SemiITE: Semi-supervised Individual Treatment Effect Estimation via Disagreement-Based Co-training

Qiang Huang[1,3,4,5], Jing Ma[2], Jundong Li[2], Huiyan Sun[1,3,5]✉, and Yi Chang[1,3,4,5]✉

[1] School of Artificial Intelligence, Jilin University
[2] University of Virginia
[3] International Center of Future Science, Jilin University
[4] Key Laboratory of Symbolic Computation and Knowledge Engineering, Jilin University
[5] Innovation Group of Marine Engineering Materials and Corrosion Control, Southern Marine Science and Engineering Guangdong Laboratory, Zhuhai
huangqiang18@mails.jlu.edu.cn,
{jm3mr,jundong}@virginia.edu,{yichang,huiyansun}@jlu.edu.cn

Abstract. Recent years have witnessed a surge of interests in Individual Treatment Effect (ITE) estimation, which aims to estimate the causal effect of a treatment (e.g., job training) on an outcome (e.g., employment status) for each individual (e.g., an employee). Various machine learning based methods have been proposed recently and have achieved satisfactory performance of ITE estimation from observational data. However, most of these methods overwhelmingly rely on a large amount of data with labeled treatment assignments and corresponding outcomes. Unfortunately, a significant amount of labeled observational data can be difficult to collect in real-world applications due to time and expense constraints. In this paper, we propose a Semi-supervised Individual Treatment Effect estimation (*SemiITE*) framework with a disagreement-based co-training style, which aims to utilize massive unlabeled data to better infer the factual and counterfactual outcomes of each instance with limited labeled data. Extensive experiments on two widely used real-world datasets validate the superiority of our *SemiITE* over the state-of-the-art ITE estimation models.

Keywords: Treatment Effect Estimation, Semi-supervised Learning

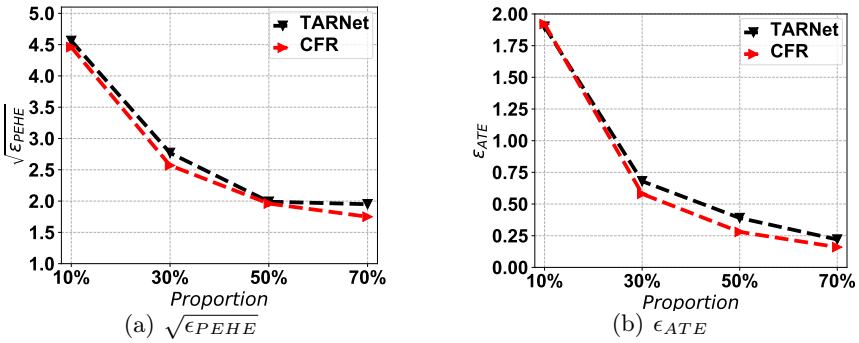
1 Introduction

Estimating individual treatment effect (ITE) is an important problem in causal inference, which aims to estimate the causal effect of a treatment on an outcome for each individual, e.g., "how would participating in a job training would influence the employment status of an employee?". ITE estimation plays an important role in a wide range of areas, such as decision making and policy evaluation

regarding healthcare [9,20], education [15], and economics [27]. A traditional solution for this problem is to conduct randomized controlled trials (RCTs), which randomly divide individuals into treatment group and control group with different treatment assignments (e.g., participating in the job training or not), and then estimate the causal effect of treatment assignment with the outcome difference over these two groups. However, performing RCTs could be costly, time-consuming, and even unethical [5,13]. To overcome these issues, different from RCTs, many machine learning based methods [6,18,26,30] have been proposed to estimate individual treatment effect directly from observational data and have achieved great success in recent years.

Despite the great success the aforementioned machine learning based models have achieved in causal effect estimation, most of them often require a large amount of *labeled observational data* (i.e., instances that come with treatment assignments and corresponding factual outcomes) in the training process. To show how the amount of *labeled observational data* affects the ITE estimation performance, we conduct an initial exploration by training two ITE estimation models CFR [26] and TARNet [17] on the IHDP dataset [4] with different proportions of labeled observational data. The ITE estimation prediction on the test data is shown in Figure 1. Here, we adopt the widely used metrics of $\sqrt{\epsilon_{PEHE}}$ and ϵ_{ATE} [13] (more details can be seen in Section 4) to evaluate the performance of the two models on ITE estimation. We can observe that the performance of these two models is poor when the proportion of training data is low, but improves significantly as the proportion increases to a large percentage. This example demonstrates the indispensability of the large amount of labeled observational data for existing state-of-the-art ITE estimation models. However, in many domains such as health care, the labeled observational data is often very scarce. The process of collecting such labeled observational data could take years and be extremely expensive, or it may face serious ethical issues [14]. Fortunately, *unlabeled observational data* (i.e., instances only with covariates) is easy to obtain, and many studies [8,34] have shown that unlabeled data is also beneficial to the performance of machine learning models. Exploiting unlabeled data for ITE estimation can greatly reduce the cost of collecting labeled observational data. Therefore, how to estimate ITE from limited labeled observational data by using unlabeled data is a pressing issue in causal inference.

To tackle the above problem, in this paper, we propose to use co-training framework [3] to harness the power of unlabeled observational data to aid ITE estimation. Co-training is a popular semi-supervised learning framework that has achieved great success in many problems. It first trains multiple diverse base learners on the limited labeled data, then the trained base learners are used to predict unlabeled data. At last, the most confident predictions of the base learners on the unlabeled data are iteratively added into the labeled data set. However, such co-training frameworks cannot be directly grafted into the ITE estimation problem, mainly because of the following difficulties. First, the existence of hidden confounders (i.e., the unobserved variables that influence both the treatment and the outcome) may result in confounding bias in ITE



(a) $\sqrt{\epsilon_{PEHE}}$ (b) ϵ_{ATE}

Fig. 1: The performance of CFR [26] and TARNet [17] for ITE estimation on the IHDP dataset with different proportions of labeled observational data (lower is better).

estimation, hence how to control the confounding bias is an issue to be addressed. Second, traditional co-training framework relies on multiple views of data (e.g., the acoustic attribute view and pictorial attribute view for a movie sample) to train multiple diverse base learners, otherwise the co-training degrades to self-training [21]. However, it is hard to collect such observational data with multiple views in causal inference, hence generating base causal models with diversity is critical for learning the individual treatment effect. Third, most of existing co-training frameworks [2,8,32] are mainly designed for classification problems. When it comes to ITE estimation, which is naturally a regression problem in most cases, base learners usually need to be re-trained to check whether the candidate instance prediction reduces their error rate in each instance selection round, which would increase the cost of computation and time. Thus designing an appropriate co-training strategy for ITE estimation to avoid re-training issue is also pressing.

