Sampled-Boosting Regression Transfer for Atmospheric Pollution Prediction

Shrey Gupta[1], Jianzaho Bi[3], Yang Liu[2], Avani WIldani[1]
[1](Department of Computer Science, [2]Deparment of Environmental Health), Emory University
[3]Department of Environmental and Occupational Health Sciences, University of Washington

INTRODUCTION

MOTIVATION

- Prediction of atmospheric pollution (PM 2.5) requires the installation of costly equipment.
- Developing countries lack investment in equipment and suffer from data-deficiency.
- **Knowledge Transfer Methodologies (Transfer Learning):** Utilize data from data-rich regions and adapt it for prediction-modeling for data-scarce regions.

GOAL

- Improve current Instance Transfer Learning (ITL) methodologies that suffer from overfitting and are domain-specific for real-world datasets.
- **Cross-domain Collaboration [AI/ML + Environmental Science]:** To use classical machine learning algorithms for better interpretability for domain experts.

METHODOLOGY

- Sampling.TBoost is a successor for TrAdaBoost.R2 [1].
- We use Importance Sampling to get source domain samples most similar to target domain samples. We use Variance Sampling on target domain samples.
- We employ AdaBoost.R2 instead of AdaBoost.R2' as it reduces the generalizability of the model.
  - AdaBoost.R2': Modified version of AdaBoost.R2 where the weights of source instances are frozen whereas the weights of target instances are updated (focussed domain-adaptation).
- The weights of the training instances are updated as:

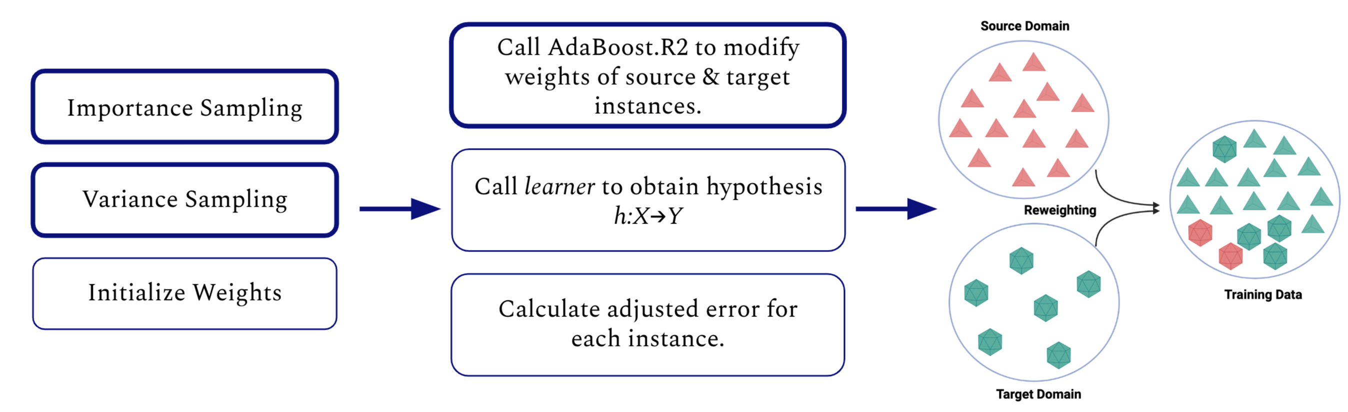

Fig 2: Pipeline showing different stages of Sampling.TBoost

$$w_i^{t+1} = \begin{cases} \frac{w_i^t \beta_t^{e_i^t} \alpha}{Z_t}, & 1 \le i \le p \\ \frac{w_i^t \beta_t^{1-e_i^t} \alpha}{Z_t}, & p \le i \le (p+q) \end{cases}$$

where β , e and Z are previously defined.
α is the fixed learning rate chosen as 0.1.
The no. of source instances are p and the no. of target instances are q.

RELEVANT CONCEPTS

AdaBoost

Adaptive Boosting is an ensemble methodology that sequentially combines (over N chosen iterations) a set of weak learners to generate a strong learner.

AdaBoost.R2 (Adaptive Boosting for Regression)

Uses *adjusted error:*

$$e_i' = \frac{e_i}{max_{i=1}^{n}|e_i|} \quad (1)$$

$$\text{where,} \quad e_i = |y(x_i) - h(x_i)| \quad (2)$$

where e_i denotes the predicted error on the hypothesis h_t and i are the number of training instances.

