Strobl, Carolin; Augustin, Thomas

Summary: Classification and regression trees are a popular and easy to interpret nonparametric regression approach, but are known to be very instable: Small changes in the learning sample can produce completely different trees. Therefore recently it has become state-of-the-art to consider ensembles (i.e., sets) of trees. The present paper contributes to the so-called TWIX approach, which produces ensembles by extra splits in additional outpoints. This approach can be considered as a compromise between the interpretable but instable single tree models and the stable but no longer interpretable ensemble methods bagging and random forests. Based on the idea to study the sensitivity of a split to some virtual, yet unseen observations, we develop a new, data driven, outpoint selection criterion, that technically turns out to be closely related to an upper entropy approach based on an imprecise Dirichlet model. Our criterion combines several attractive features: By adding extra outpoints only iff the underlying outpoint is instable, the tree is robustified parsimoniously and the computational expense of the resulting TWIX ensemble is reduced considerably. As a welcome by-product we moreover obtain a vivid diagnostic measure for the robustness of a single tree model. The rationale and benefit of our new adaptive criterion are illustrated by means of a small data example and a simulation study. Credal classification rules for robust aggregated predictions from sets of trees are briefly sketched in an outlook.

MSC:
62H30 Classification and discrimination; cluster analysis (statistical aspects)
62G08 Nonparametric regression and quantile regression
65C60 Computational problems in statistics (MSC2010)

Keywords:
classification trees; CART; C4.5; TWIX; bagging; random forests; outpoint selection; Gini index; Shannon entropy; imprecise Dirichlet model; credal classification; aggregation

Software:
C4.5; R

Full Text: DOI

References:
This reference list is based on information provided by the publisher or from digital mathematics libraries. Its items are heuristically matched to zbMATH identifiers and may contain data conversion errors. It attempts to reflect the references listed in the original paper as accurately as possible without claiming the completeness or perfect precision of the matching.