Title: Heterogeneous Statistical Transfer Learning

Speaker: Prof. Subhadeep Paul, The Ohio State University

Area: DS

Date: 16.06.2026, Venue: P22 @ 2.00PM

Abstract:

In the first part of the talk, we consider the problem of Transfer Learning (TL) under heterogeneity from a source to a new target domain for high-dimensional regression with differing feature sets. Most homogeneous TL methods assume that target and source domains share the same feature space, which limits their practical applicability. In applications, the target and source features are frequently different due to the inability to measure certain variables in data-poor target environments. Conversely, existing heterogeneous TL methods do not provide statistical error guarantees, limiting their utility for scientific discovery. Our method first learns a feature map between the missing and observed features, leveraging the vast source data, and then imputes the missing features in the target. Using the combined matched and imputed features, we then perform a two-step transfer learning for penalized regression. We develop upper bounds on estimation and prediction errors, assuming that the source and target parameters differ sparsely but without assuming sparsity in the target model. We obtain results for both when the feature map is linear and when it is nonparametrically specified as unknown functions. Our results elucidate how estimation and prediction errors of HTL depend on the model's complexity, sample size, the quality and differences in feature maps, and differences in the models across domains.

In the second part of the talk, going beyond linear models, I will discuss a transfer learning method for nonparametric regression using a random forest. The unknown source and target regression functions are assumed to differ for a small number of features. Our method obtains residuals from a source domain-trained Centered RF (CRF) in the target domain, then fits another CRF to these residuals with feature splitting probabilities proportional to feature-residual distance covariance. We derive an upper bound on the mean square error rate of the procedure that theoretically brings out the benefits of transfer learning in random forests. Our results explain why shallower trees in the residual random forest in the target domain provide implicit regularization.

Speaker Profile:

Subhadeep Paul is an Associate Professor in the Department of Statistics at The Ohio State University. He is also a faculty fellow, and previously served as a co-director of the foundations of data science and AI at the Translational Data Analytics Institute at Ohio State. He received his PhD in Statistics from the University of Illinois at Urbana-Champaign in 2017. His research focuses on statistical analysis of complex network-linked data and transfer and federated statistical learning. His research has been funded by two NSF grants from the algorithms of threat detection and mathematics of digital twins programs.

Webpage Link: https://stat.osu.edu/people/paul.963