Estimating Individual Treatment Effect from Educational Studies with Residual Counterfactual Networks

Siyuan Zhao
Worcester Polytechnic Institute
100 Institute Road
Worcester, MA 01609, USA
szhao@wpi.edu

Neil Heffernan
Worcester Polytechnic Institute
100 Institute Road
Worcester, MA 01609, USA
nth@wpi.edu

ABSTRACT

Personalized learning considers that the causal effects of a studied learning intervention may differ for the individual student. Making the inference about causal effects of studies interventions is a central problem. In this paper we propose the Residual Counterfactual Networks (RCN) for answering counterfactual inference questions, such as "Would this particular student benefit more from the video hint or the text hint when the student cannot solve a problem?". The model learns a balancing representation of students by minimizing the distance between the distributions of the control and the treated populations, and then uses a residual block to estimate the individual treatment effect based on the representation of the student. We run experiments on semi-simulated datasets and real-world educational online experiment datasets to evaluate the efficacy of our model. The results show that our model matches or outperforms the state-of-the-art.

Keywords

Counterfactual inference, deep residual learning, educational experiments, individual treatment effect

1. INTRODUCTION

The goal of personalized learning is to provide pedagogy, curriculum, and learning environments to meet the needs of individual students. For example, an Intelligent Tutor System (ITS) decides which hints would most benefit a specific student. If the ITS could infer what the student performance would be after receiving each hint, then it would simply choose the hint which leads to the best performance for the student. To make this possible, we might run an online educational experiment by randomly assigning students to one of the hints, and collect student performance. Then making predictions about causal effects of possible interventions (e.g. available hints) becomes a central problem in this case. In this paper we focus on the task of answering counterfactual questions [8] such as, "Would this particular student benefit more from the video hint or the text hint when the student cannot solve a problem?"

There are two ways of collecting data for counterfactual inference: randomized control trials (RCTs) and observational studies. In RCTs, participants (e.g. students) are randomly assigned to interventions (e.g. video hints or text hints), while participants in observational studies are not essentially randomly assigned to interventions. For example, consider the experiment of evaluating the efficacy of video hints and text hints for a certain problem. Under the design of RCT, students who need a hint would be randomly assigned to either the video hints or the text hints. In an observational study, students are assigned to one of the interventions based on their contextual information, such as knowledge level or personal preference.

[5] proposed Balancing Neural Networks (BNN) which can be applied to solve the counterfactual inference problem. They used a form of regularizer to enforce the similarity between the distributions of representations learned for populations with different interventions, for example, the representations for students who received text hints versus those who received video hints.This reduces the variance from fitting a model on one distribution and applying it to another. Because of random assignment to the interventions in RCTs, the distributions of the populations within different interventions are highly likely to be identical. However, in the observational study, we may end up with the situation where only male students receive video hints and female students receive text hints. Without enforcing the similarity between the distributions of representations for male and female students, it is not safe to make a prediction of the outcome if male students receive text hints. In machine learning, "domain adaptation" [7] refers to the dissimilarity of the distributions between the training data and the test data.

Recent work [6] has demonstrated that (deep) neural networks can be used with domain adaptation approaches to produce outstanding results on some domain adaptation benchmark datasets. Motivated by their work, we propose the Residual Counterfactual Networks (RCN) for the counterfactual inference to estimate the individual treatment effect and evaluate its efficacy in both a simulated dataset and a real-world dataset from an educational online experiment. The RCN extends the BNN by adding a residual block to estimate the individual treatment effect (ITE) based on the learned representation of participants. The idea of the residual block is originated from the state-of-the-art deep residual learning [2]. We enable the estimation of ITE by plugging several layers into neural networks to explicitly learn the residual function with reference to the learned representation.

The rest of the paper is organized as follows. Section 2 provides an overview of the problem setup of counterfactual inference for estimating the ITE. Section 3 details information of our model. Section 4 gives an overview of related work in this research area. Section 5 describes the datasets and evaluation metrics used to test our model. Section 6 presents the results of our model and compares them with other models. Finally, we discuss the results and conclude the paper.

