Treatment Effect Prediction with Generative Adversarial Networks using Electronic Health Records

Jiebin Chu[1], Wei Dong[2], Zhengxing Huang[1]

[1] College of Biomedical Engineering and Instrumental Science, Zhejiang University
2 Department of Cardiology, Chinese PLA General Hospital
zhengxinghuang@zju.edu.cn

Abstract. Treatment effect prediction (TEP) plays a vital role in disease management by ensuring that the expected clinical outcomes are obtained after performing specialized and sophisticated treatments on patients given their personalized clinical status. To address this problem, we propose an adversarial deep treatment effect prediction model by utilizing the potential of a large volume of electronic health records (EHR) data. Our model employs two auto-encoders for learning the representative and discriminative features of both patient characteristics and treatments from EHR data. The discriminative power of the learned features is further enhanced by decoding the correlational information between the patient characteristics and subsequent treatments by mean s of a generative adversarial learning strategy. Thereafter, a logistic regression layer is appended on the top of the resulting feature representation layer for TEP. The proposed model was evaluated on a real clinical dataset and the experimental results demonstrate that our proposed model achieves competitive performance compared to state-of-the-art models in tackling the TEP problem.

Keywords: Treatment Effect Prediction, Deep Learning, Adversarial learning, Electronic Health Records.

1 Introduction

Treatment effect prediction (TEP), as ensuring to obtain the expected clinical outcomes after performing specialized and sophisticated treatments on patients given their personalized clinical status, is vital for disease management. Traditional approaches to addressing this problem have mostly relied on randomized controlled trial (RCT) studies [1], which urges healthcare professionals to make treatment decisions according to the best evidence from systematic research on both the efficacy and efficiency of various therapeutic alternatives [2]. Although valuable, there are several typical limitations to RCT studies [1]. Specifically, participants in RCTs are strictly selected and tend to be a "pretty rarefied population", which is not representative of the real-world population that the scheduled treatments will eventually target [3].

Electronic health records (EHRs), with their increasingly widespread adoption in clinical practice, provide a comprehensive source for treatment effect analysis to augment traditional RCT studies [4-6].The different aspects of medical information recorded in EHR data are highly correlated and thus provide significant potential for exploitation, for example, to extract representative and discriminative features for treatment effect prediction (TEP).

In this study, we propose a novel adversarial deep treatment effect prediction (ADTEP) model to anticipate treatment effects by utilizing a large volume of EHR data. In detail, two Auto-encoders (AE) are employed to encode the physical condition and treatment information of patient samples into latent robust representations. To align the generated treatments with the actual performed treatments, we adopt an adversarial learning scheme [7] and use a discriminator to differentiate the fake generated treatments from the real performed treatments documented in the EHR data. With this adversarial learning strategy, not only the patient characteristics and subsequent treatments, but also the correlational information between them are encoded in the latent representation, making the generated features sufficiently representative to convey the essential and critical information in the EHR data.

