Medical Outcome Prediction by Adaptive Knowledge Distillation

Xu Min[1,†], **Yiqin Yu**[2,†], **Shiwan Zhao** [2], **Jing Mei** [2], **Shaochun Li** [2]

[1]IBM Cognitive Asset Engineering Services, Beijing, China
[2]IBM Research China, Beijing, China
{minxux, yuyiqin, zhaosw, meijing, lishaoc}@cn.ibm.com
[†]*Contributed equally to this work*

Abstract

Medical outcome prediction is an important task in healthcare scenarios, such as in-hospital mortality prediction. There is an increasing demand of adapting such prediction models from one domain to another one. However, the model adaptation becomes challenging when the source domain data is absent due to patient privacy and security concerns. To address this problem, we propose a method named Adaptive Knowledge Distillation (AKD). In detail, we use knowledge distillation to extract useful information from a source model which is built on source domain data. Specifically, to adaptively control how much knowledge should be injected, a dynamic strategy of the imitation parameter is designed. Our results on MIMIC-III data demonstrate the effectiveness of AKD for the adaptations of medical outcome prediction models.

Introduction

Along with the accumulation of digital healthcare data, machine learning algorithms have been widely used to build numerous models to generate insights for disease prevention, diagnosis, treatments and prognosis. In particular, medical outcome prediction is one of the most important tasks (Miotto et al. 2017), such as in-hospital mortality prediction. Meanwhile there is a growing demand of adapting such prediction models built on one domain to another new one.

Domain Adaptation (DA) is an important methodology to leverage information in one or more related *source* domains to another *target* domain. However, sharing of healthcare datasets which were used to build these models is highly restricted under most circumstances. When healthcare institutes try to build a new prediction model, it is often impossible for them to get the privilege of accessing the data of other published source models during adaptation.

Knowledge distillation (KD) (Hinton, Vinyals, and Dean 2014) has been used to extract knowledge from the source model when the source domain data is absent. Nevertheless, KD is originally used to build smaller and faster models by "compressing" large and complex models. Therefore, when applying KD in domain adaptation tasks, the concern about how much knowledge should be distilled from the source model should be reconsidered carefully.

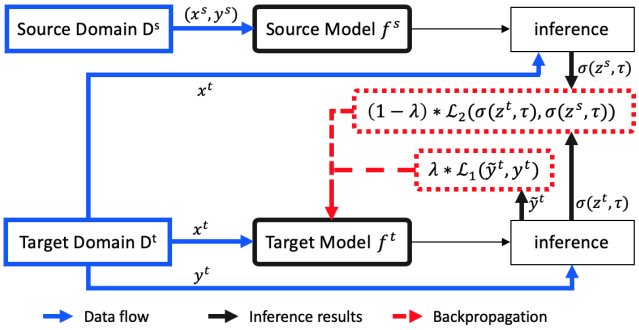

Figure 1: The AKD framework. The back-propagation algorithm considers losses from both of the hard labels ($\mathcal{L}_1$) and the soft labels ($\mathcal{L}_2$) on the target data.

In this paper, we address the domain adaptation problem in cases where the source domain data is absent and the target domain data is relatively small. We propose a novel model-agnostic approach named Adaptive Knowledge Distillation (AKD) to adaptively distill knowledge from the source model which will be described in detail below.

Adaptive Knowledge Distillation

A domain D usually contains an input space X and an output space Y. Formally, DA is defined as a mechanism of training a model f^t on the target domain $D^t = (X^t, Y^t)$ by adapting from the source domain $D^s = (X^s, Y^s)$.

Here, we use KD to address the DA problem where 1) the source model f^s is available while the source domain data (X^s, Y^s) is absent, 2) the size of target domain data (X^t, Y^t) is relatively small. We propose an AKD framework as shown in Figure 1. Basically, the overall loss function $\mathcal{L}(\theta)$ of the target model is calculated as:

$$\mathcal{L}(\theta) = \lambda * \mathcal{L}_1(\tilde{y}^t, y^t) + (1-\lambda) * \mathcal{L}_2(\sigma(z^t, \tau), \sigma(z^s, \tau)), \quad (1)$$

where θ are model parameters. On one hand, $\mathcal{L}_1$ is the cross-entropy loss computed on the hard label $\tilde{y}^t$ predicted by the target model and the corresponding ground-truth label y^t. On the other hand, $\mathcal{L}_2$ is the cross-entropy loss comparing

the soft labels $\sigma(z^t, \tau)$ predicted by the target model against the soft labels $\sigma(z^s, \tau)$ generated by the source model. λ is an imitation parameter, σ is a softmax function parameterized by temperature τ, and z^t and z^s are the logits of the target and source model respectively.

