Random KNN Classification and Regression

Shengqiao Li[1], Donald A. Adjeroh[2] and E. James Harner[1,⋆]

1. Department of Statistics, West Virginia University
2. Lane Department of Computer Science and Electrical Engineering, West Virginia University
⋆Contact author: jharner@stat.wvu.edu

Keywords: Classification, Machine Learning, K Nearest Neighbor, High Dimensional Data

High dimensional data, involving thousands of variables, are becoming increasingly available in various applications in biometrics, bioinformatics, chemometrics, and drug design. While such high dimensional data can be readily generated, successful analysis and modeling of such datasets is highly challenging. We present Random KNN, a novel generalization of traditional nearest-neighbor modeling. Random KNN consists of an ensemble of base k-nearest neighbor models, each constructed from a random subset of the input variables. We study the properties of the proposed Random KNN, including its theoretical convergence. Using different datasets, we perform an empirical analysis of its performance, and compare its results with those from recently proposed methods for high dimensional datasets. It is shown that Random KNN provides significant advantages in both the computational requirement and classification performance. The Random KNN approach can be applied to both qualitative and quantitative responses, i.e., classification and regression problems, and has applications in statistics, machine learning and pattern recognition. The Random KNN algorithms are implemented in the *rknn* R package.

References

Shengqiao Li, E. James Harner and Donald A. Adjeroh (2010). Random KNN. Submitted to *IEEE Transactions on Pattern Analysis and Machine Intelligence*.