Classification Using C5.0

Max Kuhn

1. Pfizer Global R&D, Groton, Connecticut, USA
*Contact author: mxkuhn@gmail.com

Keywords: Classification Tree, Rule–Based Model, Boosting, Feature Selection

C4.5 (Quinlan, 1993) is a classification methodology that was developed in the same era as the classification and regression tree (CART) methodology (Breiman et al., 1984) but is more widely known outside of statistics. The model can the the form of a tree–based model or a set of distinct rules. C5.0 is an evolved version of this algorithm and, although it is very similar to C4.5, very little has been published on the primary improvements: feature selection (“winnowing”) and boosting. In 2010, Quinlan created an open–source version of the software, which we adapted into the C50 package. This talk briefly describes in the improvements in C5.0 as well as the functionality available in the package.

References
