Influence Measures for CART Classification Trees

Avner Bar-Hen[1], Servane Gey[1,3], and Jean-Michel Poggi[2]

[1]Laboratoire MAP5, Université Paris Descartes France
[2]Laboratoire de Mathématiques, Université Paris Sud, Orsay, France
[3]Corresponding author, Servane Gey, e-mail Servane.Gey@parisdescartes.fr

Abstract

This paper deals with measuring the influence of observations on the results obtained with classification trees. 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.

1 Introduction

Classification And Regression Trees (CART; Breiman *et al.* (1984) [4]) 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) [26], or Bel *et al.* (2009) [2]). 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) [13]). 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 *et al.* (2009) [5] 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) [3] 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. For example, discriminant analysis has been studied by Campbell (1978) [6], Critchley and Vitiello (1991) [8] for the linear case and Croux and Joossens (2005) [9] for the quadratic one. For linear discrimination influence functions on the error rate, or the prediction error of binary classifiers, were considered in [10, 11]. Variance of the asymptotic normal distribution is generally estimated through resampling techniques. Therefore these results could be used to obtain a threshold to decide whether an observation is an outlier 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.

2 CART classification trees

The data are considered as an independent sample of the random variables $(X^1, \ldots, X^p, Y)$, where the X^ks are the explanatory variables and Y is the categorical variable to be explained. CART is a rule-based method that generates a binary tree through recursive partitioning that splits a subset (called a node) of the data set into two subsets (called sub-nodes) according to the minimization of a heterogeneity criterion computed on the resulting sub-nodes. Each split is based on a single variable. Let us consider a decision tree T. When Y is a categorical variable a class label is assigned to each terminal node (or leaf) of T. Hence T can be viewed as a mapping to assign a value $\widehat{Y}_i = T(X_i^1, \ldots, X_i^p)$ to each observation.

Among all binary partitions of each set of values of the explanatory variables at a node t, the principle of CART is to split t into two sub-nodes t_- and t_+ according to a threshold on one of the variables (or a subset of the labels for categorical variables), such that the reduction of heterogeneity between a node and the two sub-nodes is maximized. The growing procedure is stopped when there is no more admissible splitting. Each leaf is assigned to the most frequent class of its observations. Of course, such a maximal tree (denoted by T_{max}) generally overfits the training data and the associated prediction error $R(T_{max})$, with

$$R(T) = \mathbb{P}(T(X^1, \ldots, X^p) \neq Y), \tag{1}$$

is typically large. Since the goal is to build from the available data a tree T whose prediction error is as small as possible, in a second stage the tree T_{max} is pruned to produce a subtree T' whose expected performance is close to the minimum of $R(T')$ over all binary subtrees T' of T_{max}. The pruning is based on the penalized empirical risk $\hat{R}_{pen}(T)$ to balance optimistic estimates of empirical risk by adding a complexity term that penalizes larger subtrees:

$$\hat{R}_{pen}(T) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{T(X_i^1, \ldots, X_i^p) \neq Y_i} + \alpha |T| \tag{2}$$

where $\mathbb{1}$ is the indicator function, n the total number of observations, α a positive penalty constant, $|T|$ denotes the number of leaves of the tree T and Y_i is the ith random realization of Y.

The R package *rpart* provides both the sequence of subtrees pruned from a deep maximal tree and a final tree selected from this sequence by using the 1-SE rule (see [4]). The penalized criterion used in the pruning of *rpart* is $\hat{R}_{pen}$ defined by (2). The cost complexity parameter denoted by *cp* corresponds to the temperature α used in the original penalized criterion (2) divided by the misclassification rate of the root of the tree. The pruning step leads to a sequence $\{T_1; \ldots; T_K\}$ of nested subtrees (where T_K is reduced to the root of the tree) associated with a nondecreasing sequence of temperatures $\{cp_1; \ldots; cp_K\}$. Then, a tree is chosen among this sequence by cross-validation.

3 Influence measures for CART

Let $X = (X^1, \ldots, X^p) \in \mathcal{X}$ be the vector of the explanatory variables, and consider that the data are independent realizations $\mathcal{L} = \{(x_1, y_1); \ldots; (x_n, y_n)\}$ of $(X, Y) \in \mathcal{X} \times \{1; \ldots; J\}$. The dependent variable Y is assumed to be a categorical variable with J unordered categories. Influence measures quantify discrepancy between T, the tree computed with the complete sample $\mathcal{L}$ and T^{-i}, the jackknife tree computed with $\mathcal{L}^{-i}$, the whole sample minus the observation (x_i, y_i).