2. PROBLEM SETUP

Let $\mathcal{T}$ be the set of proposed interventions we wish to consider, X the set of participants, and Y the set of possible outcomes. For each proposed intervention $t \in \mathcal{T}$, let $Y_t \in Y$ be the potential outcome for x when x is assigned to the intervention t. In randomized control trial (RCT) and observed study, only one outcome is observed for a given participant x; even if the participant is given an intervention and later the other, the participant is not in the same state. In machine learning, ”bandit feedback” refers to this kind of partial feedback. The model described above is also known as the Rubin-Neyman causal model [11, 10].

We focus on a binary intervention set $\mathcal{T} = \{0, 1\}$, where intervention 1 is often referred as the ”treated” and intervention 0 is the ”control.” In this scenario the ITE for a participant x is represented by the quantity of $Y_1(x) - Y_0(x)$. Knowing the quantity helps assign participant x to the best of the two interventions when making a decision is needed, for example, choosing the best intervention for a specific student when the student has a trouble solving a problem. However, we cannot directly calculate ITE due to the fact that we can only observe the outcome of one of the two interventions.

In this work we follow the common simplifying assumption of no-hidden confounding variables. This means that all the factors determining the outcome of each intervention are observed. This assumption can be formalized as the strong ignorability condition:

$$(Y_1, Y_0) \perp t|x, 0 < p(t = 1|x) < 1, \forall x.$$

Note that we cannot evaluate the validity of strong ignorability from data, and the validity must be determined by domain knowledge.

In the ”treated” and the ”control” setting, we refer to the observed and unobserved outcomes as the factual outcome $y^F(x)$, and the counterfactual outcome $y^{CF}(x)$ respectively. In other words, when the participant x is assigned to the ”control” ($t = 0$), $y^F(x)$ is equal to $Y_1(x)$, and $y^{CF}(x)$ is equal to $Y_0(x)$. The other way around, $y^F(x)$ is equal to $Y_0(x)$, and $y^{CF}(x)$ is equal to $Y_1(x)$.

Given n samples $\{(x_i, t_i, y_i^F)\}_{i=1}^n$, where $y_i^F = t_i \cdot Y_1(x_i) + (1-t_i)Y_0(x_i)$, a common approach for estimating the ITE is to learn a function $f : X \times T \to Y$ such that $f(x_i, t_i) \approx y_i^F$. The estimated ITE is then:

$$I\hat{T}E(x_i) = \begin{cases} y_i^F - f(x_i, 1-t_i), & t_i = 1. \\ f(x_i, 1-t_i) - y_i^F, & t_i = 0. \end{cases}$$

We assume n samples $\{(x_i, t_i, y_i^F)\}_{i=1}^n$ form an empirical distribution $\hat{p}^F = \{(x_i, t_i)\}_{i=1}^n$. We call this empirical distribution $\hat{p}^F \sim p^F$ the empirical factual distribution. In order to calculate ITE, we need to infer the counterfactual outcome which is dependent on the empirical distribution $\hat{p}^{CF} = \{(x_i, 1-t_i)\}_{i=1}^n$. We call the empirical distribution $\hat{p}^{CF} \sim p^{CF}$. The p^F and p^{CF} may not be equal because the distributions of the control and the treated populations may be different. The inequality of two distributions may cause the counterfactual inference over a different distribution than the one observed from the experiment. In machine learning terms, this scenario is usually referred to as domain adaptation, where the distribution of features in test data are different than the distribution of features in training data.

3. MODEL

We proposed RCN to estimate individual treatment effect using counterfactual inference. The RCN first learns a balancing representation of deep features $\Phi : X \to R^d$, and then learns a residual mapping Δf on the representation to estimate the ITE. The structure of the RCN is shown in the left side of Figure 1.

To learn a representation of deep features Φ, the RCN uses fully connected layers with ReLu activation function, where $Relu(z) = max(0, z)$. We need to generalize from factual distribution to counterfactual distribution in the feature representation Φ to obtain accurate estimation of counterfactual outcome. The common successful approaches for domain adaptation encourage similarity between the latent feature representations w.r.t the different distributions. This similarity is often enforced by minimizing a certain distance between the domain-specific hidden features. The distance between two distributions is usually referred to as the discrepancy distance, introduced by [7], which is a hypothesis class dependent distance measure tailored for domain adaptation.

