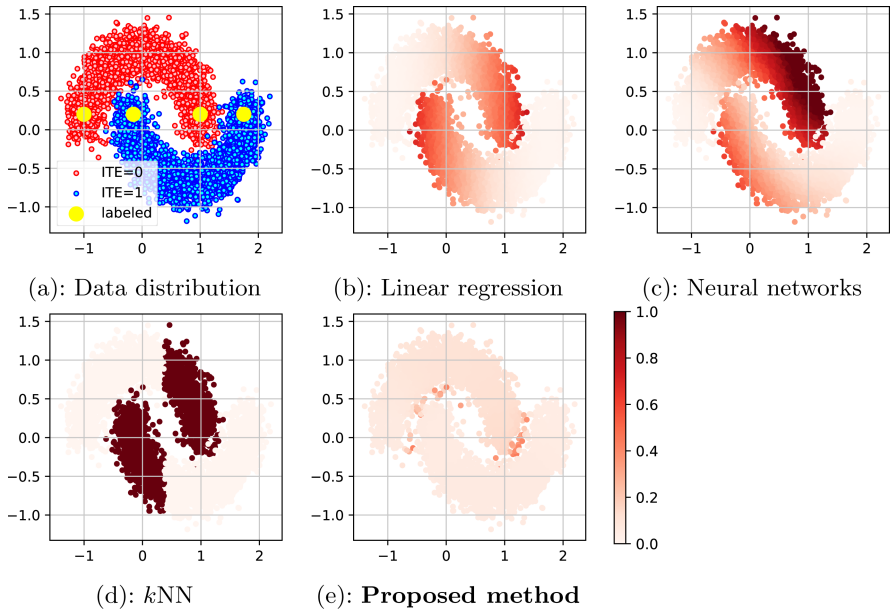
1 Introduction

One of the important roles of predictive modeling is to support decision making related to taking particular actions in responses to situations. The recent advances of in the machine learning technologies have significantly improved their predictive performance. However, most predictive models are based on passive observations and do not aim to predict the causal effects of actions that actively intervene in environments. For example, advertisement companies are interested not only in their customers' behavior when an advertisement is presented, but also in the causal effect of the advertisement, in other words, the change it causes on their behavior. There has been a growing interest in moving from this passive predictive modeling to more active causal modeling in various domains, such as education [14], advertisement [8,22], economic policy [20], and health care [12].

Taking an action toward a situation generally depends on the expected improvement in the outcome due to the action.

This is often called the *individual treatment effect (ITE)* [32] and is defined as the difference between the outcome of taking the action and that of *not* taking the action. An intrinsic difficulty in ITE estimation is that ITE is defined

© Springer Nature Switzerland AG 2021
F. Hutter et al. (Eds.): ECML PKDD 2020, LNAI 12457, pp. 542–558, 2021.
https://doi.org/10.1007/978-3-030-67658-2_31



(a): Data distribution (b): Linear regression (c): Neural networks

(d): kNN (e): **Proposed method**

Fig. 1. Illustrative example using (a) two-moon dataset. Each moons has a constant ITE either of 0 and 1. Only two labeled instances are available for each moon, denoted by yellow points, whose observed (treatment, outcome) pairs are $(0,1),(1,1),(1,1),(0,0)$ from left to right. Figures (b), (c), and (d) show the ITE estimation error (PEHE) by the standard two-model approach using different base models suffered from the lack of labeled data. The deeper-depth color indicates larger errors. The proposed semi-supervised method (e) successfully exploits the unlabeled data to estimate the correct ITEs. (Color figure online)

as the difference between the factual and counterfactual outcomes [21,26,32]; in other words, the outcome that we can actually observe is either of the one when we take an action or the one when we do not, and it is physically impossible to observe both. To address the counterfactual predictive modeling from observational data, various techniques including matching [31], inverse-propensity weighting [29], instrumental variable methods [3], and more modern deep learning-based approaches have been developed [17,33]. For example, in the matching method, matching pairs of instances with similar covariate values and different treatment assignments are determined. The key idea is to consider the two instances in a matching pair as the counterfactual instance of each other so that we can estimate the ITE by comparing the pair.

Another difficulty in ITE estimation is data scarcity. For ITE estimation, we need some labeled instances whose treatments (i.e., whether or not an action was taken on the instance) and their outcomes (depending on the treatments) as well as their covariates are given. However, collecting such labeled instances can be quite costly in terms of time and money, or owing to other reasons, such

as physical and ethical constraints [27,28]. Consequently, ITE estimation from scarcely labeled data is an essential requirement in many situations.

In the ordinary predictive modeling problem, a promising option to the scarcity of labeled data is semi-supervised learning that exploits unlabeled instances only with covariates because it is relatively easy to obtain such unlabeled data. A typical solution is the graph-based label propagation method [4,38,42], which makes predictions for unlabeled instances based on the assumption that instances with similar covariate values are likely to have a same label.

In this study, we consider a semi-supervised ITE estimation problem. The proposed solution called *counterfactual propagation* is based on the resemblance between the matching method in causal inference and the graph-based semi-supervised learning method called label propagation. We consider a weighted graph over both labeled instances with treatment outcomes and unlabeled instances with no outcomes, and estimate ITEs using the smoothness assumption of the outcomes and the ITEs.

The proposed idea is illustrated in Fig. 1. Figure 1(a) describes the two-moon shaped data distribution. We consider a binary treatment and binary outcomes. The blue points indicate the instances with a positive ITE ($= 1$), where the outcome is 1 if the treatment is 1 and 0 if the treatment is 0. The red points indicate the instances with zero ITE ($= 0$); their outcomes are always 1 irrespective of the treatments. We have only four labeled data instances shown as yellow points, whose observed $(\text{treatment}, \text{outcome})$ pairs are $(0, 1), (1, 1), (1, 1), (0, 0)$ from left to right. Since the amount of labeled data is considerably limited, supervised methods relying only on labeled data fail to estimate the ITEs. Figures 1(b), (c), (d) show the ITE estimation errors by the standard two-model approach using different base learners, which show poor performance. In contrast, the proposed approach exploits unlabeled data to find connections between the red points and those between the blue points to estimate the correct ITEs (Fig. 1(e)).