Let $T(x_k)$ (resp. $T^{-i}(x_k)$) be the the class prediction of x_k based on T (resp. T^{-i}).

3.0.1 Influence on predictions

The first natural idea is to focus on class predictions. In this case, influence measure is directly related to $\mathbb{1}_{T(x_k)\neq T^{(-i)}(x_k)}$:

$$I_1(x_i) = \frac{1}{n-1}\sum_{k=1;k\neq i}^{n} \mathbb{1}_{T(x_k)\neq T^{(-i)}(x_k)}, \quad (3)$$

$$I_2(x_i) = \mathbb{1}_{T(x_i)\neq T^{(-i)}(x_i)} \quad (4)$$

$I_1(x_i)$ is the proportion of observations for which the predicted label changes using the jackknife tree $T^{(-i)}$ instead of the reference tree T. It is closely related to the resubstitution estimate of the prediction error.
$I_2(x_i)$ indicates if x_i is classified in a different way by T and $T^{(-i)}$ and is closely related to the leave-one-out estimate of the prediction error.

3.0.2 Influence based on subtrees sequences

Another way to look at the data is to consider the complexity cost constant, penalizing bigger trees in the pruning step of the CART tree design, as a tuning parameter. It allows to scan the data and sort them with respect to their influence on the CART tree.
Let us consider on the one hand the sequence of subtrees based on the complete dataset, denoted by T_{cp_j}, and on the other hand the n jackknife sequences of subtrees based on the jackknife subsamples $\mathcal{L}^{-i}$, denoted by $T_{cp_j}^{(-i)}$. Suppose that the sequence T_{cp_j} contains K_T elements, and that each sequence $T_{cp_j}^{(-i)}$ contains $K_{T^{(-i)}}$ elements ($i = 1, \ldots, n$). This leads to a total of $N_{cp} \leqslant K_T + \sum_{1\leqslant i\leqslant n} K_{T^{(-i)}}$ distinct values $\{cp_1; \ldots; cp_{N_{cp}}\}$ of the cost complexity parameter in increasing order from cp_1 to $cp_{N_{cp}} = \max_{1\leqslant j\leqslant N_{cp}} cp_j$.

Then, for each value cp_j of the complexity and each observation x_i, we compute the binary variable $\mathbb{1}_{T_{cp_j}(x_i)\neq T_{cp_j}^{(-i)}(x_i)}$ that tells us if the reference and jackknife subtrees corresponding to the same complexity cp_j provide different predicted labels for the removed observation x_i. Thus we define influence measures I_3 and I_4 as the number of complexities for which these predicted labels differ: for $i = 1, \ldots, n$

$$I_3(x_i) = \frac{1}{(n-1)N_{cp}}\sum_{k=1;k\neq i}^{n}\sum_{j=1}^{N_{cp}} \mathbb{1}_{T_{cp_j}(x_k)\neq T_{cp_j}^{(-i)}(x_k)} \quad (5)$$

$$I_4(x_i) = \frac{1}{N_{cp}}\sum_{j=1}^{N_{cp}} \mathbb{1}_{T_{cp_j}(x_i)\neq T_{cp_j}^{(-i)}(x_i)} \quad (6)$$

3.1 Distributional results

Let I be any of the resubstitution indices (I_1, I_3), and $\hat{I}$ be the estimate of I computed as defined in equations (3) and (5).

Proposition 1. *If $\hat{\sigma}^2$ is the jackknife estimate of the variance of I, we obtain the following confidence interval for $I > 0$:*

$$\left[\, \hat{I} - \epsilon_\alpha \hat{\sigma} \;;\; \hat{I} + \epsilon_\alpha \hat{\sigma} \,\right] \quad (7)$$

where ϵ_α is the value of a standard gaussian variable that has the probability $\frac{\alpha}{2}$ to be exceeded.

Using the property of jackknife estimate (see [19]), an unbiased estimate of $\hat{\sigma}$ is given by the variance of the values $\left(\hat{I}(x_i)\right)_{1\leqslant i\leqslant n}$.

4 Exploring Paris Tax Revenues data

4.1 Dataset and reference tree

We apply the tools presented in the previous section on tax revenues of households in 2007 from the 143 cities surrounding Paris. Cities are grouped into four counties (corresponding

to the french administrative "département"). The PATARE data (PAris TAx REvenues) are freely available on `http://www.data-publica.com/data`. For confidentiality reason we do not have access to the tax revenues of the individual households but we have characteristics of the distribution of the tax revenues per city. For each city, we have the first and the 9th deciles (named respectively *D1* and *D9*), the quartiles (named respectively *Q1*, *Q2* and *Q3*), the mean, and the percentage of the tax revenues coming from the salaries and treatments (named *PtSal*).
Basically we tried to predict the county of the city with the characteristics of the tax revenues distribution.