To address the aforementioned difficulties, we propose a novel Semi-supervised Individual Treatment Effect estimation framework (*SemiITE*) via disagreement-based co-training. *SemiITE* builds a shared module to capture the hidden confounders so as to alleviate the confounding bias. To effectively enhance the ITE estimation performance under the semi-supervised setting, *SemiITE* generates three base potential outcome prediction models with diversity in a variety of ways. Moreover, to better utilize the unlabeled observational data, we design a novel co-training strategy in *SemiITE* based on the disagreement information of the three base models, which can select the most confident unlabeled instance predictions directly without re-training in each instance selection round. The following are the main contributions of our work:

– We formulate a novel research problem to utilize unlabeled observational data in a co-training manner for better individual treatment effect estimation.

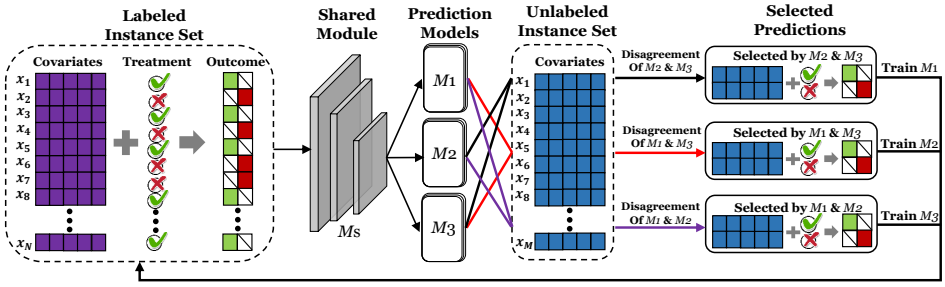


Fig. 2: An overview of the proposed semi-supervised ITE estimation framework *SemiITE* with disagreement-based co-training that utilizes unlabeled instances.

– We design a novel disagreement-based co-training framework *SemiITE* for semi-supervised individual treatment effect estimation, which can make use of unlabeled instances effectively, eliminate confounding bias, and avoid the re-training issue in each instance selection round.
– We perform extensive experiments and the results show that the proposed ITE estimation framework *SemiITE* is superior to existing state-of-the-art methods for ITE estimation when labeled observational data is limited.

2 Preliminaries

Notations. Let $L = \{(\boldsymbol{x}_1, t_1, y^{t_1}), ..., (\boldsymbol{x}_N, t_N, y^{t_N})\}$ denote the set of labeled observational instances with covariates, treatment assignments, and corresponding outcomes, where $\boldsymbol{x}_i \in \mathbb{R}^d$, $t_i \in \{0, 1\}$, y^{t_i} represent the covariates, treatment assignment, observed factual outcome given treatment assignment t_i of instance i, respectively; and N is the number of labeled instances. Let $U = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_M\}$ denote the set of unlabeled instances only with covariates, where M is the number of unlabeled instances.

Problem Statement. We develop our framework based on the potential outcome framework [24,25], which is widely used in causal inference. The individual treatment effect of instance i is defined as $\tau_i = y_i^1 - y_i^0$. Noting that in real-world scenarios, only one of the potential outcomes can be observed for each instance, and the remaining unobserved potential outcome is also known as the counterfactual outcome. Inferring the counterfactual outcome from observational data is one of the most challenging tasks in causal inference [25]. Using the above, we provide the formal problem statement as follows: given the set of labeled observational instances $L = \{(\boldsymbol{x}_i, t_i, y^{t_i})\}_{i=1}^N$ and the set of unlabeled instances $U = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_M\}$, our goal is to learn ITE τ_i for each instance i from limited labeled observational data by making use of massive unlabeled data.

Co-training. Co-training [3] is a widely used solution to utilize unlabeled data to aid prediction in semi-supervised learning. In co-training, multiple base learners will be trained on the limited labeled data. Then for each base learner, the

most confident predictions of unlabeled samples predicted by its peer base learners would be chosen to add into the labeled data set and the model will be refined using the newly labeled data set. The above steps will repeat until no base learners update or a preset number of learning rounds has been executed.

3 Proposed Method

In this section, we elaborate the proposed *SemiITE*, a novel framework which can utilize unlabeled instances for ITE estimation. Figure 2 depicts an overview of *SemiITE*. The framework mainly contains three components: a shared module, triple base potential outcome prediction models, and a co-training strategy.

More specifically, we first build a shared module to capture deep information for each instance and to balance the latent representations of treatment group and control group, then build three backbone neural network based models with different structures and initializations to infer potential outcomes, noting that the shared module and the three backbone models are integrated as an ensemble model. We first train the ensemble model using labeled observational instance set L, then in each round of the co-training, we select some instance(s) with its predicted potential outcomes by fitting the unlabeled instances to the trained ensemble model according to the disagreement information of the three backbone prediction models and add them to the labeled instance set, until all the instances in the unlabeled set U are selected or the number of training rounds reaches the preset maximum.

3.1 Model Structure of *SemiITE*

First, we illustrate the model structure of the proposed framework. Generally, the model structure of SemiITE contains a shared module which aims to capture the hidden confounders and three potential outcome prediction models for inferring individual treatment effect.

