Monte Carlo Simulation of the Impact of Distributional Properties on the Effectiveness of Cluster Boosted Regression
Faculty advisor: Prof. Mark Chignell

Clustering into patient types is a way of generating clinical predictions based on non-confidential summarized patient data (Chignell et al., 2013). Predictions made based on segmented patient types using Cluster-boosted regression can improve on predictions made using confidential raw patient data, with studies reported by Rouzbahman et al. (2017) showing around a 2 percent predication in the case of predicting length of stay and death status in an intensive care unit, and in predicting the likelihood of a visit to an emergency department within one month of assessment for late stage cancer patients.

The purpose of this project is to use Monte Carlo Simulation experiments to determine which distributional properties of multivariate data influence the magnitude of the boosting effect in cluster boosted regression. It is anticipated that this research should lead to a scientific paper that provides key insights into why cluster boosting is beneficial as well as providing criteria that can be used to determine which types of data set will stand to benefit more from the cluster boosting approach.

Required Skill: To carry out this project you should have some experience with statistical analysis and regression analysis in particular, and should be familiar with the R programming language and associated statistical and machine learning packages.

References

Rouzbahman, M., Jovicic, A., and Chignell, M. (2017). Can Cluster-Boosted Regression Improve Prediction?: Death and Length of Stay in the ICU. IEEE Journal of Biomedical and Health Informatics, 21(3), 851-858.

Chignell, M., Rouzbahman, M., Kealey, M.R., Yu, E., Samavi, R. and Sieminowski, T. Development of Non-Confidential Patient Types for Use in Emergency Medicine Clinical Decision Support. (2013). IEEE Security & Privacy, November/December, 2-8.

Contact: Mark Chignell chignel@mie.utoronto.ca

Research Area: Clinical Decision Support, Health Data Mining