Weight update takes place as:

$$w_i^{t+1} = \frac{w_i^t \beta_t^{1-e_i'^t}}{Z_t} \quad (3)$$

$$\text{where,} \quad \beta_t = \eta_t / 1 - \eta_t \quad \text{and} \quad \eta_t = \sum_{k=1}^{n} w_i^t e_i^t \quad (4)$$

where Z_t is normalizing constant, t is current iteration.

TrAdaBoost = Transfer Learning + AdaBoost

TrAdaBoost.R2 = Transfer Learning + AdaBoost.R2

Importance Sampling

Choosing samples to train upon by measuring the importance of the instances for prediction. Techniques used:

1. L_1/L_2 Norm.
2. Similarity Measure.

Variance Sampling (using k-Center Sampling)

Introducing noise (source samples) in the target dataset to increase its variability.

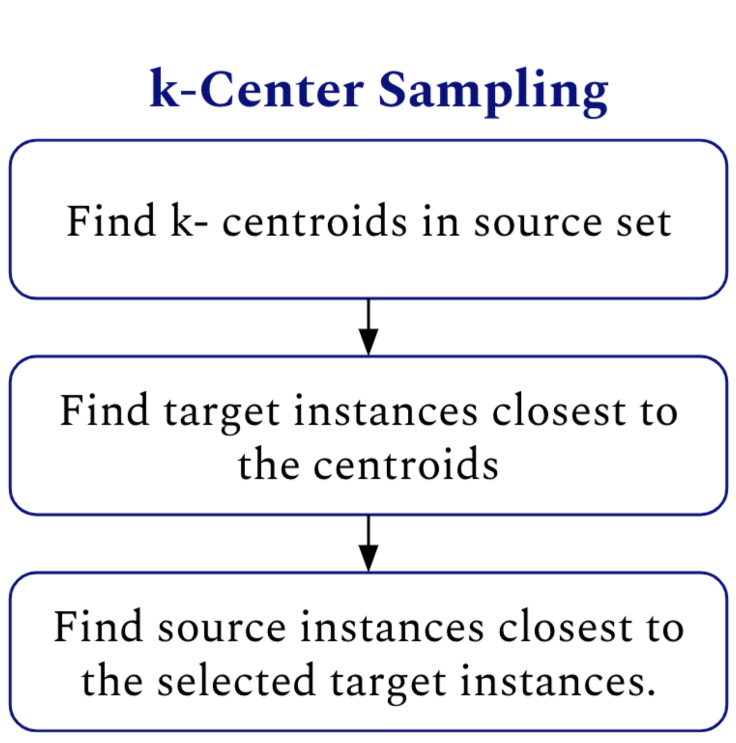

Fig. 1: Flow-chart for k-Center sampling employed for Variance Sampling in Sampling.TBoost.