We propose an efficient learning algorithm assuming the use of a neural network as the base model, and conduct experiments using semi-synthetic real-world datasets to demonstrate that the proposed method estimates the ITEs more accurately than baselines when the labeled instances are limited.

2 Semi-supervised ITE Estimation Problem

We start with the problem setting of the semi-supervised treatment effect estimation problem. Suppose we have N labeled instances and M unlabeled instances. (We usually assume $N \ll M$.) The set of labeled instances is denoted by $\{(\mathbf{x}_i, t_i, y_i^{t_i})\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^D$ is the covariates of the i-th instance, $t_i \in \{0, 1\}$ is the treatment applied to instance i, and $y_i^{t_i}$ is its outcome. Note that for each instance i, either $t_i = 0$ or $t_i = 1$ is realized; accordingly, either y_i^0 or y_i^1 is available. The unobserved outcome is called a counterfactual outcome. The set of unlabeled instances is denoted by $\{(\mathbf{x}_i)\}_{i=N+1}^{N+M}$, where only the covariates are available.

Our goal is to estimate the ITE for each instance. Following the Rubin-Neyman potential outcomes framework [32,34], the ITE for instance i is defined as $\tau_i = y_i^1 - y_i^0$ exploiting both the labeled and unlabeled sets. Note that τ_i is not known even for the labeled instances, and we want to estimate the ITEs for both the labeled and unlabeled instances.

We make typical assumptions in ITE estimation in this study. i.e., (i) stable unit treatment value: the outcome of each instance is not affected by the treatment assigned to other instances; (ii) unconfoundedness: the treatment assignment to an instance is independent of the outcome given covariates (confounder variables); (iii) overlap: each instance has a positive probability of treatment assignment.

3 Proposed Method

We propose a novel ITE estimation method that utilizes both the labeled and unlabeled instances. The proposed solution called *counterfactual propagation* is based on the resemblance between the matching method in causal inference and the graph-based semi-supervised learning method.

3.1 Matching

Matching is a popular solution to address the counterfactual outcome problem. Its key idea is to consider two similar instances as the counterfactual instance of each other so that we can estimate the causal effect by comparing the pair. More concretely, we define the similarity w_{ij} between two instances i and j, as that defined between their covariates; for example, we can use the Gaussian kernel.

$$w_{ij} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^2}\right). \tag{1}$$

The set of (i, j) pairs with w_{ij} being larger than a threshold and satisfying $t_i \neq t_j$ are found and compared as counterfactual pairs. Note that owing to definition of the matching pair, the matching method only uses labeled data.

3.2 Graph-Based Semi-supervised Learning

Graph-based semi-supervised learning methods assume that the nearby instances in a graph are likely to have similar outputs. For a labeled dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ and an unlabeled dataset $\{\mathbf{x}_i\}_{i=N+1}^{N+M}$, their loss functions for standard predictive modeling typically look like

$$L(f) = \sum_{i=1}^{N} l(y_i, f(\mathbf{x}_i)) + \lambda \sum_{i,j=1}^{N+M} w_{ij} \left(f(\mathbf{x}_i) - f(\mathbf{x}_j)\right)^2, \tag{2}$$

where f is a prediction model, l is a loss function for the labeled instances, and λ is a hyper-parameter. The second term imposes "smoothness" of the model

output over the input space characterized by w_{ij} that can be considered as the weighted adjacency matrix of a weighted graph; it can be seen the same as that used for matching (1).

The early examples of graph-based methods include label propagation [42] and manifold regularization [4]. More recently, deep neural networks have been used as the base model f [38].

3.3 Treatment Effect Estimation Using Neural Networks

We build our ITE estimation model based on the recent advances of deep-learning approaches for ITE estimation, specifically, the treatment-agnostic representation network (TARNet) [33] that is a simple but quite effective model. TARNet shares common parameters for both treatment instances and control instances to construct representations but employs different parameters in its prediction layer, which is given as:

$$f(\mathbf{x}_i, t_i) = \begin{cases} \Theta_1^\top g\left(\Theta^\top x_i\right) & (t_i = 1) \\ \Theta_0^\top g\left(\Theta^\top x_i\right) & (t_i = 0) \end{cases}, \tag{3}$$

where Θ is the parameters in the representation learning layer and Θ_1, Θ_0 are those in the prediction layers for treatment and controlled instances, respectively. The g is a non-linear function such as ReLU. One of the advantages of TARNet is that joint representations learning and separate prediction functions for both treatments enable more flexible modeling.

3.4 Counterfactual Propagation

It is evident that the matching method relies only on labeled data, while the graph-based semi-supervised learning method does not address ITE estimation; however, they are quite similar because they both use instance similarity to interpolate the factual/counterfactual outcomes or model predictions as mentioned in Sect. 3.2. Our idea is to combine the two methods to propagate the outcomes and ITEs over the matching graph assuming that similar instances would have similar outcomes.

Our objective function consists of three terms, $L_\mathrm{s}, L_\mathrm{o}, L_\mathrm{e}$, given as

$$L(f) = L_\mathrm{s}(f) + \lambda_\mathrm{o} L_\mathrm{o}(f) + \lambda_\mathrm{e} L_\mathrm{e}(f), \tag{4}$$

where λ_o and λ_e are the regularization hyper-parameters. We employ TARNet [33] as the outcome prediction model $f(\mathbf{x}, t)$. The first term in the objective function (4) is a standard loss function for supervised outcome estimation; we specifically employ the squared loss function as

$$L_s(f) = \sum_{i=1}^{N} (y_i^{t_i} - f(\mathbf{x}_i, t_i))^2. \tag{5}$$

Note that it relies only on the observed outcomes of the treatments that are observed in the data denoted by t_i.

The second term L_{o} is the outcome propagation term:

$$L_{\mathrm{o}}(f) = \sum_{t} \sum_{i,j=1}^{N+M} w_{ij}((f(\mathbf{x}_i, t) - f(\mathbf{x}_j, t))^2. \tag{6}$$

Similar to the regularization term (2) in the graph-based semi-supervised learning, this term encourages the model to output similar outcomes for similar instances by penalizing the difference between their outcomes. This regularization term allows the model to propagate outcomes over a matching graph. If two nearby instances have different treatments, they interpolate the counterfactual outcome of each other, which compares the factual and (interpolated) counterfactual outcomes to estimate the ITE. The key assumption behind this term is the smoothness of outcomes for each treatment over the covariate space. While w_{ij} indicates the adjacency between nodes i and j in the graph-based regularization, it can be considered as a matching between the instances i and j in the treatment effect estimation problem. Even though traditional matching methods have only rely on labeled instances, we combine matching with graph-based regularization which also utilizes unlabeled instances. This regularization enables us to propagate the outcomes for each treatment over the matching graph and mitigate the counterfactual problem.