In this paper we use an Integral Probability Metric (IPM) measure of distance between two distributions $p_0 = p(x|t = 0)$, and $p_1 = p(x|t = 1)$, also known as the control and treated distributions. The IPM for p_0 and p_1 is defined as

$$\text{IPM}_{\mathcal{F}}(p_0, p_1) := \sup_{f \in \mathcal{F}} \left| \int_S f dp_0 - \int_S f dp_1 \right|,$$

where $\mathcal{F}$ is a class of real-valued bounded measurable functions on S.

The choice of functions is the crucial distinction between IPMs [15]. Two specific IPMs are used in our experiments: the Maximum Mean Discrepancy (MMD), and the Wasserstein distance. When $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leqslant 1\}$, where $\mathcal{H}$ represents a reproducing kernel Hilbert space (RKHS) with k as its reproducing kernel, $\text{IPM}_{\mathcal{F}}$ is called MMD. In other words, the family of norm-1 reproducing kernel Hilbert space

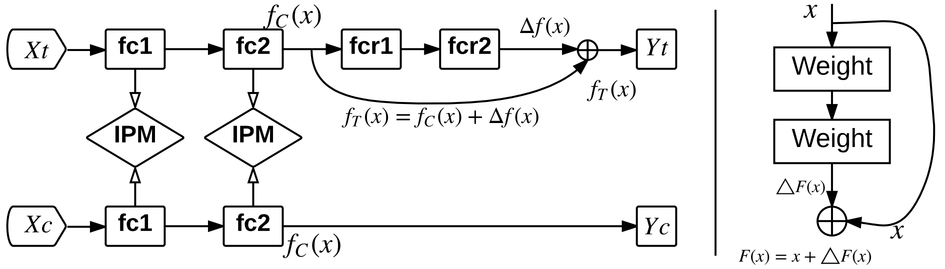

Figure 1: (left) Residual Counterfactual Networks for counterfactual inference. IPM is adopted on layers fc1 and fc2 to minimize the discrepancy distance of the deep features of the control and the treated populations. For the treated group, we add a residual block fcr1-fcr2 so that $f_T(x) = f_C(x) + \Delta f(x)$; (right) Residual block

(RKHS) functions lead to the MMD. The family of 1-Lipschitz functions $\mathcal{F} = \{f : \|f\|_L \leq 1\}$, where $\|f\|_L$ is the Lipschitz semi-norm of a bounded continuous real-valued function f, make IPM the Wasserstein distance. Both the Wasserstein and MMD metrics have consistent estimators which can be efficiently computed in the finite sample case [14]. The important property of IPM is that $p_0 = p_1$ iff $\mathrm{IPM}_{\mathcal{F}}(p_0, p_1) = 0$.

The representation with reduction of the discrepancy between the control and the treated populations helps the model to focus on balancing features across two populations when inferring the counterfactual outcomes. For instance, if in an experiment, almost no male student ever received intervention A, inferring how male students would react to intervention A is highly prone to error and a more conservative use of the gender feature might be warranted.

After balancing the feature representations of the control and the treated populations, the next step is to infer the treatment effect for participant x. We adopt the residual block [2] to estimate the treatment effect.

As shown in the right side of Figure 1, $F(x)$ is the underlying desired function mapping. Instead of stacking a number of layers to fit the desired $F(x)$, we let stacked fully connected layers learn the residual mapping $\Delta f(x) = F(x) - x$. Then the origin mapping is converted into $\Delta f(x) + x$. The operation $\Delta f(x) + x$ is performed by a shortcut connection and an element-wise addition. Learning residual mapping is favored over fitting the desired mapping directly, because it is easier to find the residual with reference to an identity mapping than to learn the mapping as new.