Shared module. To conduct unbiased ITE estimation [18], we capture the hidden confounders in the proposed framework by building a shared module with a multi-layer neural network. This shared module maps the original covariates to latent space and generates shared latent representation of each instance for the following base potential outcome prediction models. Furthermore, as proved in [26], the representations with closer distance between treatment and control groups can help mitigate the biases in causal effect estimation, thus we refine the representations generated by the shared module to obtain balanced representations between treatment group and control group towards more unbiased ITE estimation, which will be introduced later. For the shared module which is denoted as M_s, we aim to learn a representation learning function $f_s : \mathcal{X} \in \mathbb{R}^d \rightarrow \mathbb{R}^m$, which maps the observed covariates to an m-dimensional latent space. Specifically, we parameterize the representation learning function f_s by stacking L_s neural network layers. The representations generated by the shared

module for instance i can be formulated as follows:

$$\boldsymbol{h}_i = f_s(\boldsymbol{x}_i) = \varphi(\boldsymbol{W}_{L_s}...\varphi(\boldsymbol{W}_1\boldsymbol{x}_i + b_1) + b_{L_s}), \tag{1}$$

where $\boldsymbol{x}_i$ is the original covariates of instance i and $\boldsymbol{h}_i \in \mathbb{R}^m$ is the learned representation of instance i by function f_s, $\varphi(\cdot)$ denotes the ReLU activation function, $\boldsymbol{W}_S$ and b_S $(S = 1, 2, ..., L_s)$ are the learning weight matrix and bias term of the S-th layer, respectively.

Triple base prediction models with diversity. To ensure the effectiveness of the co-training strategy, the base models should be diverse, which can help address the limited label issue [21,8], because if all of the base learners are identical, the training process with multiple learners will degrade to self-training with a single learner. Conventional co-training based methods typically require sufficient and redundant views of data to train the base learners in order to diversify them [19,22]. Given that the requirement for data with sufficient views is too stringent to meet in causal inference due to the expensive and time-consuming nature of observational data collection, in this work we build three outcome prediction models M_1, M_2, and M_3 by stacking multiple neural network layers to infer individual treatment effect for each instance and achieve the diversity of the three outcome prediction models from the following several aspects. First, we let the network structures of the three potential outcome prediction models differ. We assign different values of the depth and width, i.e., the number of hidden layers and neurons in each layer for each base model. Second, we let the learning weights of network architecture of each base model to be initialized by different methods. We take Gaussian random initialization, uniform random initialization, and Glorot initialization [10] on M_1, M_2, and M_3, respectively. Third, we use different optimization methods to update the three potential outcome prediction models. In particular, we use stochastic gradient descent to optimize M_1 and M_2, Adam optimization to optimize M_3. Fourth, we let M_s and M_1 to be updated together after adding selected instances with outcomes predicted by M_2 and M_3 to the set of labeled instances while fixing the other modules. M_2 and M_3 will be updated separately while fixing the other modules, including the shared module M_s.

Specifically, we aim to learn a prediction function $f_v : \mathbb{R}^m \to \mathbb{R}$ $(v = 1, 2, 3)$ for each potential outcome prediction model M_v and parameterize the prediction function f_v by stacking multiple layers of neural network. With the representation $\boldsymbol{h}_i$ of instance i by the aforementioned representation learning function $f_s(\cdot)$ and the corresponding treatment assignment $t_i \in \{0, 1\}$, the predicted outcome y_i with treatment t_i of instance i in model M_v $(v = 1, 2, 3)$ can be computed by the function f_v as:

$$f_v(\boldsymbol{h}_i, t_i) = \begin{cases} \hat{y}_i^{t_i=0} = f_v^0(\boldsymbol{h}_i) & \text{if } t_i = 0 \\ \hat{y}_i^{t_i=1} = f_v^1(\boldsymbol{h}_i) & \text{if } t_i = 1, \end{cases} \tag{2}$$

where f_v^0 and f_v^1 are parameterized by L_v fully connected layers followed by an output layer for $t_i = 0$ and $t_i = 1$, respectively. More specifically, the formulation

of function f_v^t $(t = 0, 1)$ can be written as:

$$f_v^t(\boldsymbol{h}_i) = \boldsymbol{w}^t \varphi(\boldsymbol{W}_{L_v}^t ... \varphi(\boldsymbol{W}_1^t \boldsymbol{h}_i + c_1) + c_{L_v}) + c^t, \tag{3}$$

where $\varphi(\cdot)$ denotes the ReLU activation function, $\boldsymbol{W}_K^t$ and c_K $(K = 1, 2..., L_v)$ are the weight matrix and bias term for the K-th hidden layer, respectively. $\boldsymbol{w}^t$ is the weight vector and c^t is the corresponding bias term of the final prediction layer.

Besides, we propose to utilize HSIC (Hilbert-Schmidt Independence Criterion) [12] to enhance and quantify the diversity of the three base models by measuring the dependence of the predicted factual outcomes from the three prediction models. Assuming that the predicted factual outcomes by M_1, M_2, and M_3 for n samples are $\boldsymbol{Y}_1$, $\boldsymbol{Y}_2$, and $\boldsymbol{Y}_3 \in \mathbb{R}^n$, respectively, the diversity of the three base models can be calculated as follows:

$$\begin{aligned} \mathcal{L}_{div} &= \sum_{i,j \in \{1,2,3\}, i<j} HSIC(\boldsymbol{Y}_i, \boldsymbol{Y}_j), \\ \text{where} \quad HSIC(\boldsymbol{Y}_i, \boldsymbol{Y}_j) &= \frac{1}{(n-1)^2} \boldsymbol{K}_i \boldsymbol{J} \boldsymbol{K}_j \boldsymbol{J}. \end{aligned} \tag{4}$$

Here, $\boldsymbol{K}_i$ and $\boldsymbol{K}_j$ are kernel matrices for prediction vectors $\boldsymbol{Y}_i$ and $\boldsymbol{Y}_j$, respectively. In this work, we use Gaussian kernel for computing the kernel matrix. $\boldsymbol{J} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}$, where $\boldsymbol{I}$ is identity matrix and $\boldsymbol{1}$ is the vector with all elements of 1. Smaller HSIC value indicates stronger independence between the two variables. More details about the derivation can be found in [12]. Furthermore, based on the three potential outcome prediction models with diversity, we can use the disagreement information between them to design a novel co-training strategy to avoid the re-training issue in each instance selection round. More details can be found in section 3.3.

