Abstract

Random Forest (RF) is an ensemble supervised machine learning technique. Based on bagging and random feature selection, number of decision trees (base classifiers) is generated and majority voting is taken among them. The size of RF is subjective and varies from one dataset to another. Furthermore due to the randomization induced during creation, and its huge size, RF has at best been described as black-box. Various changes based on the experimental results have been proposed in the algorithm since then to optimize the performance of RF. To this end, we aim to find a subset, having accuracy comparable to original RF but having a much smaller size. In this paper, we show that the problem of selection of optimal subset of random forest follows the dynamic programming paradigm. Applying this approach to various UCI data-sets, corresponding subsets are obtained and studied. We conclude that such subsets do exist and that they are not unique. Moreover the size of these subsets is small fraction of the original RF (in the range of tens) and that accuracy of these subsets is a discrete valued function over its range.
An Approach towards Optimizing Random Forest using Dynamic Programming Algorithm

- Grahn H, Lavesson N, Lapajne M and Slat D, A CUDA implementation of Random Forest Early Results, Master Thesis Software Engineering, School of Computing, Blekinge Institute of Technology, Sweden
- Cormen, T., Leiserson, C., Rivest, R., Stein, C.: "Introduction To


**Index Terms**

Computer Science  
Algorithms

**Keywords**

Ensemble classifier  
Random Forest  
Decision tree  
Dynamic Programming  
Optimal subset