The goal of the residual block is to approximate a residual function Δf such that $f_T(x) = f_C(x) + \Delta f(f_C(x))$, where f_C is the deep representation of participant x before being fed into the output layer, and f_T is the input to the output layer for the treated population. The output layer is a ridge linear regression to generate the final outcome. From the definition of the residual function Δf, we see that $\Delta f(x)$ is the estimated treatment effect for participant x, which is our interest in a control and treated experiment. With the residual block directly connected to fc2, the residual function $\Delta f(x)$ is dependent on the feature representation of participant x.

We plug in the residual block (shown in Figure 1) between fc2 layer and final output layer for the treated population in order to estimate the ITE. There is no residual block plugged in between fc2 layer and the final output layer for the control population. The final output layer $\varphi(\cdot)$ is a linear regression to calculate the predicted outcome, such that $Yc = \varphi(f_C(x))$, and $Yt = \varphi(f_T(x))$.

Recall the problem setup described above that there exist n samples $\{(x_i, t_i, y_i^F)\}_{i=1}^n$, where $y_i^F = t_i \cdot Y_1(x_i) + (1 - t_i)Y_0(x_i)$. In the control and the treated setting, we assume that $n_c(n_c > 0)$ samples $\left\{(x_i, 0, y_i^{(0)})\right\}_{i=1}^{n_c} \sim D_c$ are assigned to the control ($t = 0$), and $n_t(n_t > 0)$ samples $\left\{(x_i, 1, y_i^{(1)})\right\}_{i=1}^{n_t} \sim D_t$ are assigned to the treated ($t = 1$), such that $n = n_c + n_t$. As described above, RCN is an integration of deep feature learning, feature representation balancing, and treatment effect estimation in an end-to-end fashion with the loss function as such:

$$\begin{aligned}\min_{f_T=f_S+\Delta f(f_S)} \quad & \frac{1}{n_c}\sum_{i=1}^{n_c} L(f_c(\mathbf{x}_i), y_i^{(0)}) \\ & + \frac{1}{n_t}\sum_{i=1}^{n_t} L(f_t(\mathbf{x}_i), y_i^{(1)}) \\ & + \lambda \cdot \mathrm{IPM}(D_c, D_t),\end{aligned}$$

where λ is the tradeoff parameter for the IPM penalty, L is the loss function of the model. In the case of binary classification, L is the standard cross entropy. In the case of regression, L is root-mean-square error (RMSE). During the training, the model only has the access to the factual outcome.

4. RELATED WORK

From a conceptual point of view, our work is inspired by the work on domain adaptation and deep residual learning. [6] proposed the Residual Transfer Network that adopt MMD distance to learn transferable deep features from labeled data in the source domain and unlabeled data in the target domain and adds a residual block to transfer the prediction classifier from the target domain to the source domain. The structure of our model is similar to that of their model. Deep residual learning is introduced by [2], the winner of the ImageNet ILSVRC 2015 challenge, to ease the training of deep networks. The residual block is designed to learn residual functions $\Delta F(\mathbf{x})$ with reference to the layer input $\mathbf{x}$. Reformulating layers to the residual block makes the training easier than directly learning the original functions $F(\mathbf{x}) = \Delta F(\mathbf{x}) + \mathbf{x}$.

Our model extends the work by [5, 13], where the authors build a connection between domain adaptation and counterfactual inference. They use IPMs, such as MMD and wasserstein distance, to learn a representation of the data which balances the control and treated distributions. The treatment assignment is concatenated with the representation to predict the factual outcome as while the reverse treatment assignment is concatenated with the representation to predict the counterfactual outcome. Compared to their work, we add a residual block to estimate the individual treatment effect based on the representation. [17, 1] proposed random causal forests (RCF) which is built upon the idea of random forests to estimate the heterogeneous treatment effect.

5. EXPERIMENTS

5.1 Evaluation Metrics

To compare among various models, we report the RMSE of estimated individual treatment effect, denoted

$$\epsilon_{ITE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((Y_1(x_i) - Y_0(x_i)) - I\hat{T}E(x_i))^2},$$

and the absolute error in average treatment effect

$$\epsilon_{ATE} = \left|\frac{1}{n}\sum_{i=1}^{n}(f_t(x_i) - f_s(x_i)) - \frac{1}{n}\sum_{i=1}^{n}(Y_1(x_i) - Y_0(x_i))\right|.$$

Following [4, 5], we report the Precision in Estimation of Heterogeneous Effect (PEHE),

$$PEHE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((Y_1(x_i) - Y_0(x_i)) - (\hat{y_1}(x_i) - \hat{y_0}(x_i))^2}.$$

Compared to the fact that achieving a small RMSE of estimated ITE needs the accurate estimation of counterfactual responses, a good (small) PEHE requires the accurate estimation of both factual and counterfactual responses.