2 Methods

We consider a typical clinical study of TEP, in which the EHR data record patient features, treatment interventions, and achieved treatment outcomes. For each patient sample u, we observe a set of patient features $\boldsymbol{x}_{\mathrm{u}}$, a set of treatment interventions $\boldsymbol{a}_{\mathrm{u}}$ conditioned on $\boldsymbol{x}_{\mathrm{u}}$, and the achieved treatment outcome y_u. The EHR dataset can be described as $\mathcal{D} = \{(\boldsymbol{x}_u, \boldsymbol{a}_u, y_u)|u = 1, \cdots, N_{\mathcal{D}}\}$. We propose the ADTEP model to address the aforementioned problem. The proposed ADTEP contains seven components: a patient feature encoder $\mathrm{E_x}$, a treatment intervention encoder $\mathrm{E_a}$, a patient feature decoder $\mathrm{G_x}$, a treatment intervention decoder $\mathrm{G_a}$, a treatment intervention generator $\mathrm{G_{xa}}$, a treatment intervention discriminator $\mathrm{D_a}$, and a logistic regression layer for TEP $\mathrm{C_y}$. In detail, given a patient sample $(\boldsymbol{x}, \boldsymbol{a}, y)$, two encoder layers $\mathrm{E_x}$ and $\mathrm{E_a}$ are first employed to extract the latent features $\boldsymbol{h}_x$ and $\boldsymbol{h}_a$ from $\boldsymbol{x}$ and $\boldsymbol{a}$, respectively. The reconstructed features $\boldsymbol{x}'$ and $\boldsymbol{a}'$ can then be estimated from the latent features $\boldsymbol{h}_x$ and $\boldsymbol{h}_a$, using the decoders $\mathrm{G_x}$ and $\mathrm{G_a}$. Note that $\mathrm{E_x}$ and $\mathrm{G_x}$ form an AE for patient feature observations, and for $\mathrm{E_a}$ and $\mathrm{G_a}$ to reconstruct treatment interventions. Both AEs $\mathrm{E_x}$-$\mathrm{G_x}$/$\mathrm{E_a}$-$\mathrm{G_a}$ are adopted to capture robust and discriminative patient feature/treatment representations in the latent feature vector $\boldsymbol{h}_{\mathrm{x}}$/$\boldsymbol{h}_{\mathrm{a}}$. Consequently, the latent feature vectors $\boldsymbol{h}_{\mathrm{x}}$ and $\boldsymbol{h}_{\mathrm{a}}$ are concatenated to form the input of $\mathrm{C_y}$ for TEP.

We measure the reconstruction performance for patient feature $\boldsymbol{x}$ conducted by the encoder $\mathrm{E_x}$ and decoder $\mathrm{G_x}$. For efficient learning of the encoder-decoder, standard practice is to use the Euclidean distance between the input and the generated output to minimize the patient feature reconstruction loss, that is,

$$\mathcal{L}_x = \mathbb{E}_{\boldsymbol{x},\boldsymbol{a},y\sim P_{data}(\boldsymbol{x},\boldsymbol{a},y)} \left|\left|\boldsymbol{x} - G_x\big(E_x(\boldsymbol{x})\big)\right|\right|_2^2. \tag{1}$$

The reconstruction performance for treatment vector $\boldsymbol{a}$ is measured by means of the encoder $\mathrm{E_a}$ and decoder $\mathrm{G_a}$. Similarly to the patient feature reconstruction loss $\mathcal{L}_x$, the treatment reconstruction loss $\mathcal{L}_a$ can be measured as follows:

$$\mathcal{L}_a = \mathbb{E}_{\boldsymbol{x},\boldsymbol{a},y\sim P_{data}(\boldsymbol{x},\boldsymbol{a},y)} \left|\left|\boldsymbol{a} - G_a\left(E_a(\boldsymbol{a})\right)\right|\right|_2^2. \quad (2)$$

To encourage the reconstruction of treatments from discriminative patient features that are similar to real ones, so that the prediction performance can be enriched, we design a treatment discriminator $\mathrm{D_a}$ to differentiate the reconstructed treatment vector $\tilde{a}$ from the true observed treatment a. In particular, we employ a binary classifier to categorize the given input as "real" if the input is the actual treatment vector performed on patients, and "fake" otherwise. The adversarial loss $\mathcal{L}_{GAN}$ is defined as:

$$\mathcal{L}_{GAN} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{a},y\sim p_{data}(\boldsymbol{x},\boldsymbol{a},y)}[log\, D_a(\boldsymbol{a})] + \mathbb{E}_{\boldsymbol{x},\boldsymbol{a},y\sim p_{data}(\boldsymbol{x},\boldsymbol{a},y)}[log\left(1 - D_a\left(G_{xa}\left(E_x(\boldsymbol{x})\right)\right)\right)]. \quad (3)$$

