Boosting Transfer Learning with Survival Data from Heterogeneous Domains

ALEXIS BELLOT AND MIHAELA VAN DER SCHAAR,
ALAN TURING INSTITUTE, UNIVERSITY OF CAMBRIDGE

INTRODUCTION

Motivation:

- ▷ Most models do **not** generalize to populations outside the marginal and conditional distributions observed in training.
- ▷ In many applications **small samples** do not allow training of flexible models.
- ▷ Problem is exacerbated in **survival analysis** - *censoring*.

Example: NHS hospitals in the UK have on average 89 beds per hospital- which for rare diseases can lead to fewer than 50 registered outcomes of interest per year.

Objective:

Can we leverage data from other domains to improve time to event predictions?

$$S(t|\boldsymbol{x}_i) = \mathbb{P}(T_i > t|\boldsymbol{x}_i) \quad (1)$$

PROBLEM SET-UP AND CHALLENGES

Data:

$\mathcal{D}_{ta} := (X, \delta T + (1-\delta)C, \delta)$ from a domain $\mathcal{X}_{ta} \times \mathcal{T}_{ta} \times \{0,1\}$ with joint probability p_{ta}

$\mathcal{D}_{so} := (X, \delta T + (1-\delta)C, \delta)$ from a domain $\mathcal{X}_{so} \times \mathcal{T}_{so} \times \{0,1\}$ with joint probability p_{so}

1) Feature mismatch Differing domains will often result in heterogeneous feature spaces - $\mathcal{X}_{so} \neq \mathcal{X}_{ta}$. The same variables will not be observed everywhere.

2) Distribution mismatch: Marginal - $p_{so}(X) \neq p_{ta}(X)$ - and conditional distributions - $p_{so}(T|X) \neq p_{ta}(T|X)$ - may differ between domains.

Solution:

- ▷ Propose a iterative non-parametric approach to avoid directly modelling distributional differences, and uses variable **missingness itself** if predictive for the target population.
- ▷ Carefully modify data distribution such as to **(1)** leverage *most similar patients from auxiliary domains*, and **(2)** *encourage good predictions in target domain.*

Objective: Learn model that implicitly corrects for distribution shift, without distributional assumptions, guided by in-sample prediction performance.

OUR APPROACH: TRANSFER LEARNING THROUGH BOOSTING

Step 1: ***Define weak learners***

• Recursively partition population into homogeneous subgroups - like trees with leaves and nodes.

- ▷ **Nodes** Use the likelihood ratio statistic as a homogeneity measure. Expand splitting rules to *include missingness of a variable itself as a valid partition*.
- ▷ **Leaves** Predictions in each leaf j of the resulting tree are made with the Kaplan-Meier estimator.

$$\hat{h}_j(t) = \prod_{i \in \mathcal{C}_j : t_i \leq t} \left(1 - \frac{N_j(t_i)}{Y_j(t_i)}\right)$$

Step 2: ***Define a measure of patient similarity***

• Source instances *similar* to target if predictions agree. New notion of prediction correctness for survival with consistent estimate e_i of,

$$\frac{1}{\tau}\int_0^\tau \mathbb{E}_{(T,X)\sim p_{ta}}\left[\left(I(T>t) - \hat{h}(t;X)\right)^2\right] dt$$

Step 3: ***Ensemble learning***

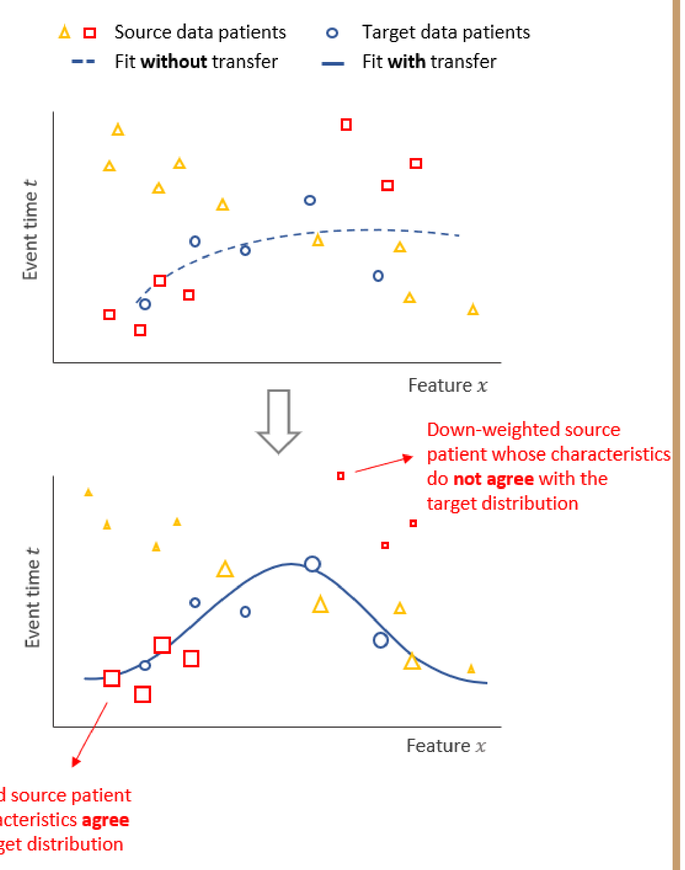
• Iteratively modify the patient distribution,

$$w_i^{(m+1)} \propto \begin{cases} w_i^{(m)} (\beta_T^{(m)})^{-e_i^{(m)}}, & i \in \mathcal{D}_{ta} \\ w_i^{(m)} (\beta_S^{(m)})^{e_i^{(m)}}, & i \in \mathcal{D}_{so} \end{cases}$$