The third term L_{e} is the ITE propagation term defined as

$$L_{\mathrm{e}}(f) = \sum_{i,j=1}^{N+M} w_{ij}(\hat{\tau}_i - \hat{\tau}_j)^2, \tag{7}$$

where $\hat{\tau}_i$ is the ITE estimate for instance i:

$$\hat{\tau}_i = f(\mathbf{x}_i, 1) - f(\mathbf{x}_i, 0).$$

$L_{\mathrm{e}}(f)$ imposes the smoothness of the ITE values in addition to that of the outcomes imposed by outcome propagation (6). In comparison to the standard supervised learning problems, where the goal is to predict the outcomes, as stated in Sect. 2, our objective is to predict the ITEs. This term encourages the model to output similar ITEs for similar instances. We expect that the outcome propagation and ITE propagation terms are beneficial especially when the available labeled instances are limited while there is an abundance of unlabeled instances, similar to semi-supervised learning.

3.5 Estimation Algorithm

As mentioned earlier, we assume the use of neural networks as the specific choice of the outcome prediction model f based on the recent successes of deep neural networks in causal inference. For computational efficiency, we apply a sampling approach to optimizing Eq. (4). Following the existing method [38], we employ

Algorithm 1: Counterfactual propagation

Input: labeled instances $\{(\mathbf{x}_i, t_i, y_i^{t_i})\}_{i=1}^{N}$, unlabeled instances $\{(\mathbf{x}_i)\}_{i=N+1}^{N+M}$, a similarity matrix $w = (w_{ij})$, and mini-batch sizes b_1, b_2.
Output: estimated outcome(s) for each treatment $\hat{y}_i^1$ and/or $\hat{y}_i^0$ using Eq. (3).
while *not converged* **do**
  # Approximating the supervised loss
  Sample b_1 instances $\{(\mathbf{x}_i, t_i, y_i)\}$ from the labeled instances
  Compute the supervised loss (5) for the b_1 instances
  # Approximating propagation terms
  Sample b_2 pairs of instances $\{(\mathbf{x}_i, \mathbf{x}_j)\}$
  Compute the outcome propagation terms $\lambda_{\mathrm{o}} L_{\mathrm{o}}$ for b_2 pairs of instances
  Sample b_2 pairs of instances $\{(\mathbf{x}_i, \mathbf{x}_j)\}$
  Compute the ITE propagation terms $\lambda_{\mathrm{e}} L_{\mathrm{e}}$ of b_2 pairs of instances
  Update the parameters to minimize $L_s + \lambda_o L_o + \lambda_e L_e$ for the sampled instances
end

the Adam optimizer [18], which is based on stochastic gradient descent to train the model in a mini-batch manner.

Algorithm 1 describes the procedure of model training, which iterates two steps until convergence. In the first step, we sample a mini-batch consisting of b_1 labeled instances to approximate the supervised loss (5). In the second step, we compute the outcome propagation term and the ITE propagation terms using a mini-batch consisting of b_2 instance pairs. Note that in order to make the model more flexible, we can employ different regularization parameters for the treatment outcomes and the control outcomes. The b_1 and b_2 are considered as hyper-parameters; the details are described in Sect. 4. In practice, we optimize only the supervised loss for the first several epochs, and decrease the strength of regularization as training proceeds, in order to guide efficient training [38].

4 Experiments

We test the effectiveness of the proposed semi-supervised ITE estimation method in comparison with various supervised methods, especially when the available labeled data are strictly limited. We first conduct experiments using two semi-synthetic datasets based on public real datasets. We also design some experiments varying the magnitude of noise on outcomes to explore how the noisy outcomes affect the proposed method. Our implementation is available on Github[1].

[1] https://github.com/SH1108/CounterfactualPropagation.

4.1 Datasets

Owing to the counterfactual nature of ITE estimation, we rarely access real-world datasets including ground truth ITEs, and therefore cannot directly evaluate ITE estimation methods like the standard supervised learning methods using cross-validation. Therefore, following the existing work [17], we employ two semi-synthetics datasets whose counterfactual outcomes are generated through simulations. Refer to the original papers for the details on outcome generations [14,17].

News Dataset is a dataset including opinions of media consumers for news articles [17]. It contains 5,000 news articles and outcomes generated from the NY Times corpus[2]. Each article is consumed on desktop ($t = 0$) or mobile ($t = 1$) and it is assumed that media consumers prefer to read some articles on mobile than desktop. Each article is generated by a topic model and represented in the bag-of-words representation. The size of the vocabulary is 3,477.

IHDP Dataset is a dataset created by randomized experiments called the Infant Health and Development Program (IHDP) [14] to examine the effect of special child care on future test scores. It contains the results of 747 subjects (139 treated subjects and 608 control subjects) with 25 covariates related to infants and their mothers. Following the existing studies [17,33], the ground-truth counterfactual outcomes are simulated using the NPCI package [10].

4.2 Experimental Settings

Since we are particularly interested in the situation when the available labeled data are strictly limited, we split the data into a training dataset, validation dataset, and a test dataset by limiting the size of the training data. We change the ratio of the training to investigate the performance; we use $10\%, 5\%$, and 1% of the whole data from the News dataset, and use $40\%, 20\%$, and 10% of those from the IHDP dataset for the training datasets. The rest 80% and 10% of the whole News data are used for test and validation, respectively. Similarly, 50% and 10% of the whole IHDP dataset are used for test and validation, respectively. We report the average results of 10 trials on the News dataset and 50 trials on the IHDP dataset.