Specifically, in medical outcome prediction tasks, the first part of the overall loss $\mathcal{L}_1$ relates to medical data information in the target domain, whereas the second part $\mathcal{L}_2$ relates to the medical knowledge encoded in the source prediction model. Here the imitation parameter λ controls how much proportion of knowledge distilled from the source model will be injected into the training process of the target prediction model. Intuitively, we hope to grasp the information of the probability vector generated by the source model, when this vector is consistent with the ground-truth label of the target data. Practically, we propose a dynamic strategy for λ which can be adaptive according to the target data as following:

$$\lambda = \lambda_0 + \delta \cdot \mathbb{1}(y^t \neq \arg\max \sigma(z^s)), \tag{2}$$

where $0 \leq \lambda_0 \leq 1, 0 \leq \delta \leq 1$, and $0 \leq \lambda_0 + \delta \leq 1$.

To explain more clearly, the value of λ will be close to 0 when $\sigma(z^s)$ is consistent with y^t, which means we learn more from the source model for such samples. Otherwise, when the $\sigma(z^s)$ and y^t are inconsistent, the weight will be close to 1, which means we put more attention on the target data itself instead of the source model. Through this dynamic mechanism, we force the target model to mimic the source model when it generates correct prediction on target data. The adaptive choice of λ in distilling knowledge is intrinsically to improve the generalization ability of target model with the help of source models.

Medical Outcome Prediction Experiments

We evaluate our AKD approach on MIMIC-III benchmarks (Harutyunyan et al. 2019) to demonstrate its potential in real-world healthcare scenarios. All experiments are implemented in TensorFlow and run with NVIDIA Tesla P100 GPUs. Based on MIMIC-III (Johnson et al. 2016), a freely accessible critical care database, Harutyunyan et al. (2019) constructed benchmark machine learning datasets. In particular, they defined an "In-Hospital Mortality" (IHM) task as predicting whether an ICU patient will die at discharge given the first 48 hours observation of the ICU stay.

For medical outcome prediction experiments, we first split MIMIC-III data into two non-overlapping domains according to different types of ICU admission. As shown in Figure 2, the source domain contains patients with admission type "Emergency/Urgent", while the target domain contains patients with admission type "Elective". The source domain has 11800, 2616 and 2524 samples in the training, validation and test set, respectively. Meanwhile the target domain has 1417, 319, and 356 samples in the training, validation and test set, respectively. Each sample has 48 timestamps of 76 features, with a label indicating mortality at discharge.

There exists significant class imbalance in this IHM task. In detail, the mortality rate in the source domain is 15.15% and that in the target domain is only 3.63%. Therefore with regard to model evaluation, auPRC (area under the Precision Recall Curve) is a more convincing measurement than accuracy and auROC (area under the Receiver Operating Characteristic curve). To keep consistence with Harutyunyan et al. (2019), we adopt bidirectional RNN (Recurrent Neural Networks) with LSTM (Long-Short Term Memory) units to implement all networks.

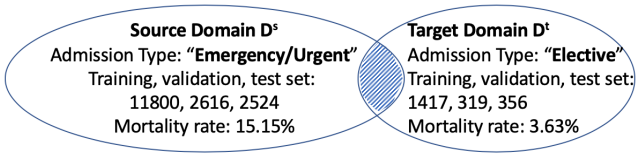

Figure 2: MIMIC-III IHM datasets split into source and target domains. Those patients with both emergency and elective admissions are removed (the slashed area).

Table 1: Performance on MIMIC-III in IHM task.

-	test_acc	auROC	auPRC
teacher on source	0.8843	0.8314	0.4936
teacher on target	0.9578	0.8650	0.2482
baseline	0.9494	0.8633	0.2014
KD ($\lambda = 0.5$)	0.9550	0.8658	0.2336
AKD ($\lambda_0 = 0, \delta = 0.9$)	0.9550	0.8304	**0.2647**

We report our results in Table 1, and find that: 1) the source model trained on source domain generates good performance on the source test set, with auPRC of 0.4936. 2) The source model achieves reasonable performance on the target test set, with auPRC of 0.2482. 3) If we train a model only using a small size of target data, the classification performance is worse with auPRC of only 0.2014. 4) If we train a new model on the target data with the help of the source model using original KD, the auPRC is 0.2336 between the previous two results. 5) The proposed AKD method achieves an auPRC of 0.2647 which is better than only using the source model or training from scratch.

In conclusion, our result demonstrates that our method AKD can effectively facilitate healthcare outcome prediction tasks by adpatively leveraging source domain models.

References

Harutyunyan, H.; Khachatrian, H.; Kale, D. C.; Ver Steeg, G.; and Galstyan, A. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific data* 6(1):96.

Hinton, G.; Vinyals, O.; and Dean, J. 2014. Distilling the knowledge in a neural network. In *NIPS 2014 Deep Learning and Representation Learning Workshop*.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Li-wei, H. L.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3:160035.

Miotto, R.; Wang, F.; Wang, S.; Jiang, X.; and Dudley, J. T. 2017. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics* 19(6):1236–1246.