• Confidence attributed to weak learner m,

$$\beta_T^{(m)} = \frac{\epsilon^{(m)}}{1-\epsilon^{(m)}}$$

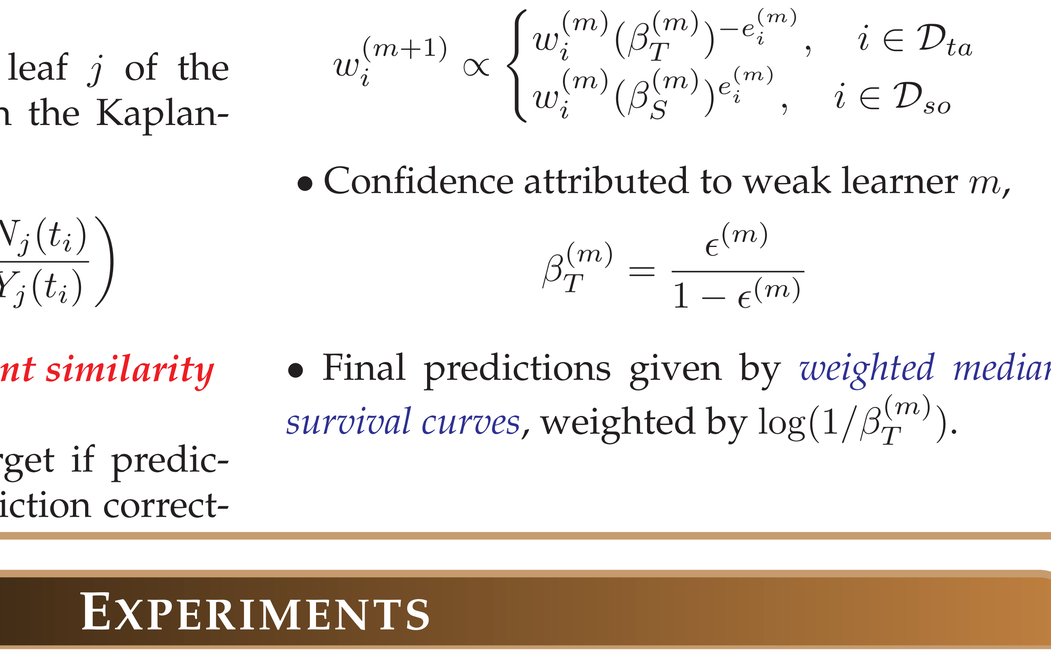
• Final predictions given by *weighted median survival curves*, weighted by $\log(1/\beta_T^{(m)})$.



EXPERIMENTS

Synthetic Simulations:

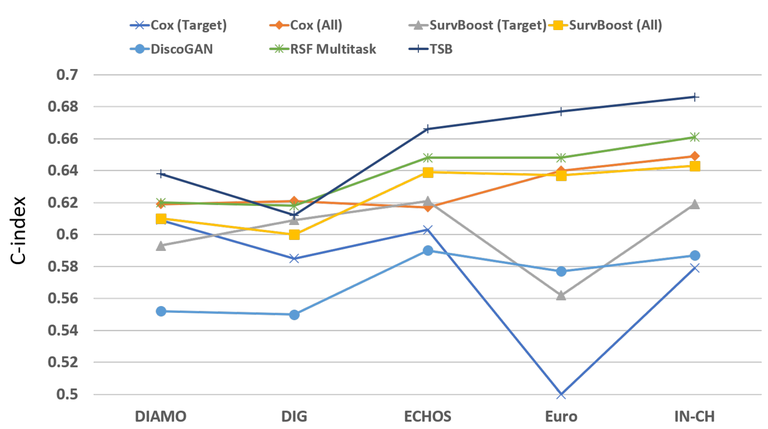
- ▷ Vary marginal source and target distributions
- ▷ Vary conditional source and target distributions



Real data: MAGGIC data from 30 different medical studies containing patients that experienced heart failure.

- ▷ *Can we learn from all studies to target a specific one?*

Benchmarks: Cox, Survival Boosting, trained on "target only" and "both source and target" data, multitask random forest and a GAN-based algorithm for domain translation.



CONCLUSIONS

Transfer learning is successful for moderate discrepancies in joint distributions.

Technical Significance:

- ▷ Convergence guarantees analogous to original Adaboost method.
- ▷ No distributional assumptions on the underlying populations.
- ▷ Splitting rules of trees is expanded to include "missingness" of a variable itself as a valid splitting rule.

Medical Significance:

- ▷ Clinicians can benefit from accurate decision support mechanisms even when the target population at the local hospital is small.