In addition to the evaluation under labeled data scarcity, we also test the robustness against label noises. As pointed out in previous studies, noisy labels in training data can severely deteriorate predictive performance, especially in semi-supervised learning. Following the previous work [14,17], we add the noise $\epsilon \sim \mathcal{N}(0, c^2)$ to the observed outcomes in the training data, where $c \in \{1, 3, 5, 7, 9\}$. In this evaluation, we use 1% of the whole data as the training data for the News dataset and 10% for the IHDP dataset, respectively, since we are mainly interested in label-scarce situations.

[2] https://archive.ics.uci.edu/ml/datasets/Bag+of+Words.

The hyper-parameters are tuned based on the prediction loss using the observed outcomes on the validation data. We calculate the similarities between the instances by using the Gaussian kernel; we select σ^2 from $\{5 \times 10^{-3}, 1 \times 10^{-3}, \ldots, 1 \times 10^2, 5 \times 10^2\}$, and select λ_o and λ_e from $\{1 \times 10^{-3}, 1 \times 10^{-2}, \ldots, 1 \times 10^2\}$. Because the scales of treatment outcomes and control outcomes are not always the same, we found scaling the regularization terms according to them is beneficial; specifically, we scale the regularization terms with respect to the treatment outcomes, the control outcomes, and the treatment effects by $\alpha = 1/\sigma_{y^1}^2, \beta = 1/\sigma_{y^0}^2$, and $\gamma = 1/(\sigma_{y^1}^2 + \sigma_{y^0}^2)$, respectively. We apply principal component analysis to reduce the input dimensions before applying the Gaussian kernel; we select the number of dimensions from $\{2, 4, 6, 8, 16, 32, 64\}$. The learning rate is set to 1×10^{-3} and the mini-batch sizes b_1, b_2 are chosen from $\{4, 8, 16, 32\}$.

As the evaluation metrics, we report the *Precision in Estimation of Heterogeneous Effect (PEHE)* used in the previous research [14]. PEHE is the estimation error of individual treatment effects, and is defined as

$$\epsilon_{\text{PEHE}} = \frac{1}{N+M} \sum_{i=1}^{N+M} (\tau_i - \hat{\tau}_i)^2.$$

Following the previous studies [33,40], we evaluate the predictive performance for labeled instances and unlabeled instances separately. Note that, although we observe the factual outcomes of the labeled data, their true ITEs are still unknown because we cannot observe their counterfactual outcomes.

4.3 Baselines

We compare the proposed method with several existing supervised ITE estimation approaches. (i) Linear regression (Ridge, Lasso) is the ordinary linear regression models with ridge regularization or lasso regularization. We consider two variants: one that includes the treatment as a feature (denoted by 'Ridge-1' and 'Lasso-1'), and the other with two separated models for treatment and control (denoted by 'Ridge-2' and 'Lasso-2'). (ii) k-nearest neighbors (kNN) is a matching-based method that predicts the outcomes using nearby instances. (iii) Propensity score matching with logistic regression (PSM) [29] is a matching-based method using the propensity score estimated by a logistic regression model. We also compared the proposed method with tree models such as (iv) random forest (RF) [6] and its causal extension called (v) causal forest (CF) [37]. In CF, trees are trained to predict propensity score and leaves are used to predict treatment effects. (vi) TARNet [33] is a deep neural network model that has shared layers for representation learning and different layers for outcome prediction for treatment and control instances. (vii) Counterfactual regression (CFR) [33] is a state-of-the-art deep neural network model based on balanced representations between treatment and control instances. We use the Wasserstein distance.

Table 1. The performance comparison of different methods on News dataset. The † indicates that our proposed method (CP) performs statistically significantly better than the baselines by the paired t-test ($p < 0.05$). The bold results indicate the best results in terms of the average.

$\sqrt{\epsilon_{\text{PEHE}}}$	News 1%		News 5%		News 10%	
Method	Labeled	Unlabeled	Labeled	Unlabeled	Labeled	Unlabeled
Ridge-1	$^{\dagger}4.494_{\pm 1.116}$	$^{\dagger}4.304_{\pm 0.988}$	$^{\dagger}4.666_{\pm 1.0578}$	$^{\dagger}3.951_{\pm 0.954}$	$^{\dagger}4.464_{\pm 1.082}$	$^{\dagger}3.607_{\pm 0.943}$
Ridge-2	$2.914_{\pm 0.797}$	$^{\dagger}2.969_{\pm 0.814}$	$^{\dagger}2.519_{\pm 0.586}$	$^{\dagger}2.664_{\pm 0.614}$	$^{\dagger}2.560_{\pm 0.558}$	$^{\dagger}2.862_{\pm 0.621}$
Lasso-1	$^{\dagger}4.464_{\pm 1.082}$	$^{\dagger}3.607_{\pm 0.943}$	$^{\dagger}4.466_{\pm 1.058}$	$^{\dagger}3.367_{\pm 0.985}$	$^{\dagger}4.464_{\pm 1.0822}$	$^{\dagger}3.330_{\pm 0.984}$
Lasso-2	$^{\dagger}3.344_{\pm 1.022}$	$^{\dagger}3.476_{\pm 1.038}$	$^{\dagger}2.568_{\pm 0.714}$	$^{\dagger}2.848_{\pm 0.751}$	$^{\dagger}2.269_{\pm 0.628}$	$^{\dagger}2.616_{\pm 0.663}$
kNN	$^{\dagger}3.678_{\pm 1.250}$	$^{\dagger}3.677_{\pm 1.246}$	$^{\dagger}3.351_{\pm 1.004}$	$^{\dagger}3.434_{\pm 1.018}$	$^{\dagger}3.130_{\pm 0.752}$	$^{\dagger}3.294_{\pm 0.766}$
PSM	$^{\dagger}3.713_{\pm 1.149}$	$^{\dagger}3.662_{\pm 1.127}$	$^{\dagger}3.363_{\pm 0.901}$	$^{\dagger}3.500_{\pm 0.961}$	$^{\dagger}3.260_{\pm 0.734}$	$^{\dagger}3.526_{\pm 0.832}$
RF	$^{\dagger}4.494_{\pm 1.116}$	$^{\dagger}3.691_{\pm 0.878}$	$^{\dagger}4.466_{\pm 1.058}$	$^{\dagger}2.975_{\pm 0.874}$	$^{\dagger}4.464_{\pm 1.082}$	$^{\dagger}2.657_{\pm 0.682}$
CF	$^{\dagger}3.691_{\pm 1.082}$	$^{\dagger}3.607_{\pm 0.943}$	$^{\dagger}3.196_{\pm 0.901}$	$^{\dagger}3.215_{\pm 0.910}$	$^{\dagger}3.101_{\pm 0.806}$	$^{\dagger}3.129_{\pm 0.818}$
TARNET	$^{\dagger}3.166_{\pm 0.742}$	$^{\ddagger}3.160_{\pm 0.722}$	$^{\dagger}2.670_{\pm 0.796}$	$^{\dagger}2.666_{\pm 0.773}$	$^{\dagger}2.589_{\pm 0.894}$	$^{\dagger}2.598_{\pm 0.869}$
CFR	$2.908_{\pm 0.752}$	$2.925_{\pm 0.746}$	$^{\dagger}2.590_{\pm 0.772}$	$^{\dagger}2.546_{\pm 0.796}$	$^{\dagger}2.570_{\pm 0.519}$	$^{\dagger}2.451_{\pm 0.547}$
CP (proposed)	$\mathbf{2.844}_{\pm \mathbf{0.683}}$	$\mathbf{2.823}_{\pm \mathbf{0.656}}$	$\mathbf{2.310}_{\pm \mathbf{0.430}}$	$\mathbf{2.446}_{\pm \mathbf{0.471}}$	$\mathbf{2.003}_{\pm \mathbf{0.393}}$	$\mathbf{2.153}_{\pm \mathbf{0.436}}$