RESULTS

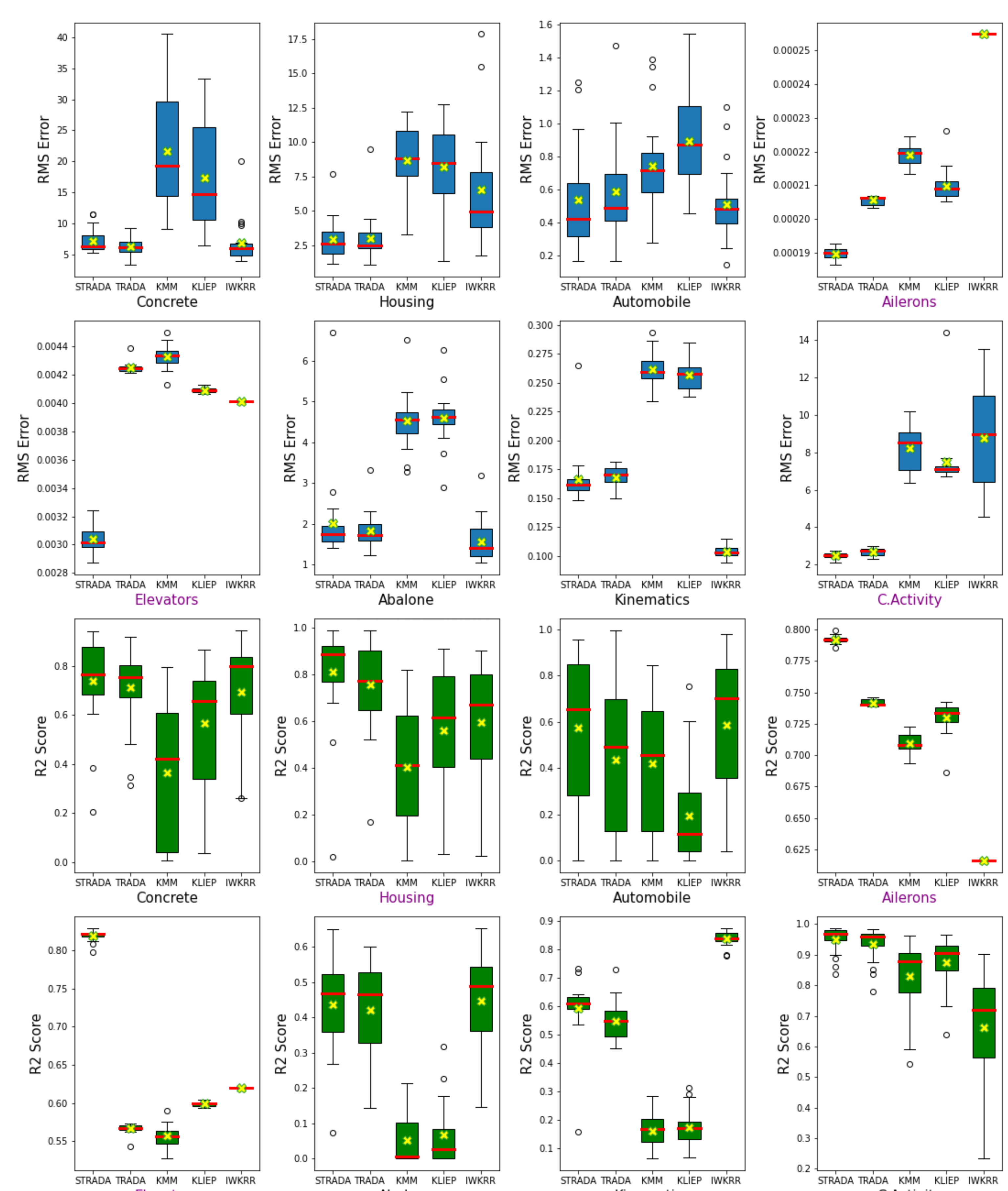

Fig 3: Comparison of transfer learning algorithms– TRADA: TrAdaBoost, STRADA: Sampling.TBoost, KMM: Kernel Mean Matching, and KLIEP: Kullback-Leibler Importance Estimation, IWKRR: Importance Weighted-Kernel Ridge Regression. The Interquartile Range (IQR), mean value (marker: yellow "X"),and median value (marker: red line) for each algorithm over the iterations have been highlighted. The datasets for which Sampling.TBoost performs particularly well are marked (marker: purple).

DATASET

- We chose 8 regression datasets from the UCI machine learning repository [2] as shown in Fig 3.
- The datasets were divided into source, target, and test sets using the splitting methodology used by Pardoe et al. [1].
- Splitting Methodology [Conceptual Split]:
  - Identifying moderately correlated feature (F_M) with the target variable.
  - Split into source-target based on the range of values of F_M.
- **Simulated a real-world Transfer Learning Problem:**

$Size_{Target}$ <<< $Size_{Source}$

ANALYSIS

- Sampling.TBoost consistently performs well -- low RMSE and **high R-squared score.**
- Methodologies like IW-KRR.TL and TTR2 sometimes outperform Sampling.TBoost but fluctuate highly in their performance.
- **TTR2 is the baseline algorithm** for this study.
- Sampling.TBoost outperforms TTR2 on:
  - 5/8 datasets for Root Mean Squared Error.
  - 8/8 datasets for R-squared Score.

CONCLUSION

- We introduce Sampling.TBoost, a complexity-tolerant, domain-agnostic, boosting-based transfer learning algorithm.
- Sampling.TBoost uses Importance Sampling and unconstrained weight update strategy to outperform competitive transfer learning methodologies.
- Sampling.TBoost improves the average performance by 12% across all diverse distribution regression datasets.
- The changes we propose to TrAdaBoost.R2 are modest enough to function as a succeful replacement.

ACKNOWLEDGEMENT

This research is funded by Laney Graduate School and Rollins Department of Public Health at Emory University. The work of Y. Liu was partially supported by the National Institute of Environmental Health Sciences of the National Institute of Health.

REFERENCES

1. Pardoe D, Stone P (2010) Boosting for regression transfer. In: Proceedings of the 27thInternational Conference on Machine Learning, pp 863–870.
2. Asuncion A, Newman D (2007) Uci machine learning repository.

Created with BioRender Poster Builder