3.2 Loss Function

With the above network structure including M_s, M_1, M_2, and M_3, we design a loss term to combine these components for inferring potential outcomes in an end-to-end manner.

Loss for predicted potential outcomes of M_1, M_2, and M_3. First we need to minimize the difference between the inferred factual outcome by function f_v $(v = 1, 2, 3)$ and the observed factual outcome. We use mean squared error (MSE) function to evaluate the predicted outcomes to approximate the observed factual outcomes. Given the set of labeled observational instances $L = \{(\boldsymbol{x}_i, t_i, y^{t_i})\}_{i=1}^N$, the loss term for the three outcome prediction models can be written as:

$$\mathcal{L}_{mse} = \sum_{v=1}^{3} \sum_{i=1}^{N} MSE(f_v(f_s(\boldsymbol{x}_i), t_i), y^{t_i}), \tag{5}$$

Loss for representation balancing. Due to the fact that we only minimize the error of factual outcomes in Eq. (5), whereas the counterfactual distribution generally differs from the factual distribution, which lead to a biased ITE

estimation [17]. Therefore, it is necessary to balance the distributions of representations generated by the shared module M_s between the treatment group and the control group, which will help with the unbiased ITE estimation. Here, following previous work [14], we use integral probability metric (IPM) to measure the difference between the representation distributions of instances in the treatment group and those in the control group. We denote the balance term as $\mathcal{L}_{IPM}$.

Diversity Term. Besides, we add the HSIC term $\mathcal{L}_{div}$ shown in Eq. (4) into the overall loss function to control and ensure the diversity between the three prediction base models in the training process.

Overall loss. To sum up, the final loss function of our proposed framework *SemiITE* can be written as follows:

$$\mathcal{L} = \mathcal{L}_{mse} + \alpha\mathcal{L}_{IPM} + \beta\mathcal{L}_{div} + \gamma\|\theta\|_2^2, \tag{6}$$

where α, β, and γ are hyperparameters to control the trade-off between corresponding loss term and other terms. $\|\theta\|_2^2$ is the regular term imposed on all learning parameters θ to avoid over-fitting.

3.3 Co-training Strategy via Disagreement

In this subsection, we introduce the co-training strategy of *SemiITE*, in which the framework chooses the predictions of unlabeled instances predicted by the three different potential outcome prediction models based on their disagreement information, and the three models will be refined by these chosen unlabeled instances iteratively. In addition, we illustrate that *SemiITE* avoids re-training issue in each instance selection round, while such issue exists in previous work [35,22] for regression with co-training.

Before introducing the proposed co-training strategy of *SemiITE*, we present the re-training issue in traditional co-training for regression problem, which increases the computational cost greatly. We take an example of traditional co-training regression method [35] to present the issue. Two different regression models are used in this method, which estimates the prediction confidence for each unlabeled sample based on the following principle: whether the error rate of regression model is reduced after adding new predicted sample from unlabeled data set to the training data set. Thus, the method needs to calculate the error reduction rate $\Delta_{\boldsymbol{x}_u} = \sum_{\boldsymbol{x}_i \in L}(y_i - h(\boldsymbol{x}_i))^2 - (y_i - h'(\boldsymbol{x}_i))^2$ on the training set for each candidate unlabeled instance $\boldsymbol{x}_u$, where h is the original learner and h' is the newly trained learner by adding $\boldsymbol{x}_u$ to the training set. Finally, the instance with the largest positive $\Delta_{\boldsymbol{x}_u}$ would be chosen to add into the training set. One can see that we need to re-train the model $|U|$ times in each instance selection round when the two regression models are parameterized (e.g., SVM, Neural networks), which demands a lot of computational resources.

To address the above problem, we utilize the disagreement information of the three outcome prediction models to select unlabeled instance without re-training the models in each instance selection round. Next we introduce the co-training

Algorithm 1 *SemiITE*

Input: The labeled set $L = \{(\boldsymbol{x}_i, t_i, y^{t_i})\}_{i=1}^{N}$, unlabeled set $U = \{\boldsymbol{x}_i\}_{j=1}^{M}$, and the maximum number of instance selection rounds T
Output: The shared module M_s and three outcome prediction models M_1, M_2, and M_3

Initialization:
Build network modules M_s, M_1, M_2, and M_3
L_1, L_2, $L_3 = L$
Train M_s, M_1, M_2, and M_3 based on $\mathcal{L}$ in Eq. (6)
Training:
for $t = 1 \rightarrow T$ **do**
  for $v = 1 \rightarrow 3$ **do**
    $CL_v = \emptyset$
    $\boldsymbol{x}_u \leftarrow$ chosen unlabeled instance based on Eq. (7)
    $CL_v = CL_v \cup (\boldsymbol{x}_u, 1, f_k(f_s(\boldsymbol{x}_u), 1))$ $(k \neq v)$
    $\boldsymbol{x}_w \leftarrow$ chosen unlabeled instance based on Eq. (8)
    $CL_v = CL_v \cup (\boldsymbol{x}_w, 0, f_h(f_s(\boldsymbol{x}_w), 0))$ $(h \neq k, v)$
    $\widetilde{L}_v = L_v \cup CL_v$
    if $v = 1$ **then**
      Train M_s and M_1 on $\widetilde{L}_v$
    else
      Train M_v on $\widetilde{L}_v$
    end if
  end for
end for
return M_s, M_1, M_2 and M_3