4.4 Results and Discussions

We discuss the performance of the proposed method compared with the baselines by changing the size of labeled datasets, and then investigate the robustness against the label noises.

We first see the experimental results for different sizes of labeled datasets and sensitivity to the choice of the hyper-parameters that control the strength of label propagation. Tables 1 and 2 show the PEHE values by different methods for the News dataset and the IHDP dataset, respectively. Overall, our proposed method exhibits the best ITE estimation performance for both labeled and unlabeled data in both of the datasets; the advantage is more significant in the News dataset. The News dataset is a relatively high-dimensional dataset represented using a bag of words. The two-model methods such as Ridge-2 and Lasso-2 perform well in spite of their simplicity, and in terms of regularization types, the Lasso-based methods perform relatively better due to the high-dimensional nature of the dataset.

The proposed method also performs the best in the IHDP dataset; however, the performance gain is rather moderate, as shown by the no statistical significance against CFR [33] with the largest 40%-labeled data, which is the most powerful baseline method. The reason for the moderate improvements is probably because of the difficulty in defining appropriate similarities among instances, because the IHDP dataset has various types of features including continuous variables and discrete variables. The traditional baselines such as Ridge-1, Lasso-1, k-NN matching, and the tree-based models show limited performance; in contrast, the deep learning based methods such as TARNet and CFR demonstrate remarkable performance. Generally, the performance gain by the proposed method is larger on labeled data than on unlabeled data.

Our proposed method has two different propagation terms, the outcome propagation term and the ITE propagation term, as regularizers for semi-supervised

Table 2. The performance comparison of different methods on IHDP dataset. The † indicates that our proposed method (CP) performs statistically significantly better than the baselines by the paired t-test ($p < 0.05$). The bold results indicate the best results in terms of average.

$\sqrt{\epsilon_{\text{PEHE}}}$	IHDP 10%		IHDP 20%		IHDP 40%	
Method	Labeled	Unlabeled	Labeled	Unlabeled	Labeled	Unlabeled
Ridge-1	$^{\dagger}5.484_{\pm 8.825}$	$^{\dagger}5.696_{\pm 7.328}$	$^{\dagger}5.067_{\pm 8.337}$	$^{\dagger}4.692_{\pm 6.943}$	$^{\dagger}4.80_{\pm 8.022}$	$^{\dagger}4.448_{\pm 6.874}$
Ridge-2	$^{\dagger}3.426_{\pm 5.692}$	$^{\dagger}3.357_{\pm 5.177}$	$^{\dagger}2.918_{\pm 4.874}$	$^{\dagger}2.918_{\pm 4.730}$	$^{\dagger}2.605_{\pm 4.314}$	$^{\dagger}2.639_{\pm 4.496}$
Lasso-1	$^{\dagger}6.685_{\pm 10.655}$	$^{\dagger}6.408_{\pm 9.900}$	$^{\dagger}6.435_{\pm 10.147}$	$^{\dagger}6.2446_{\pm 9.639}$	$^{\dagger}6.338_{\pm 9.704}$	$^{\dagger}6.223_{\pm 9.596}$
Lasso-2	$^{\dagger}3.118_{\pm 5.204}$	$^{\ddagger}3.292_{\pm 5.725}$	$^{\dagger}2.684_{\pm 4.428}$	$^{\dagger}2.789_{\pm 4.731}$	$^{\dagger}2.512_{\pm 4.075}$	$^{\dagger}2.571_{\pm 4.379}$
kNN	$^{\dagger}4.457_{\pm 6.957}$	$^{\dagger}4.603_{\pm 6.629}$	$^{\dagger}4.023_{\pm 6.193}$	$^{\dagger}4.370_{\pm 6.244}$	$^{\dagger}3.623_{\pm 5.316}$	$^{\dagger}4.109_{\pm 5.936}$
PSM	$^{\dagger}6.506_{\pm 10.077}$	$^{\dagger}6.982_{\pm 10.672}$	$^{\dagger}6.277_{\pm 9.708}$	$^{\dagger}7.209_{\pm 11.077}$	$^{\dagger}6.065_{\pm 9.362}$	$^{\dagger}7.181_{\pm 9.362}$
RF	$^{\dagger}6.924_{\pm 10.620}$	$^{\dagger}5.356_{\pm 8.790}$	$^{\dagger}6.854_{\pm 10.471}$	$^{\dagger}4.845_{\pm 8.241}$	$^{\dagger}6.928_{\pm 10.396}$	$^{\dagger}4.549_{\pm 7.822}$
CF	$^{\dagger}5.389_{\pm 8.736}$	$^{\dagger}5.255_{\pm 8.070}$	$^{\dagger}4.939_{\pm 7.762}$	$^{\dagger}4.955_{\pm 7.503}$	$^{\dagger}4.611_{\pm 7.149}$	$^{\dagger}4.764_{\pm 7.448}$
TARNET	$^{\dagger}3.827_{\pm 5.315}$	$^{\dagger}3.664_{\pm 4.888}$	$^{\dagger}2.770_{\pm 3.617}$	$^{\dagger}2.770_{\pm 3.542}$	$^{\dagger}2.005_{\pm 2.447}$	$^{\dagger}2.267_{\pm 2.825}$
CFR	$^{\dagger}3.461_{\pm 5.1444}$	$^{\dagger}3.292_{\pm 4.619}$	$^{\dagger}2.381_{\pm 3.126}$	$^{\dagger}2.403_{\pm 3.080}$	$1.572_{\pm 1.937}$	$1.815_{\pm 2.204}$
CP (proposed)	$\mathbf{2.427}_{\pm \mathbf{3.189}}$	$\mathbf{2.652}_{\pm \mathbf{3.469}}$	$\mathbf{1.686}_{\pm \mathbf{1.838}}$	$\mathbf{1.961}_{\pm \mathbf{2.343}}$	$\mathbf{1.299}_{\pm \mathbf{1.001}}$	$\mathbf{1.485}_{\pm \mathbf{1.433}}$