However, calculating ϵ_{ITE}, ϵ_{ATE}, and PEHE requires the "ground truth" of the ITE for each participant in the experiment. We cannot gather the counterfactual outcomes from RCTs and observational studies, and thus do not have the ITE of each participant. We cannot evaluate ϵ_{ITE} and PEHE on these datasets. In order to evaluate the performance on these datasets across various models, we use a measure, called policy risk, introduced by [13]. Given a model f, the participant x is assigned to the treatment $\pi_f(x) = 1$ if $f(x,1) - f(x,0) > \lambda$ (in the case of RCN, $\Delta f > \lambda$), where λ is the treatment threshold, and to the control $\pi_f(x) = 0$ otherwise. The risk policy is defined as:

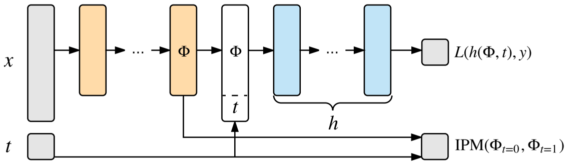

Figure 2: CFR for ITE estimation. L is a loss function, IPM is an integral probability metric

$$R_{Pol}(\pi_f) = 1 - (\mathbb{E}[Y_1|\pi_f(x) = 1] \cdot p(\pi_f = 1) \\ + \mathbb{E}[Y_0|\pi_f(x) = 0] \cdot p(\pi_f = 0)).$$

The empirical estimator of the risk policy on a dataset is calculated by:

$$\hat{R}_{Pol}(\pi_f) = 1 - (\mathbb{E}[Y_1|\pi_f(x) = 1, t = 1] \cdot p(\pi_f = 1) \\ + \mathbb{E}[Y_0|\pi_f(x) = 0, t = 0] \cdot p(\pi_f = 0)).$$

To obtain the policy risk, we use the method introduced by [16]. We select a subset of participants in the dataset where the treatment recommendation inferred by the model is the same as the treatment assignment in the experiment and then calculate the average loss from the subset of the data (see Table 1 for illustrative data).

For the datasets without the "ground truth" on ITE, we also calculate the average treatment effect on the treated by ATT $= \frac{1}{n_t}\sum_{i=1}^{n_t} y_i^{(1)} - \frac{1}{n_s}\sum_{i=1}^{n_s} y_i^{(0)}$, and report the error on ATT as $\epsilon_{ATT} = \left|\text{ATT} - \frac{1}{n_t}\sum_{i=1}^{n_t}(f_t(x_i) - f_s(x_i))\right|$.

5.2 Baselines

Balancing Neural Networks (BNN) is a neural networks-based model for counterfactual inference. Compared to RCN, it has exactly the same fc1 and fc2 layers with IPM regularizer to learn the representation $\Phi(x)$ of the participant x. However, instead of using residual block to estimate treatment effect, it concatenates the treatment assignment t_i to the output of fc2 layer $\Phi(x)$ and feeds $[\Phi(x_i), t_i]$ to another two fully connected layers to generate the predicted outcome. We refer to this particular structure of BNN as BNN-2-2, following [5].

The Counterfactual Regression (CFR) [13] is built on the BNN. The important difference between these two models is that the CFR uses a more powerful distribution metric in the form of IPMs to learn a balancing representation. We compare our model with BNN-2-2 and CFR to verify the efficacy of residual block in terms of estimating individual treatment effect.

We introduce a simple neural networks baseline model to evaluate the efficacy of the IPM regularizer and residual mapping. This baseline model is a feed-forward neural networks model with four hidden layers, trained to predict the factual outcome based on X and t, without the IPM regularizer and the residual block. We refer to this as NN-4.