strategy of *SemiITE* to avoid the re-training issue when estimating individual treatment effect. First, we utilize the set of labeled observational instances L with treatment assignments and corresponding factual outcomes to train the initialized holistic model, then we can obtain a trained inference model denoted as $\mathcal{M}$. Then we begin to conduct the co-training procedure to choose instances from unlabeled set U for T rounds. In each round, we first fit the unlabeled instances only with covariates into the trained inference model $\mathcal{M}$, then we can obtain the potential outcomes $y_i^{t_i=1}|v$ and $y_i^{t_i=0}|v$ $(v = 1, 2, 3)$ by the function $f_v(\boldsymbol{h}_i, t_i)$ for each unlabeled instance i. After that, we choose unlabeled instances by the disagreement information between M_1, M_2, and M_3 and add the selected instances with their corresponding treatment assignment and predicted outcome into the training set. More specifically, we take a strategy that if the disagreement (i.e., the difference) between two outcome prediction models (e.g., M_2 and M_3) on the prediction of instance i from unlabeled set U is minimal, then the two outcome prediction models will teach the third outcome prediction model (e.g., M_1) on this instance. Then we add the instance i with its covariates, treatment assignment, and the corresponding outcome predicted by M_2 and M_3 into the

training set of instances to train the third prediction model M_1. One can see that there are two predicted outcomes (e.g., by M_2 and M_3) for an instance i with t_i, here we randomly choose one of the two outcomes as the prediction to add into the trained set L. We take an example that chooses unlabeled instances for training prediction model M_1 by the disagreement of M_2 and M_3. The selection formulation can be written as:

$$\arg\min_{\boldsymbol{x}_u \in U} \|f_2(f_s(\boldsymbol{x}_u), t=1) - f_3(f_s(\boldsymbol{x}_u), t=1)\|^2 \tag{7}$$

$$\arg\min_{\boldsymbol{x}_v \in U} \|f_2(f_s(\boldsymbol{x}_v), t=0) - f_3(f_s(\boldsymbol{x}_v), t=0)\|^2, \tag{8}$$

then we will add the selected instances with pseudo-labels and treatment assignments $(\boldsymbol{x}_u, t_u = 1, f_2(f_s(\boldsymbol{x}_u), 1))$ and $(\boldsymbol{x}_v, t_v = 0, f_3(f_s(\boldsymbol{x}_v), 0))$ into the trained set L to train M_s and M_1 based on the loss function in Eq. (6) without MSE losses of M_2 and M_3. Similarly, M_2 and M_3 will be updated by such instance selection procedure while the shared module would be maintained when training M_2 and M_3.

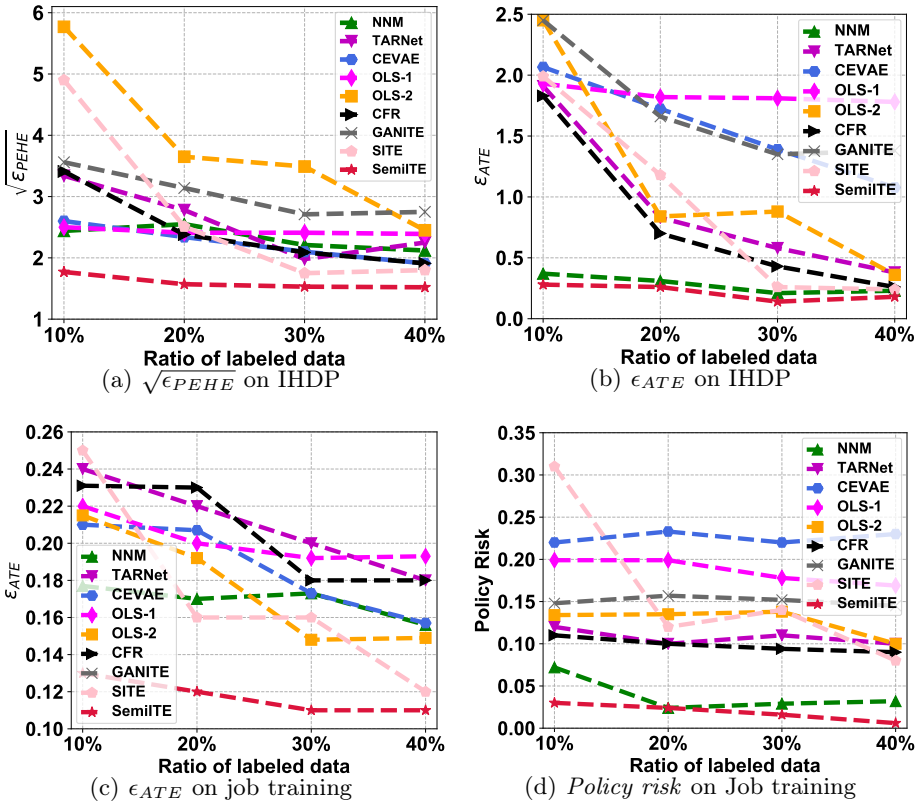
The summary of the proposed framework*SemiITE* is shown in Algorithm 1. Noting that the proposed framework *SemiITE* avoids the re-training issue by utilizing the disagreement information of the three outcome prediction models in each instance selection round, which can greatly reduce the computational cost of the co-training framework. After finishing the co-training procedure, we can infer the individual treatment effect for an unseen instance by any of the three potential outcome prediction models. We use M_1 as an example to infer the ITE of an unseen instance i: $\hat{\tau}_i = f_1(f_s(\boldsymbol{x}_i), 1) - f_1(f_s(\boldsymbol{x}_i), 0)$.

4 Experiments

In this section, we present the experimental results of the proposed framework *SemiITE*, including ITE performance evaluation, ablation study, and hyperparameter study.

4.1 Experimental Setting

Datasets. We conduct the experiments on two benchmark real-world datasets IHDP [4] and Job training [7], which have been widely used in previous works of causal inference [26,17]. In IHDP, each instance's covariates include 25 variables that measure various aspects of children and their mothers. The treatment group's infants receive intensive high-quality childcare and specialist home visits, while the control group's infants do not, and the outcome is the infants' cognitive test scores. In Job training, each instance is an employee with 17 covariates such as age, education, and ethnicity. Instances in the treatment group participate in job training, while those in the control group do not. The outcome of each instance is the employment status.



(a) $\sqrt{\epsilon_{PEHE}}$ on IHDP (b) ϵ_{ATE} on IHDP

(c) ϵ_{ATE} on job training (d) *Policy risk* on Job training

Fig. 3: ITE estimation performance comparison results for different methods on the IHDP data and the Job training data. The horizontal axis represents the proportion of labeled data, and the vertical one denotes the values of metrics (the lower the better).

