A Transfer Learning Method for Multi-Center Prognostic Prediction Analysis *

Kota Matsui[1], **Kenta Kanamori**[2], **Wataru Kumagai**[3], **Mitsuaki Nishikimi**[4] and **Shigeyuki Matsui**[1]
[1] Department of Biostatistics, Nagoya University Graduate School of Medicine, Japan
[2] Graduate School of Engineering, Nagoya Institute of Technology, Japan
[3] RIKEN Center for Advanced Intelligence Project, Japan
[4] Department of Emergency and Critical Care, Nagoya University Graduate School of Medicine, Japan
{matsui.k, smatsui}@med.nagoya-u.ac.jp, kanamori.k.mllab.nit@gmail.com, wataru.kumagai@riken.jp

Abstract

We present a transfer learning method for the prognostic prediction analysis using multi-center data. The proposed method improves prediction accuracy by explicitly incorporating varying background-information distributions across multiple center into the model, in multi-center observational study. Experimental results show the effectiveness of the proposed method.

1 Introduction

Prognostic analysis of disease, which is useful for risk stratification and treatment selection of patients, has been frequently carried out as multi-center studies. In a multi-center study, a relatively large number of patients can be collected in a short period of time, and it is effective to expand the applicability or generalizability of the developed predictor through accommodating multiple centers with diverse characteristics. However, the analysis of multiple centers at the same time can raise a difficulty in prognostic prediction since the background information of patients can considerably vary across multiple centers. Generally, patients in base hospitals tend to be more serious, compared with those in (relatively small) municipal hospitals.

In this study, we will see that the multi-center prognostic analysis can be formulated as a *transfer learning under the covariate shift* (Shimodaira [2000]; Sugiyama *et al.* [2012]). Specifically, by explicitly incorporating information on the within-center covariate distributions into the learning phase, we propose a novel framework for prognostic prediction model to enhance the prediction performance for individual patients in a particular target center. This is contrast to the standard framework of prognostic prediction, where a common predictor is being used for patients in the target center without regard to the center's characteristics.

2 Problem Formulation

Multi-center prognostic prediction at $K+1$ centers is often formulated as binary classification problems. Let $\boldsymbol{x}_{i_k} = (x_{i_k 1}, ..., x_{i_k d}) \in \mathcal{X}_k$ ($i_k = 1, ..., n_k$, $k = 1, ..., K$) be the covariate vector of i_k-th patient in k-th center and $y_{i_k} \in \mathcal{Y} = \{1, -1\}$ be the binary outcome, that is $y_{i_k} = 1$ represents the bad prognosis of i_k-th patient and $y_{i_k} = -1$ represents good prognosis. Then we want to develop a prediction model $f : \mathcal{X}_t \to \mathcal{Y}$ using the data $\{(\boldsymbol{x}_{i_k}, y_{i_k})\}$ from K source domains that predicts good or bad prognosis of a new patient in a target center $\mathcal{X}_t$. This is an ordinary binary classification problem and there are many approaches to develop f. In such study, logistic regression is often used due to interpretability, that uses the linear score function $g_{\boldsymbol{\theta},b}(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{x} + b$ and minimizes the logistic loss $\ell(g_{\boldsymbol{\theta},b}(X), Y) = \log(1 + \exp(-Y g_{\boldsymbol{\theta},b}(X)))$. Then the prognosis of a new patient is probabilistically predicted with some estimator $\hat{\boldsymbol{\theta}}, \hat{b}$ as follows: $f(\boldsymbol{x}) = 1$ if $\frac{1}{1+\exp\{-g_{\hat{\boldsymbol{\theta}},\hat{b}}\}} > 0.5$ and $f(\boldsymbol{x}) = -1$ otherwise. In tradition, parameter estimation is executed via empirical risk minimization

$$\hat{\boldsymbol{\theta}}, \hat{b} = \underset{\boldsymbol{\theta},b}{\arg\min} \frac{1}{\sum_{k=1}^K n_k} \sum_{k=1}^K \sum_{i_k=1}^{n_k} \ell(g_{\boldsymbol{\theta},b}(\boldsymbol{x}_{i_k}), y_{i_k}). \quad (1)$$

However $\hat{\boldsymbol{\theta}}, \hat{b}$ estimated by (1) may perform poorly due to *covariate shift*, that we describe in detail in the following.

Covariate Shift. Let $p_k(\boldsymbol{x}, y)$ and $p_t(\boldsymbol{x}, y)$ be joint distribution over $\mathcal{X} \times \mathcal{Y}$ at k-th source center and that of target center. Then, we can write the marginal distribution of each center as $p_k(\boldsymbol{x})$, $p_t(\boldsymbol{x})$ and the conditional distribution of each center as $p_k(y \mid \boldsymbol{x})$, $p_t(y \mid \boldsymbol{x})$ respectively. Under the above notation, the covariate shift can be formulated as follows : for $1 \le k \le K$ and $k \ne k'$,

$$\begin{gathered} p_k(\boldsymbol{x}) \ne p_{k'}(\boldsymbol{x}),\ p_k(\boldsymbol{x}) \ne p_t(\boldsymbol{x}), \\ p_k(y \mid \boldsymbol{x}) = p_{k'}(y \mid \boldsymbol{x}),\ p_k(y \mid \boldsymbol{x}) = p_t(y \mid \boldsymbol{x}). \end{gathered} \quad (2)$$

This is a reasonable assumption in prognostic prediction analysis. For example, considering a prognostic study in which large hospitals and community hospitals coexist, it is easy to imagine the situation in which serious patients concentrate in the former and mild patients concentrate in the latter (this situation is corresponding to the first two condition in (2)). It would also be reasonable to imagine that patients who have similar clinical information follow a similar prognosis (this is corresponding to the last two condition in (2)). Since covariate shifts can significantly deteriorate the performance of the prediction model (1) at the target center (Sugiyama *et al.* [2012]), it is necessary to modify it adequately.

