Influence Measures for CART Classification Trees

Avner Bar-Hen

Université Paris Descartes, France Avner.Bar-Hen@parisdescartes.fr

Servane Gey*

Université Paris Descartes, France Servane.Gey@parisdescartes.fr

Jean-Michel Poggi

Université Paris Sud, Orsay, France Jean-Michel.Poggi@math.u-psud.fr

Classically, robustness deals with model stability, considered globally. In data analysis, one may focus on individual observations diagnosis issues rather than model properties or variable selection problems. Some individuals are influential, which are not necessarily outliers or atypical individuals. Hence the question of measuring the influence of observations on the results obtained with classification trees is of interest.

To define the influence of individuals on the analysis, we propose criterions to measure the sensitivity of the Classification And Regression Trees (CART) analysis. The proposals are based on predictions and use jackknife trees. The analysis is extended to the pruned subtrees sequences of CART to produce specific notions of influence. Using the framework of influence functions, distributional results are derived.

A real dataset relating the administrative classification of cities surrounding Paris, France, to the characteristics of their tax revenues distribution, is analyzed using the new influence-based tools.

Key Words: Data analysis, Influential individuals, Decision Trees.