Baselines. We compare the performance of the proposed framework *SemiITE* on ITE estimation with the following state-of-the-art causal inference models: (1) OLS-1 [26] is the ordinary linear regression model which treats the treatment assignment as a covariate of an instance. (2) OLS-2 [26] are two separated linear regression models for treatment ($t = 1$) and control ($t = 0$) instances. (3) Nearest neighbor matching (NNM) [6] is a matching-based method that infers the potential outcomes of an instance by using its nearby instances. Here, we use the Euclidean distance to measure the similarity between two instances. (4) Causal Effect Variational Autoencoder (CEVAE) [18] follows the causal structure of inference with proxies and builds deep latent variable model to estimate the unknown latent space that summarizes the confounders and the causal effect. (5) Counterfactual Regression (CFR) [26] is a multilayer perceptron based method to infer the counterfactual outcome and minimize the imbalance between treatment and control group. Here we use the Wasserstein-1 distance. (6)

TARNet [17] is a variant of the CFR model, which does not have a built-in representation balancing component. (7) GANITE [31] is a generative adversarial net based model to infer ITE. (8) SITE [30] infers the ITE by capturing hidden confounders and preserving local similarity of data. Due to the fact that the baselines are full-supervised methods, to ensure the fairness of the comparison, we adopt the manifold assumption [16] (i.e., the samples with similar inputs should get similar outputs) and design a corresponding term to add into the loss function for each baseline (except NNM). Assuming that the function $f(\boldsymbol{x}_i, t)$ denotes the predicted outcomes for a certain causal inference model (e.g., CFR) where $\boldsymbol{x}_i, t$ denote the original covariates and treatment assignment of unit i respectively, then a corresponding term based on manifold assumption is added into the loss function of the model:

$$\mathcal{L}_m = \sum_{t} \sum_{i,j=1}^{N+M} w_{ij}(f(\boldsymbol{x}_i, t) - f(\boldsymbol{x}_j, t))^2, \tag{9}$$

where N and M represent the number of labeled and unlabeled samples, respectively. w_{ij} denotes the similarity between two units $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Here, we use the Gaussian kernel to compute the similarity for each unit pair.

Evaluation Metrics. For the IHDP dataset, we adopt two widely used metrics in causal inference for ITE estimation: (1) Rooted Precision in Estimation of Heterogeneous Effect $\sqrt{\epsilon_{PEHE}} = \sqrt{\frac{1}{n}\sum_{i=1}(\tau_i - \hat{\tau}_i)^2}$, where $\tau_i = y_i^{t_i=1} - y_i^{t_i=0}$ and $\hat{\tau}_i = \hat{y}_i^{t_i=1} - \hat{y}_i^{t_i=0}$ are the ground truth ITE and the inferred ITE, respectively; and (2) Mean Absolute Error on ATE $\epsilon_{ATE} = \frac{1}{n}|\sum_{i=1}\hat{\tau}_i - \sum_{i=1}\tau_i|$. For the Job training dataset, we use ϵ_{ATE} and *policy risk*, which is detailed in previous work [17]. Lower values of all metrics denote better performance.

We spilt the data into training set, unlabeled data set, validation set, and test set, where the size of training set is limited. For both IHDP and Job training, we use different ratios of training data to evaluate the performance. We use {10%, 20%, 30%, 40%} of the whole data for labeled set of instances, {70%, 60%, 50%, 40%} for unlabeled set of instances to be selected in co-training rounds, and the rest of the data is used for validation and test set (5% for validation and 15% for test). Regarding the hyperparameters of the proposed framework, we utilize the grid search strategy to find the optimal hyperparameters combination based on the results of validation set. Specifically, for the shared module M_s, we set the number of hidden layer as 3, and the dimension of each hidden layer as 100. For the potential outcome prediction models M_v ($v = 1, 2, 3$), the number of hidden layer varies in {2, 3, 4} and the dimension of each layer ranges in {200, 300, 400, 500}. The trade-off hyperparameters α, β, and γ are set in range $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$. The maximum number of unlabeled instance selection rounds $T = 500$. We use the predictions of M_1 as the final inferred potential outcomes and run the experiments 10 times and the average performance of each method is reported. Besides, all codes are implemented by Python and we use Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz 264G Memory, and NVIDIA Corporation GP100GL.

4.2 ITE Estimation Performance

We compare the proposed framework *SemiITE* against the aforementioned baseline methods with respect to the ITE estimation performance. The results are shown in Figure 3. By analyzing the experiment results we can conclude that:

- (**i**) The supervised state-of-the-art baselines have unsatisfactory performance of ITE estimation when the proportion of labeled observation is low (e.g., below 20%), but their performance gradually improves with the increasing of proportion of labeled data and achieve satisfactory level when the proportion of labeled data is over 40%, which demonstrates that the existing methods require a plenty of labeled observational instances to support their effectiveness for ITE estimation.
- (**ii**) *SemiITE* clearly outperforms several supervised causal inference models with different ratios of labeled instances. And it is worthy to note that the lower the proportion of labeled observational data is, the greater the superiority of the proposed *SemiITE* over other causal inference models will exhibit, which illustrates that *SemiITE* can utilize those unlabeled instances effectively and extract the useful information from them for ITE estimation.
- (**iii**) The performance of *SemiITE* does not change significantly when the proportion of labeled data changes, indicating that *SemiITE* is stable, because *SemiITE* can utilize unlabeled instances to infer potential outcomes more precisely with limited labeled data. Its performance is not greatly affected by the proportion of labeled data, which illustrates that *SemiITE* would still be effective even if the labeled observational data is limited.

