Influence Measures for CART Classification Trees

Avner Bar-Hen
Laboratoire MAP5, Université Paris Descartes, France

Servane Gey
Laboratoire MAP5, Université Paris Descartes, France

Jean-Michel Poggi
Laboratoire de Mathématiques, Université Paris Sud and Université Paris Descartes, France

Abstract: This paper deals with measuring the influence of observations on the results obtained with CART classification trees. To define the influence of individuals on the analysis, we use influence measures to propose criterions to quantify the sensitivity of the CART classification tree analysis. The proposals are based on predictions and use jackknife trees. The analysis is extended to the pruned sequences of CART trees to produce CART specific notions of influence. Using the framework of influence functions, distributional results are derived.

A numerical example, the well known spam dataset, is presented to illustrate the notions developed throughout the paper. A real dataset relating the administrative classification of cities surrounding Paris, France, to the characteristics of their tax revenues distribution, is finally analyzed using the new influence-based tools.

Keywords: Influential individuals; Influence functions; Decision trees; CART.

Authors’ Addresses: A. Bar-Hen, email: Avner.Bar-Hen@mi.parisdescartes.fr; S. Gey, email: Servane.Gey@parisdescartes.fr; J.-M. Poggi, email: Jean-Michel.poggi@math.u-psud.fr.

Published online: 11 March 2015
1. Introduction

Classification And Regression Trees (CART; Breiman, Friedman, Olshen, and Stone 1993) have proven to be very useful in various applied contexts mainly because models can include numerical as well as nominal explanatory variables and because models can be easily represented (see for example Zhang and Singer 2010, or Bel, Allard, Laurent, Cheddadi, and Bar-Hen 2009). Because CART is a nonparametric method as well as it provides data partitioning into distinct groups, such tree models have several additional advantages over other techniques: for example input data do not need to be normally distributed, predictor variables are not supposed to be independent, and non-linear relationships between predictor variables and observed data can be handled.

It is well known that CART appears to be sensitive to perturbations of the learning set. This drawback is even a key property to make resampling and ensemble-based methods (as bagging and boosting) effective (see Gey and Poggi 2006). To preserve interpretability of the obtained model, it is important in many practical situations to try to restrict to a single tree. The stability of decision trees is then clearly an important issue and then it is important to be able to evaluate the sensitivity of the data on the results. Briand, Ducharme, Parache, and Mercat-Rommens (2009) proposed a similarity measure between trees to quantify it and use it from an optimization perspective to build a less sensitive variant of CART. This view of instability related to bootstrap ideas can be also examined from a local perspective. Following this line, Bousquet and Elisseeff (2002) studied the stability of a given method by replacing one observation in the learning sample with another one coming from the same model.

Many authors derived asymptotic normality of the influence functions under weak assumptions (see for example Campell 1978; Critchley and Vitiello 1991; Croux and Joossens 2005; Croux, Filzmoser, and Joossens 2008; and Croux, Haesbroeck, and Joossens 2008). These results can be used to obtain a threshold to decide whether an observation is influential or not.

The aim of this paper is to focus on individual observations diagnosis issues rather than model properties or variable selection problems. The use of an influence measure is a classical diagnostic method to measure the perturbation induced by a single element, in other terms we examine stability issue through jackknife. We use decision trees to perform diagnosis on observations. We do not propose a solution to the eternal accuracy/interpretability dilemma which is typical within tree-based methods framework since our goal is the analysis of influence observations after the tree-structure is built.

The outline is the following. Section 2 recalls first some general background on the so-called CART method. Then it introduces some influence