learning. Table 3 investigates the contributions by the different propagation terms. The proposed method using the both propagation terms (denoted by CP) shows better results than the one only with the ITE propagation denoted by CP ($\lambda_o = 0$); on the other hand, the improvement over the one only with the outcome regularization is marginal. This observation implies the outcome propagation contributes more to the predictive performance than the ITE propagation.

We also examine the sensitivity of the performance to the regularization hyper-parameters. Figure 2 reports the results using 40% and 10% of the whole data as the training data of the News and IHDP datasets, respectively. The proposed method seems rather sensitive to the strength of the regularization terms, particularly on the IHDP dataset, which suggests that the regularization parameters should be carefully tuned using validation datasets in the proposed method. In our experimental observations, slight changes in the hyper-parameters sometimes caused significant changes of predictive performance. We admit the hyper-parameter sensitivity is one of the current limitations in the proposed method and efficient tuning of the hyper-parameters should be addressed in future.

Finally, we compare the proposed method with the state-of-the-art methods by varying the magnitude of noises added to the outcomes. Fig 3 shows the performance comparison in terms of $\sqrt{\epsilon_{\text{PEHE}}}$. Note that the results when $c = 1$ correspond to the previous results in Tables 1 and 2 . The proposed method stays tolerant of relatively small magnitude of noises; however, with larger label noises, it suffers more from wrongly propagated outcome information than the baselines. This is consistent with the previous studies reporting the vulnerability of semi-supervised learning methods against label noises [7,11,23,35].

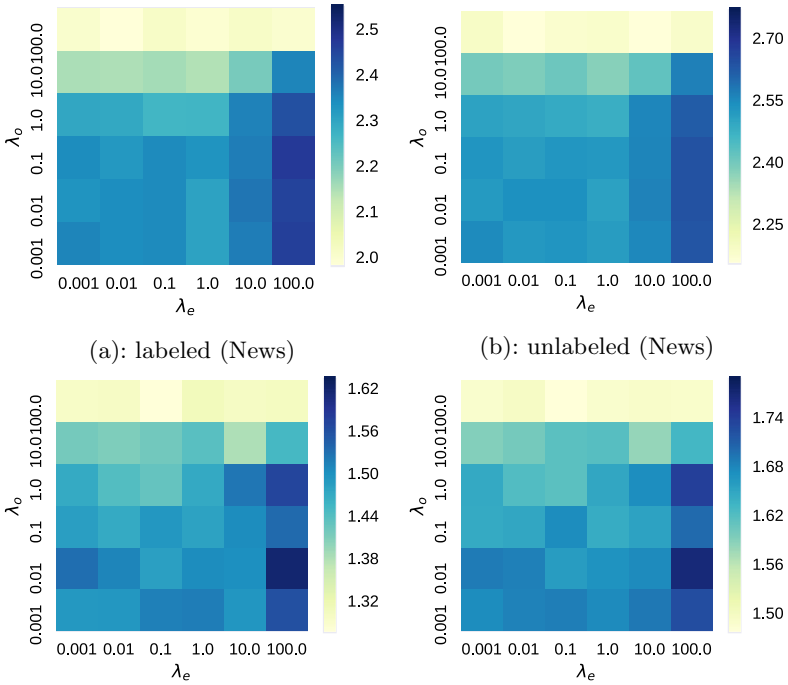
Table 3. Investigation of the contributions by the outcome propagation and the ITE propagation in the proposed method. The upper table shows the results for the News dataset, and the lower for the IHDP dataset. The $\lambda_o = 0$ and $\lambda_e = 0$ indicate the proposed method (CP) with only the ITE propagation and the outcome propagation, respectively, The † indicates that our proposed method (CP) performs statistically significantly better than the baselines by the paired t-test ($p < 0.05$). The bold numbers indicate the best results in terms of the average.

$\sqrt{\epsilon_{\text{PEHE}}}$	News 1%		News 5%		News 10%	
Method	Labeled	Unlabeled	Labeled	Unlabeled	Labeled	Unlabeled
CP ($\lambda_o = 0$)	$\mathbf{2.812_{\pm 651}}$	$\mathbf{2.806_{\pm 0.598}}$	$^\dagger 2.527_{\pm 0.474}$	$^\dagger 2.531_{\pm 0.523}$	$^\dagger 2.400_{\pm 0.347}$	$^\dagger 2.410_{\pm 0.450}$
CP ($\lambda_e = 0$)	$2.879_{\pm 0667}$	$2.885_{\pm 0.609}$	$2.351_{\pm 0.450}$	$2.483_{\pm 0.481}$	$\mathbf{1.996_{\pm 0.338}}$	$2.221_{\pm 0.455}$
CP	$2.844_{\pm 0.683}$	$2.823_{\pm 0.656}$	$\mathbf{2.310_{\pm 0.430}}$	$\mathbf{2.446_{\pm 0.471}}$	$2.003_{\pm 0.393}$	$\mathbf{2.153_{\pm 0.436}}$
$\sqrt{\epsilon_{\text{PEHE}}}$	IHDP 10%		IHDP 20%		IHDP 40%	
Method	Labeled	Unlabeled	Labeled	Unlabeled	Labeled	Unlabeled
CP ($\lambda_o = 0$)	$^\dagger 2.883_{\pm 3.708}$	$^\dagger 3.004_{\pm 4.071}$	$^\dagger 1.972_{\pm 1.930}$	$^\dagger 2.144_{\pm 2.465}$	$1.574_{\pm 1.392}$	$1.674_{\pm 1.874}$
CP ($\lambda_e = 0$)	$2.494_{\pm 3.201}$	$2.698_{\pm 3.461}$	$1.728_{\pm 2.194}$	$1.977_{\pm 2.450}$	$1.344_{\pm 1.383}$	$1.585_{\pm 1.923}$
CP	$\mathbf{2.427_{\pm 3.189}}$	$\mathbf{2.652_{\pm 3.469}}$	$\mathbf{1.686_{\pm 1.838}}$	$\mathbf{1.961_{\pm 2.343}}$	$\mathbf{1.299_{\pm 1.001}}$	$\mathbf{1.485_{\pm 1.433}}$

