Abstract

Random Forests are an effective ensemble method which is becoming increasingly popular, particularly for binary classification prediction problems. One of the most popular algorithms for implementing the Random Forest model is the Breiman and Cutler’s algorithm and this forms the basis of the "randomForest" package in R. However, a Random Forest model implemented using this package has a limitation, especially in a milieu which has limited computational power, that it cannot handle highly categorical data. In this paper, we present one of the many techniques we tried to improve the performance of a Random Forest Model using highly categorical data. The performance improvement was solely achieved using advanced pre-processing techniques like Optimal Scaling, hence the title of the paper.
Learning by-Teaching with SimStudent Technology Carnegie Mellon University.
- Mertz, J. S. 1997, Using Simulated Student for Instructional Design.

**Index Terms**

Computer Science  
Information Sciences

**Keywords**

Ensemble Methods  
Random Forest  
Prediction with Categorical Variables

Optimal Scaling

Classification

Machine Learning

Non-Linear Categorical Prediction.