Given a testing patient sample with patient feature vector $\boldsymbol{x}$, treatment vector $\boldsymbol{a}$ conditioned on $\boldsymbol{x}$, and an unknown treatment outcome label y, we can learn the representative and informative features $\boldsymbol{h}_\mathrm{x}$ and $\boldsymbol{h}_\mathrm{a}$ with respect to the patient characteristics, and subsequently the treatments performed on the patient, respectively, and then concatenate these as $[\boldsymbol{h}_\mathrm{x}, \boldsymbol{h}_\mathrm{a}]$ to be fed into the treatment effect predictor $\mathrm{C_y}$. Let y' is the predicted treatment outcome, the loss can be measured using cross-entropy as follows:

$$\mathcal{L}_{pred} = \frac{1}{N_D}\sum_{u=1}^{N_D}(y_u\, log\, y'_u + (1 - y_u)\, log(1 - y'_u)). \quad (4)$$

As demonstrated in the section above, our training is defined by four loss functions: 1) loss of GAN $\mathcal{L}_{GAN}$, loss of patient feature reconstruction $\mathcal{L}_x$, loss of treatment reconstruction $\mathcal{L}_a$, and loss of treatment outcome prediction $\mathcal{L}_{pred}$. In summary, the objective function of the ADTEP is expressed as:

$$\min_{E_x,E_a,G_x,G_a,G_{xa},G_y} \max_{D_a} L_{pred} + \alpha(L_x + L_a) + \beta L_{GAN}, \quad (5)$$

where α and β are trade-off parameters for balancing the importance of the corresponding components.

3 Experiments

We conducted a clinical case study in cooperation with the Cardiology Department of the Chinese PLA General Hospital. The primary investigated major adverse event prediction (MACE) after acute coronary syndrome (ACS). ACS refers to a group of conditions resulting from decreased blood flow in the coronary arteries, whereby that part of the heart muscle is unable to function properly or dies [8]. Regarding the indicators of treatment effects for ACS patient samples, we select the MACE after ACS as the label for treatment effects. To conduct the case study, we collaborated with the clinicians of the cardiology department, and extracted a collection of 3,463 ACS patient samples from the hospital EHR system.

Table 1. Experimental results on ACS experimental dataset

Method	Accuracy	AUC	Precision	Recall	F1 score
LR	0.744±0.016	0.648±0.026	0.505±0.078	0.198±0.034	0.284±0.044
SVM	0.716±0.010	0.621±0.014	0.402±0.032	0.219±0.026	0.283±0.027
DTEP	0.747±0.010	0.653±0.021	0.524±0.056	0.181±0.025	0.268±0.031
ADTEP	0.746±0.012	0.662±0.020	0.515±0.058	0.210±0.036	0.297±0.042