5 Related Work

5.1 Treatment Effect Estimation

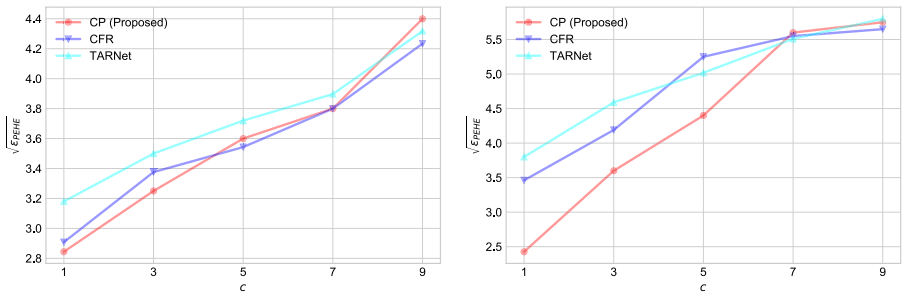
Treatment effect estimation has been one of the major interests in causal inference and widely studied in various domains. Matching [1,31] is one of the most basic and commonly used treatment effect estimation techniques. It estimates the counterfactual outcomes using its nearby instances, whose idea is similar to that of graph-based semi-supervised learning. Both methods assume that similar instances in terms of covariates have similar outcomes. To mitigate the curse of dimensionality and selection bias in matching, the propensity matching method relying on the one-dimensional propensity score was proposed [29,30]. The propensity score is the probability of an instance to get a treatment, which is modeled using probabilistic models like logistic regression, and has been successfully applied in various domains to estimate treatment effects unbiasedly [24]. Tree-based methods such as regression trees and random forests have also been well studied for this problem [9,37]. One of the advantages of such models is that they can build quite expressive and flexible models to estimate treatment effects. Recently, deep learning-based methods have been successfully applied to treatment effect estimation [17,33]. Balancing neural networks (BNNs) [17] aim to obtain balanced representations of a treatment groups and a control group by minimizing the discrepancy between them, such as the Wasserstein distance [33]. Most recently, some studies have addressed causal inference problems on network-structured data [2,13,36]. Alvari et al. applied the idea of manifold regularization using users activities as causality-based features to detect harmful users in social media [2]. Guo et al. considered treatment effect estimation on social networks using graph convolutional balancing neural networks [13]. In contrast with their work assuming the network structures are readily available, we do not assume them and considers matching network defined using covariates.



(a): labeled (News) (b): unlabeled (News)

(c): labeled (IHDP) (d): unlabeled (IHDP)

Fig. 2. Sensitivity of the results to the hyper-parameters. The colored bars indicate $\sqrt{\epsilon_{\text{PEHE}}}$ for (a)(b) the News dataset and (c)(d) the IHDP dataset when using the largest size of labeled data. The deeper-depth color indicates larger errors. It is observed that the proposed method is somewhat sensitive to the choice of hyper-parameters, especially, the strength of outcome regularizations (λ_o). (Color figure online)



(a): News (b): IHDP

Fig. 3. Performance comparisons for different levels of noise c added to the labels on (a) News dataset and (b) IHDP dataset. Note that the results when $c = 1$ correspond to the previous resuls (Tables 1 and 2).

5.2 Semi-supervised Learning

Semi-supervised learning, which exploits both labeled and unlabeled data, is one of the most popular approaches, especially in scenarios when only limited labeled data can be accessed [5,15]. Semi-supervised learning has many variants, and because it is almost impossible to refer to all of them, we mainly review the graph-based regularization methods, known as label propagation or manifold regularization [4,38,42]. Utilizing a given graph or a graph constructed based on instance proximity, graph-based regularization encourages the neighbor instances to have similar labels or outcomes [4,42]. Such idea is also applied to representation learning in deep neural networks [7,16,19,38,39]; they encourage nearby instances not only to have similar outcomes, but also have similar intermediate representations, which results in remarkable improvements from ordinary methods. One of the major drawbacks of semi-supervised approaches is that label noises in training data can be quite harmful; therefore, a number of studies managed to mitigate the performance degradation [11,23,25,35].

One of the most related work to our present study is graph-based semi-supervised prediction under sampling biases of labeled data [41]. The important difference between this work and ours is that they do not consider intervention and we do not consider the sampling biases of labeled data.

6 Conclusion

We addressed the semi-supervised ITE estimation problem. In comparison to the existing ITE estimation methods that only rely on labeled instances including treatment and outcome information, we proposed a novel semi-supervised ITE estimation method that also utilizes unlabeled instances. The proposed method called counterfactual propagation is built on two ideas from causal inference and semi-supervised learning, namely, matching and label propagation, respectively; accordingly, we devised an efficient learning algorithm. Experimental results using the semi-simulated real-world datasets revealed that our methods performed better in comparison to several strong baselines when the available labeled instances are limited. However, this method had issues related to reasonable similarity design and hyper-parameter tuning.

One of the possible future directions is to make use of balancing techniques such as the one used in CFR [33], which can be also naturally integrated into our model. Our future work also includes addressing the biased distribution of labeled instances. As mentioned in Related work, we did not consider such sampling biases for labeled data. Some debiasing techniques [41] might also be successfully integrated into our framework. In addition, robustness against noisy outcomes under semi-supervised learning framework is still the open problem and will be addressed in the future.