4.3 Ablation Study

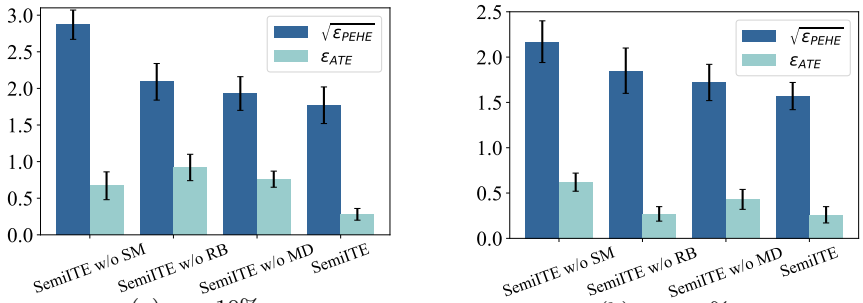
Here we develop the following three variants of *SemiITE* to explore three components of the framework for the individual treatment effect estimation.

- ***SemiITE w/o Shared Module***: This variant does not contain the shared module M_s in the framework, which means that we directly train the three potential outcome models without capturing some hidden confounders. We denote this variant as *SemiITE w/o SM*.
- ***SemiITE w/o representation balance***: This variant does not balance the distribution of representations generated by shared module M_s for the treatment and control groups, i.e., the variant does not add the loss $\mathcal{L}_{IPM}$ into the overall loss function. We denote this variant as *SemiITE w/o RB*.
- ***SemiITE w/o model diversity***: This variant does not consider the diversity of the three prediction models and the diversity term $\mathcal{L}_{div}$ is not added into the loss function. The same network structure, initialization, and optimization method are adopted for the three prediction models. We denote this variant as *SemiITE w/o MD*.

We conduct the ablation study experiment to compare the performance of the proposed *SemiITE* with the aforementioned variants. The results are shown in

Figure 4. Due to the page limit, we only report the results on the IHDP dataset with labeled instance proportion $p = \{10\%, 20\%\}$, but similar observations can also be found in the other datasets and other settings. We have the following observations:

- *SemiITE w/o SM* performs the worst, which demonstrates the importance of capturing hidden confounders in individual treatment effect estimation.
- The performance of *SemiITE w/o RB* is also degraded, because it fails to control the confounding bias, which is a common problem in causal inference.
- *SemiITE w/o MD* also performs worse than the original framework *SemiITE* because the diversity of the three prediction models cannot be achieved, which is important in co-training based semi-supervised learning.



(a) $p = 10\%$ (b) $p = 20\%$

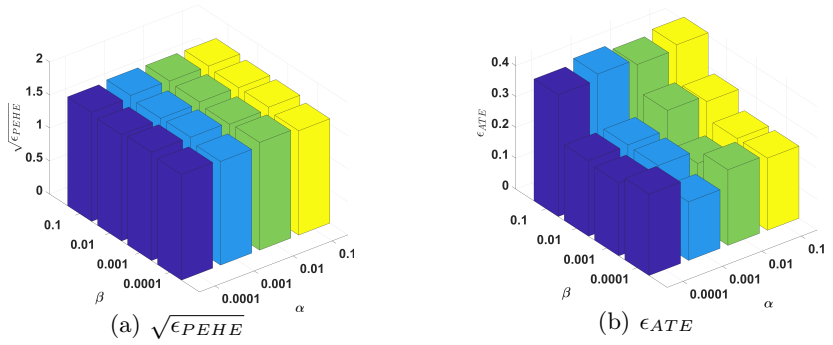
Fig. 4: Ablation study of *SemiITE* on the IHDP dataset.

4.4 Hyperparameter Study

We further explore the impact of two important hyperparameters α and β in Eq. (6) on the performance of the proposed co-training based semi-supervised ITE estiamtion framework. We set the range of the two hyperparameters as $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ and the hyperparameter study results are shown in Figure 5. Due to the page limit, we only report the results for $\sqrt{\epsilon_{PEHE}}$ and ϵ_{ATE} on IHDP with labeled data proportion $p = 10\%$. We have similar results for other datasets with different settings of p. We can observe that the performance is generally stable when the two hyperparameters vary, and the performance is relatively better when α and β range in $\{0.001, 0.01\}$, which demonstrates the robustness of the proposed framework.

5 Related Work

Causal inference with machine learning. Machine learning based causal inference methods have been shown to be effective in observational studies [13,29].



(a) $\sqrt{\epsilon_{PEHE}}$ (b) ϵ_{ATE}

Fig. 5: Hyperparameter study of *SemiITE* on IHDP with labeled data proportion $p = 10\%$.

Among them, k-NN [23] is adopted as a matching strategy to find the instance pair with the closest distance in covariates space but different treatment assignments to obtain causal effect. OLS-1 and OLS-2 [28] infer the causal effect by predicting the potential outcomes using linear regression models. Counterfactual Regression (CFR) [26] casts counterfactual inference as a type of domain adaptation problem and estimates individual treatment effect using neural network by learning balanced representations for instances in control and treatment groups. CEVAE [18] captures the hidden confounders to estimate unbiased causal effect by mapping the original covariates to latent space with variational autoencoder [1]. Yao et al. [30] proposed a local similarity preserved individual treatment effect (SITE) estimation method based on deep representation learning, which can capture the hidden confounders and preserve local similarity of data. However, all these methods are supervised in nature and require massive labeled observational data.

Semi-supervised learning with co-training. The original co-training algorithm was proposed by [3], which assumes that there are two independent, sufficient, and redundant natural views in the sample space, then two separated models can be trained on these two views. Without multiple views of data sample space, many co-training algorithms are proposed to mine useful information from unlabeled data by a single view. For example, Goldman and Zhou [11] proposed to train two different decision tree models from a single view. Zhou and Li [33] adopted a re-sampling strategy to generate three sub-datasets from the original dataset and train three diverse classification models on each generated dataset. Zhou et al. [35] trained two K-NN regressors with different distance orders and chose the predicted unlabeled samples by making the error rate of the regressor reduced most after adding the unlabeled samples into training set. Chen et al. [8] proposed to train three different CNN models with model initialization, diversity augmentation, and pseudo-label editing in a co-training framework. However, most co-training methods are designed for classification and cannot be directly applied to the causal inference problem.