Table 1: Hypothetical data for some example students. The predicted outcome is the probability that the student would complete the assignment. Students in bold are those whose randomized treatment assignment is congruent with the recommendation of the counterfactual inference model. Data from these students would be used to calculate the policy risk.

ID	Group	Completion	Predicted outcome if treated	Predicted outcome if not treated	Treatment effect	Treat?
1	Control	1	0.8	0.75	0.05	1
2	**Control**	**0**	**0.3**	**0.45**	**-0.15**	**0**
3	**Treatment**	**0**	**0.50**	**0.38**	**0.12**	**1**
4	Treament	1	0.91	0.99	-0.08	0

5.3 Simulation based on real data - IHDP

The Infant Health and Development Program (IHDP) dataset was a semi-simulated dataset introduced by [4]. The dataset consists of a number of covariates from a real randomized experiment. The goal of the experiment is to study the impact of superior child care and home visits on future cognitive test scores. [4] discarded a biased subset of the treated population in order to introduce imbalance between treated and control subjects and used a simulated counterfactual outcome. Eventually, there are 747 subjects (139 treated, 608 control), each represented by 25 covariates assessing the attributes of the children and their mothers.

5.4 ASSISTments dataset

The ASSISTments online learning platform [3] is a free web-based platform utilized by a large user-base of teachers and students. The platform has been the subject of a recent study within the state of Maine [9], demonstrating significant learning gains for students using the platform. The dataset used in this work comes from one of 22 randomized controlled experiments [12] collected within the platform. This experiment was run in assignment types known as ”skill builders” in which students are given problems until a threshold of understanding is reached; within ASSISTments, this threshold is traditionally three consecutive correct responses. Reaching this threshold denotes sufficient performance and completion of the assignment. In addition to this experimental data, information of the students prior to condition assignment is also provided in the form of problem-level log data providing a breadth of student information at fine levels of granularity.

In this experiment, there are two kinds of hints (video versus text) available for each problem from the assignment when students answer the problem incorrectly. The assignment to the video hint and the text video was random. Video content was designed to mirror text hint in an attempt to provide identical assistance. There are 147 students who received the video hint and 237 students who received the text hint. The dataset includes 15 covariates such as student past-performance history, class-past performance history. We solve a binary classification task which is to predict the completion of the assignment for each student.

6. RESULTS

The results of IHDP is presented in Table 2 when the treatment threshold $\lambda = 0$. We see that our proposed RCN performs the best on the dataset in terms of estimating ITE, ATE and PEHE. There is an especially large improvement on estimating ITE. These results indicate that the residual block $\Delta f(x)$ helps accurately predict the value of ITE based on the feature representation $\Phi(x)$ for a given participant x.

Table 2: Results of IHDP

Model	ϵ_{ITE}	ϵ_{ATE}	**PEHE**
NN-4	2.0	0.5	1.9
BNN-2-2	1.7	0.3	1.6
CFR	1.4	0.2	1.6
RCN	1.1	0.05	1.4

The results of ASSISTments dataset are the interest of our work since we hope to apply the RCN to educational experiments in order to support decision making in terms of personalized learning. The results in terms of policy risk and the average treatment effect on the treated are shown in Table 3 when the treatment threshold $\lambda = 0$. The model TA means ”Treated All” where all students are assigned to the treatment while the model NT means ”Not Treated” where all students are assigned to the control. Without considering that the effects of an intervention may differ for individual students, the model with the better performance out of these two models would be adopted when a choice must be made between these two interventions. The RCN, which considers the individual treatment effect, outperforms the TA and the NT. This indicates that taking the individual effect into account helps make a better choice of interventions. The comparison between the CFR and the RCN suggests that the RCN performs better than the CFR does in terms of risk policy and ATT.

To investigate the correlation between policy risk and treatment threshold λ, we plot the value of policy risk as a function of treatment threshold λ in Figure 3. For the results of the ASSISTments dataset from the CFR, the maximum predicted ITE in the dataset is 0.44. Once the threshold λ is larger than 0.44, the CFR is converted to ”Not Treated” where all students are assigned to the control. Since the maximum predicted ITE in the ASSISTments dataset from the CFR is 0.18, the CFR is converted to ”Not Treated” once the treatment threshold λ is larger than 0.18.