Acknowledgments. This work was partially supported by JSPS KAKENHI Grant Number 20H04244.

References

1. Abadie, A., Imbens, G.W.: Large sample properties of matching estimators for average treatment effects. Econometrica **74**(1), 235–267 (2006)
2. Alvari, H., Shaabani, E., Sarkar, S., Beigi, G., Shakarian, P.: Less is more: semi-supervised causal inference for detecting pathogenic users in social media. In: Proceedings of the 2019 World Wide Web Conference (WWW), pp. 154–161 (2019)
3. Baiocchi, M., Cheng, J., Small, D.S.: Instrumental variable methods for causal inference. Stat. Med. **33**(13), 2297–2340 (2014)
4. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. J. Machine Learn. Res. **7**, 2399–2434 (2006)
5. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 153–160 (2007)
6. Breiman, L.: Random forests. Machine Learn. **45**(1), 5–32 (2001)
7. Bui, T.D., Ravi, S., Ramavajjala, V.: Neural graph learning: training neural networks using graphs. In: Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM), pp. 64–71 (2018)
8. Chan, D., Ge, R., Gershony, O., Hesterberg, T., Lambert, D.: Evaluating online ad campaigns in a pipeline: causal models at scale. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 7–16 (2010)
9. Chipman, H.A., George, E.I., McCulloch, R.E., et al.: Bart: Bayesian additive regression trees. Ann. Appl. Stat. **4**(1), 266–298 (2010)
10. Dorie, V.: NPCI: Non-parametrics for causal inference. https://www.github.com/vdorie/npci (2016)
11. Du, B., Xinyao, T., Wang, Z., Zhang, L., Tao, D.: Robust graph-based semisupervised learning for noisy labeled data via maximum correntropy criterion. IEEE Trans. Cyber. **49**(4), 1440–1453 (2018)
12. Glass, T.A., Goodman, S.N., Hernán, M.A., Samet, J.M.: Causal inference in public health. Annual Rev. Public Health **34**, 61–75 (2013)
13. Guo, R., Li, J., Liu, H.: Learning individual causal effects from networked observational data. In: Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM), pp. 232–240 (2020)
14. Hill, J.L.: Bayesian nonparametric modeling for causal inference. J. Comput. Graph. Stat. **20**(1), 217–240 (2011)
15. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Comput. **18**(7), 1527–1554 (2006)
16. Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Label propagation for deep semi-supervised learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5070–5079 (2019)
17. Johansson, F., Shalit, U., Sontag, D.: Learning representations for counterfactual inference. In: Proceedings of the 33rd International Conference on Machine Learning (ICML), pp. 3020–3029 (2016)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)

20. LaLonde, R.J.: Evaluating the econometric evaluations of training programs with experimental data. The American Economic Review, pp. 604–620 (1986)
21. Lewis, D.: Causation. J. Philosophy **70**(17), 556–567 (1974)
22. Li, S., Vlassis, N., Kawale, J., Fu, Y.: Matching via dimensionality reduction for estimation of treatment effects in digital marketing campaigns. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI), pp. 3768–3774 (2016)
23. Liu, W., Wang, J., Chang, S.F.: Robust and scalable graph-based semisupervised learning. Proc. IEEE **100**(9), 2624–2638 (2012)
24. Lunceford, J.K., Davidian, M.: Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat. Med. **23**(19), 2937–2960 (2004)
25. Pal, A., Chakrabarti, D.: Label propagation with neural networks. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM), pp. 1671–1674 (2018)
26. Pearl, J.: Causality. Cambridge University Press (2009)
27. Pombo, N., Garcia, N., Bousson, K., Felizardo, V.: Machine learning approaches to automated medical decision support systems. In: Handbook of Research on Artificial Intelligence Techniques and Algorithms, pp. 183–203. IGI Global (2015)
28. Radlinski, F., Joachims, T.: Query chains: learning to rank from implicit feedback. In: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD), pp. 239–248. ACM (2005)
29. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983)
30. Rosenbaum, P.R., Rubin, D.B.: Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am. Stat. **39**(1), 33–38 (1985)
31. Rubin, D.B.: Matching to remove bias in observational studies. Biometrics, pp. 159–183 (1973)
32. Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. **66**(5), 688 (1974)
33. Shalit, U., Johansson, F.D., Sontag, D.: Estimating individual treatment effect: generalization bounds and algorithms. In: Proceedings of the 34th International Conference on Machine Learning (ICML), pp. 3076–3085. JMLR.org (2017)
34. Splawa-Neyman, J., Dabrowska, D.M., Speed, T.: On the application of probability theory to agricultural experiments. essay on principles. Section 9. Statistical Science, pp. 465–472 (1990)
35. Vahdat, A.: Toward robustness against label noise in training deep discriminative neural networks. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 5596–5605 (2017)
36. Veitch, V., Wang, Y., Blei, D.: Using embeddings to correct for unobserved confounding in networks. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 13769–13779 (2019)
37. Wager, S., Athey, S.: Estimation and inference of heterogeneous treatment effects using random forests. J. Am. Stat. Assoc. **113**(523), 1228–1242 (2018)
38. Weston, J., Ratle, F., Mobahi, H., Collobert, R.: Deep learning via semi-supervised embedding. In: Montavon, G., Orr, G.B., Müller, K.-R. (eds.) Neural Networks: Tricks of the Trade. LNCS, vol. 7700, pp. 639–655. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35289-8_34
39. Yang, Z., Cohen, W.W., Salakhutdinov, R.: Revisiting semi-supervised learning with graph embeddings. arXiv preprint arXiv:1603.08861 (2016)

40. Yao, L., Li, S., Li, Y., Huai, M., Gao, J., Zhang, A.: Representation learning for treatment effect estimation from observational data. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 2633–2643 (2018)
41. Zhou, F., Li, T., Zhou, H., Zhu, H., Jieping, Y.: Graph-based semi-supervised learning with non-ignorable non-response. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 7013–7023 (2019)
42. Zhu, X., Ghahramani, Z., Lafferty, J.D.: Semi-supervised learning using gaussian fields and harmonic functions. In: Proceedings of the 20th International conference on Machine learning (ICML), pp. 912–919 (2003)