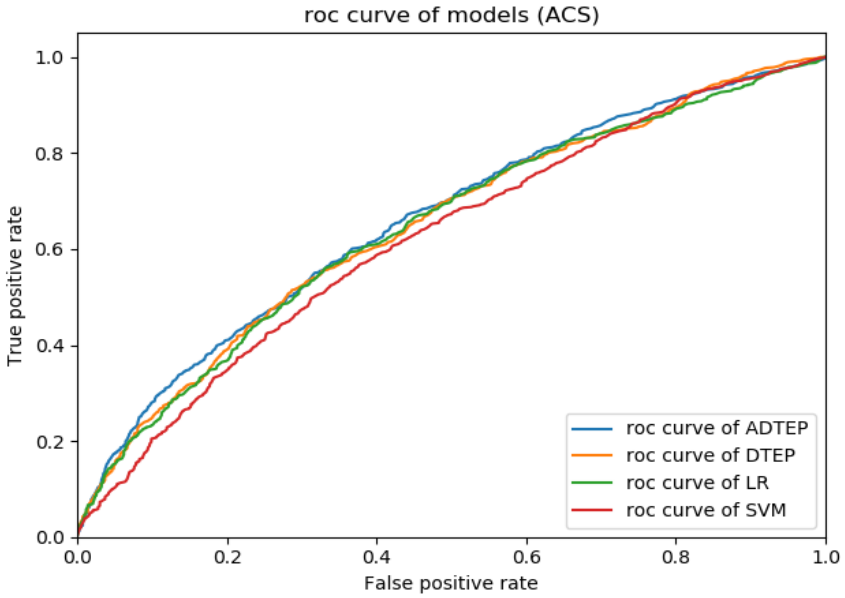

Fig. 1. ROC curves for MACE prediction after ACS

To demonstrate the effectiveness of our proposed model, we compare the proposed ADTEP with the proposed model without adversarial learning, namely the DTEP model. For the DTEP, we use AEs to generate the latent representations of both the patient characteristics and the subsequent treatments, concatenate the derived latent features, and then feed the obtained feature vector into a logistic regression layer, yielding a TEP model. Moreover, we compare the proposed model to state-of-the-art models using the experimental datasets, including logistic regression (LR) and the support vector machine (SVM).

The performance was evaluated by the Area Under the receiver operating characteristic (ROC) curve (AUC), accuracy, precision, recall and F1 score. We repeated the experiments five times to validate the performance of each model on the experimental dataset. As a result, we obtained a group of experimental results for each model, on which the mean value and confidence intervals were calculated.

Table 1 presents the TEP performance achieved on the experimental ACS dataset. As can be observed from Table 1, the proposed model achieved superior performance compared to benchmark models on the experimental dataset. ADTEP performed slightly better than DTEP in terms of both the AUC and F1. Although DTEP outperformed ADTEP in terms of the average accuracy, the performance gain was marginal. These findings indicate that the incorporation of correlational information between patient characteristics and treatments by means of the adversarial learning strategy was useful in predicting the treatment effects of ACS patient samples. Figure 1 illustrates the ROC curves for MACE prediction after ACS, also demonstrating that the proposed ADTEP achieved comparative performance with benchmark models. In particular, ADTEP exhibited 1.4%, 2.2%, and 6.3% performance gains for MACE prediction in terms of AUC over DTEP, LR, and SVM, respectively.

4 Conclusions

In this work, we have addressed quite a challenging problem in medical informatics, namely utilizing a large volume of observational data for TEP. Our proposed model was evaluated on a real clinical dataset, and the experimental results demonstrate significant improvements in TEP compared to state-of-the-art methods.

Acknowledgments

This work was partially supported by the National Key Research and Development Program of China under Grant No. 2016YFC1300303 and the National Nature Science Foundation of China under Grant No. 61672450.

References

1. Concato, J., Shah, N., Horwitz, R. I.: Randomized, controlled trials, observational studies, and the hierarchy of research designs. New England journal of medicine 342(25), 1887-1892 (2000).
2. 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).
3. Cartwright, N., Munro, E.: The limitations of randomized controlled trials in predicting effectiveness. Journal of evaluation in clinical practice 16(2), 260-266 (2010).
4. Xiao, C., Choi, E., Sun, J.: Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. JAMIA 25(10), 1419-1428 (2018).
5. Shalit, U., Johansson, F. D., Sontag, D.: Estimating individual treatment effect: generalization bounds and algorithms. In: Proc. 34th Int. Conf. Mach. Learn., vol. 70, pp. 3076–3085, (2017).
6. Yoon, J., Jordon, J., van der Schaar, M.: GANITE: Estimation of individualized treatment effects using generative adversarial nets. In: Int. Conf. Learning Representations (2018).
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., and et al.: Generative Adversarial Networks. arXiv:1406.2661 (2014).
8. Huang, Z., Dong, W.: Adversarial MACE Prediction after Acute Coronary Syndrome using Electronic Health Records. IEEE Journal of Biomedical and Health Informatics (2018).