*This research was supported by JST CREST, grant number JPMJCR1412.

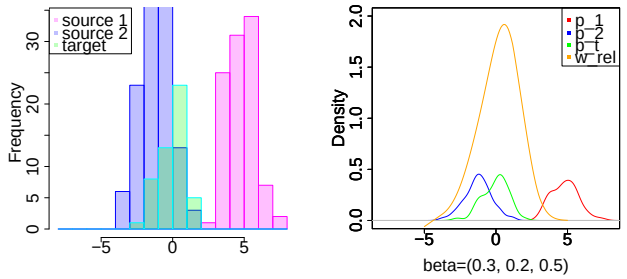

Figure 1: left panel : two source data (blue and red) and one target data (green). right panel : $p_1(\boldsymbol{x})$ (red), $p_2(\boldsymbol{x})$ (blue), $p_t(\boldsymbol{x})$ (green) and estimated $w_{rel}(\boldsymbol{x};\boldsymbol{\beta})$(orange).

Table 1: Synthetic data experiments (30 trials).

method	no importance	with importance
average accuracy	0.626	**0.877**

3 Method

At first, we introduce generalized importance for multi- center study. Inspired from the *relative importance* proposed by Yamada *et al.* [2013], we define the *multi-center relative importance* as follows:

$$w_{rel}(\boldsymbol{x};\boldsymbol{\beta}) = \frac{p_t(\boldsymbol{x})}{\beta p_t(\boldsymbol{x}) + \sum_{k=1}^{K} \beta_k p_k(\boldsymbol{x})},$$

where $\boldsymbol{\beta} = (\beta, \beta_1, ..., \beta_K)$ are hyperparameters such that $\beta + \sum_{k=1}^{K} \beta_k = 1$. Especially, β_k, $k = 1, ..., K$ can be interpreted as an index to what extent the source center k contributes the performance of prognostic model in the target center. Figure 1 illustrates the estimation of $w_{rel}(\boldsymbol{x};\boldsymbol{\beta})$. Then our learning problem is formulated as an weighted empirical risk minimization

$$\min_{\boldsymbol{\theta},b} \frac{1}{\sum_{k=1}^{K} n_k} \sum_{k=1}^{K} \sum_{i_k=1}^{n_k} w_{rel}(\boldsymbol{x}_{i_k};\boldsymbol{\beta}) \ell(g_{\boldsymbol{\theta},b}(\boldsymbol{x}_{i_k}), y_{i_k}). \quad (3)$$

The learning algorithm is roughly divided into the following two steps.

Step 1 Estimate $w_{rel}(\boldsymbol{x};\boldsymbol{\beta})$ from $\{\boldsymbol{x}_{i_k}\}_{i_k=1}^{n_k}$, and $\{\boldsymbol{x}_{i_t}\}_{i_t=1}^{n_t}$ by kernel density ratio estimation technique.

Step 2 Solve (3) using estimated importance to get $\hat{\boldsymbol{\theta}}$ and $\hat{b}$.

Furthermore, our algorithm includes the optimization of $\boldsymbol{\beta}$ by an appropriate method such as cross validation.

Through the proposed method, we can learn highly accurate prediction models even if there is no label information in the data of the target domain thanks to covariate shift assumption.

4 Experimental Results and Discussion

We conducted two experiments (one is by synthetic data and another is by real data).

Synthetic Experiment. We consider the domains (one target, three sources) with two dimensional input space. We sample the target data from the normal distribution with mean $\mu_t \sim N(\boldsymbol{0}, I_2)$. Similarly, the source data are sampled from normal distribution with mean $\mu_1 \sim N(\boldsymbol{2}, I_2)$, $\mu_2 \sim N(-\boldsymbol{2}, I_2)$ and $\mu_3 \sim N(\boldsymbol{10}, I_2)$ respectively. Here, I_2 is the identity matrix and each μ_k is interpreted as a parameter representing each center. We generate 150 target data and 200 of each source data, and the class label is generated using true logistic model. We compared proposed method (3) to conventional method (1). The results are summarized in Table 1. In this example, our proposed method showed the better result.

Table 2: Real data experiments.

method	no importance	with importance
accuracy	0.727	**0.909**

Table 3: #centers vs accuracy (30 trials).

#centers	4	5	6
average accuracy	0.877	0.963	0.977

Real Data Analysis. We evaluate the prediction accuracy of the proposed method using clinical research data of prognostic predictions conducted in Nishikimi *et al.* [2017]. The data was acquired from four hospitals (two large hospitals and two city hospitals) and is composed of clinical information of 15 items from 152 patients. Our goal is developing a predictive model that well performs in the target hospital from the source data. The results are summarized in Table 2. We can see the significant improvement in accuracy of our method compared to the case without importance weighting.

Discussion. We observe that our proposed method greatly improved the accuracy of the prognostic prediction. Finally, consider the relationship between the number of centers and the accuracy. Intuitively, it can be expected that the accuracy improves as the number of centers increases. We can observe this tendency by the experimental result with two dimensional synthetic data shown in Table 3.

References

Mitsuaki Nishikimi, Naoyuki Matsuda, Kota Matsui, et al. A novel scoring system for predicting the neurologic prognosis prior to the initiation of induced hypothermia in cases of post-cardiac arrest syndrome: the cast score. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 25(1):49, 2017.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural computation*, 25(5):1324–1370, 2013.