6 Conclusion

In this work, we propose a semi-supervised individual treatment effect estimation framework *SemiITE* via disagreement based co-training, which can effectively utilize unlabeled instances to aid ITE estimation. *SemiITE* chooses the most confident unlabeled instance predictions and then add them into the labeled instance set by the disagreement information of the base prediction models via a co-training strategy, which can avoid the re-training issue in each unlabeled instance selection round. Finally, extensive experiments on two public real-world datasets show the superiority of *SemiITE* on ITE estimation over the existing ITE estimation methods when the labeled data is limited.

7 Acknowledgements

Qiang Huang, Huiyan Sun, and Yi Chang are supported in part by the National Natural Science Foundation of China (No.U19A2065, No.61902144, No.61976102).

References

1. Auto-encoding variational bayes (2014)
2. Appice, A., Guccione, P., Malerba, D.: A novel spectral-spatial co-training algorithm for the transductive classification of hyperspectral imagery data. Pattern Recognition **63**, 229–245 (2017)
3. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the eleventh annual conference on Computational learning theory. pp. 92–100 (1998)
4. Brooks-Gunn, J., Liaw, F.r., Klebanov, P.K.: Effects of early intervention on cognitive function of low birth weight preterm infants. The Journal of pediatrics **120**(3), 350–359 (1992)
5. Cheng, L., Guo, R., Moraffah, R., Candan, K.S., Raglin, A., Liu, H.: A practical data repository for causal learning with big data. In: Proceedings of the 2019 International Symposium on Benchmarking, Measuring and Optimization. pp. 234–248 (2019)
6. Cui, P., Shen, Z., Li, S., Yao, L., Li, Y., Chu, Z., Gao, J.: Causal inference meets machine learning. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining. pp. 3527–3528 (2020)
7. Dehejia, R.H., Wahba, S.: Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American statistical Association **94**(448), 1053–1062 (1999)
8. Dong-DongChen, W., WeiGao, Z.H.: Tri-net for semi-supervised deep learning. In: Proceedings of twenty-seventh international joint conference on artificial intelligence. pp. 2014–2020 (2018)
9. Glass, T.A., Goodman, S.N., Hernán, M.A., Samet, J.M.: Causal inference in public health. Annual review of public health **34**, 61–75 (2013)
10. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 249–256. JMLR Workshop and Conference Proceedings (2010)

11. Goldman, S., Zhou, Y.: Enhancing supervised learning with unlabeled data. In: International conference on machine learning. pp. 327–334. Citeseer (2000)
12. Gretton, A., Bousquet, O., Smola, A., Schölkopf, B.: Measuring statistical dependence with hilbert-schmidt norms. In: ALT. pp. 63–77. Springer (2005)
13. Guo, R., Cheng, L., Li, J., Hahn, P.R., Liu, H.: A survey of learning causality with data: problems and methods. ACM Computing Surveys (CSUR) **53**(4), 1–37 (2020)
14. 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. pp. 232–240 (2020)
15. Hill, J.L.: Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics **20**(1), 217–240 (2011)
16. Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Label propagation for deep semi-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5070–5079 (2019)
17. Johansson, F., Shalit, U., Sontag, D.: Learning representations for counterfactual inference. In: International conference on machine learning. pp. 3020–3029. PMLR (2016)
18. Louizos, C., Shalit, U., Mooij, J., Sontag, D., Zemel, R., Welling, M.: Causal effect inference with deep latent-variable models. arXiv preprint arXiv:1705.08821 (2017)
19. Ma, F., Meng, D., Dong, X., Yang, Y.: Self-paced multi-view co-training. Journal of Machine Learning Research (2020)
20. Ma, J., Dong, Y., Huang, Z., Mietchen, D., and Li, J.: Assessing the Causal Impact of COVID-19 Related Policies on Outbreak Dynamics: A Case Study in the US. In: Proceedings of the ACM Web Conference 2022. pp. 2678–2686 (2022)
21. Nigam, K., Ghani, R.: Analyzing the effectiveness and applicability of co-training. In: Proceedings of the ninth international conference on Information and knowledge management. pp. 86–93 (2000)
22. Ning, X., Wang, X., Xu, S., Cai, W., Zhang, L., Yu, L., Li, W.: A review of research on co-training. Concurrency and computation: practice and experience p. e6276 (2021)
23. Peterson, L.E.: K-nearest neighbor. Scholarpedia **4**(2), 1883 (2009)
24. Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology **66**(5), 688 (1974)
25. Rubin, D.B.: Causal inference using potential outcomes: design, modeling, decisions. Journal of the American Statistical Association **100**(469), 322–331 (2005)
26. Shalit, U., Johansson, F., et al: Estimating individual treatment effect: generalization bounds and algorithms. In: International conference on machine learning. pp. 3076–3085 (2017)
27. Varian, H.R.: Causal inference in economics and marketing. Proceedings of the National Academy of Sciences **113**(27), 7310–7315 (2016)
28. Weisberg, S.: Applied linear regression, vol. 528. John Wiley & Sons (2005)
29. Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., Zhang, A.: A survey on causal inference. arXiv preprint arXiv:2002.02770 (2020)
30. 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 **31** (2018)
31. Yoon, J., Jordon, J., Van Der Schaar, M.: Ganite: Estimation of individualized treatment effects using generative adversarial nets. In: International Conference on Learning Representations (2018)

32. Zhang, X., Song, Q., Liu, R., Wang, W., Jiao, L.: Modified co-training with spectral and spatial views for semisupervised hyperspectral image classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **7**(6), 2044–2055 (2014)
33. Zhou, Z.H., Li, M.: Tri-training: Exploiting unlabeled data using three classifiers. IEEE Transactions on knowledge and Data Engineering **17**(11), 1529–1541 (2005)
34. Zhou, Z.H., Li, M.: Semi-supervised learning by disagreement. Knowledge and Information Systems **24**(3), 415–439 (2010)
35. Zhou, Z.H., Li, M., et al.: Semi-supervised regression with co-training. In: Proceedings of international joint conference on artificial intelligence. vol. 5, pp. 908–913 (2005)