7. CONCLUSION

As online educational experiments become popular and easy to conduct, and machine learning becomes a major tool for researchers, counterfactual inference gains a lot of interest for the purpose of personalized learning. In this paper we

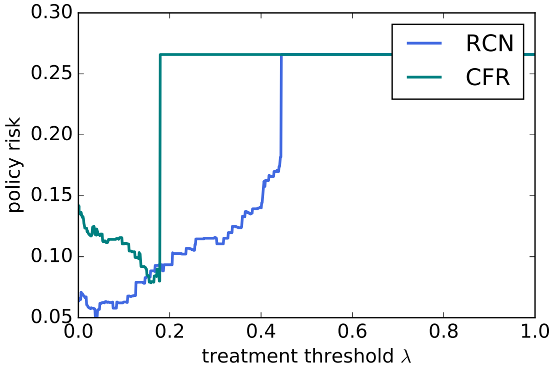

Figure 3: Treatment threshold versus policy risk on ASSISTments dataset. The lower policy risk is the better.

Table 3: Results of the ASSISTments Dataset

Model	$R_{\mathbf{POL}}$	ϵ_{ATT}
TA	0.14	-
NT	0.27	-
CFR	0.14	0.08
RCN	0.08	0.03

propose the Residual Counterfactual Networks (RCN) to estimate the individual treatment effect. Because of the dissimilarity between the distributions of the control and the treated populations, the RCN uses IPMs, such as Wasserstein and MMD distance, to learn balancing deep features from the data. A residual block is adopted on the deep features to learn the individual treatment effect (ITE) so that estimation of the ITE is dependent on the deep features. We apply our model to both synthetic datasets and real-world datasets from online educational experiment, indicating that our model achieves the state-of-the-art.

One open question for the future work is how to generalize our model for the situations where there is more than one treatment in the experiment. Integral Probability Metric (IPM) can only measure the distance between two distributions. We could use pair-wised IPM if there are more than two distributions. But this would be computationally time-consuming if the number of distributions increases. Since running experiments is expensive and collecting enough data for the model to make a reliable prediction is difficult, we need a better optimization algorithm which allows us to train the model efficiently.

8. ACKNOWLEDGMENTS

We acknowledge funding from multiple NSF grants (ACI-1440753, DRL-1252297, DRL-1109483, DRL-1316736 & DRL-1031398), the U.S. Department of Education (IES R305A120125 & R305C100024 and GAANN), the ONR, and the Gates Foundation.

9. REFERENCES

[1] S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 5 July 2016.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. pages 770–778, 2016.

[3] N. T. Heffernan and C. L. Heffernan. The assistments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24(4):470–497, 2014.

[4] J. L. Hill. Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Stat.*, 20(1):217–240, 2011.

[5] F. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 3020–3029, 2016.

[6] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems 29*, pages 136–144. Curran Associates, Inc., 2016.

[7] Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. 2009.

[8] J. Pearl. Causal inference in statistics: An overview. *Stat. Surv.*, 3(0):96–146, 2009.

[9] J. Roschelle, M. Feng, R. F. Murphy, and C. A. Mason. Online mathematics homework increases student achievement. *AERA Open*, 1 Oct. 2016.

[10] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, 66(5):688, Oct. 1974.

[11] D. B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *J. Am. Stat. Assoc.*, 2005.

[12] D. Selent, T. Patikorn, and N. Heffernan. ASSISTments dataset from multiple randomized controlled experiments. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, L@S '16, pages 181–184, New York, NY, USA, 2016. ACM.

[13] U. Shalit, F. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. 13 June 2016.

[14] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, G. R. Lanckriet, et al. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

[15] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet. On integral probability metrics, φ-divergences and binary classification. 18 Jan. 2009.

[16] A. J. Vickers, M. W. Kattan, and S. Daniel. Method for evaluating prediction models that apply the results of randomized trials to individual patients. *Trials*, 8(1):14, 5 June 2007.

[17] S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. 14 Oct. 2015.