The reference tree shows that the extreme quantiles are sufficient to separate richest from poorest counties while more global predictors are useful to further discriminate between intermediate cities.
Surprisingly, the predictions given by the reference tree are generally correct (the resubstitution misclassification rate calculated from the confusion matrix is equal to 24.3%). Since the cities within each county are very heterogeneous, we look for the cities which perturb the reference tree.

4.2 Influential observations

4.2.1 Presentation

The threshold given by the critical value of the unilateral test at level 95% for I_1 highlights ten cities, which are in descending order of influence: *Villemomble*, *Neuilly-Plaisance*, *Chevilly-Larue*, *Vitry-sur-Seine*, *Villejuif*, *Creteil*, *Choisy-le-Roi*, *Champigny-sur-Marne*, *Arcueil* and *Noisy-le-Grand*. Let us note that only three of them are highlighted by index I_2.

There are 29 different values of complexities in the reference and jackknife trees sequences. Index I_3 selects nine cities using the threshold given by the critical value of the unilateral test at level 97.5% : roughly the same except that *Neuilly-Plaisance* and *Noisy-le-Grand* are not selected while *Bonneuil-sur-Marne* is selected.

The influential cities, according to index I_4 at level 95% et 97.5%, are in descending order of influence the 5 following cities: *Paris 13eme*, *Villemomble*, *Asnieres-sur-Seine*, *Rueil Malmaison* and *Bry-sur-Marne*.
So, influence indices I_1 and I_3 deliver similar conclusions while I_4 highlights a different subset of observations.

4.2.2 Interpretation

Index I_4 highlights cities for which two parts of the city can be distinguished: a popular one with a low social level and a rich one of high social level. They are located in the right part of the reference tree (for the higher values of I_4: *Asnieres sur Seine, Villemomble, Paris 13eme, Bry-sur-Marne* and *Rueil-Malmaison*) as well as in the left part (for moderate values of I_4: *Chatenay-Malabry, Clamart, Fontenay aux Roses, Gagny, Livry-Gargan, Vanves, Chevilly-Larue, Gentilly, Le Perreux sur Marne, Le Pre-Saint-Gervais, Maisons-Alfort, Villeneuve-le-Roi, Vincennes* and the particularly interesting city *Villemomble* (see Figure 1).
To explore the converse, we inspect now the list of the 51 cities associated with lowest values of I_4 which can be considered as the less influential, the more stable. It can be easily seen that it corresponds to the 16 rich district of Paris downtown (*Paris 1er* to *12eme* and *Paris 14eme* to *Paris 16eme*) and mainly cities near Paris or directly connected by the RER line transportation.

In addition, one may notice that influential observations for PCA (Principal Component analysis) are not related to influential cities detected using I_4 index (see Figure 2). To end this study, the map in Figure 2 shows that Paris is stable, and that each surrounding county contains a stable area: the richest or the poorest cities. These areas are clustered.

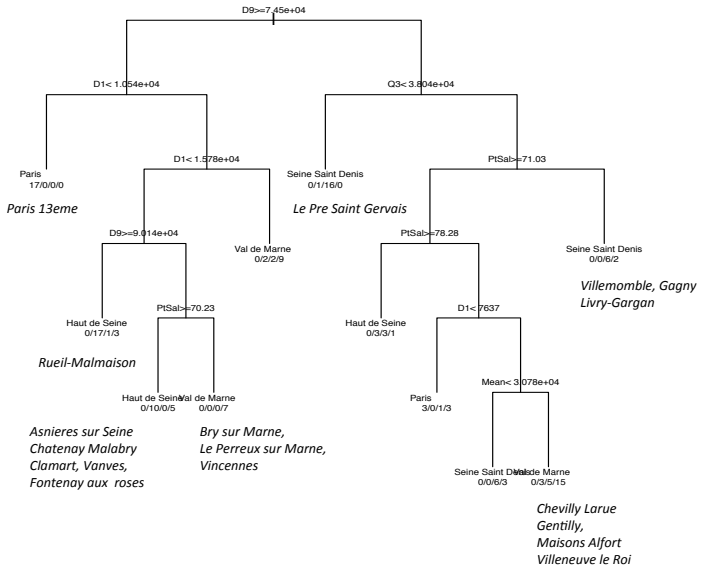

Figure 1: Influential cities located on the CART reference tree.

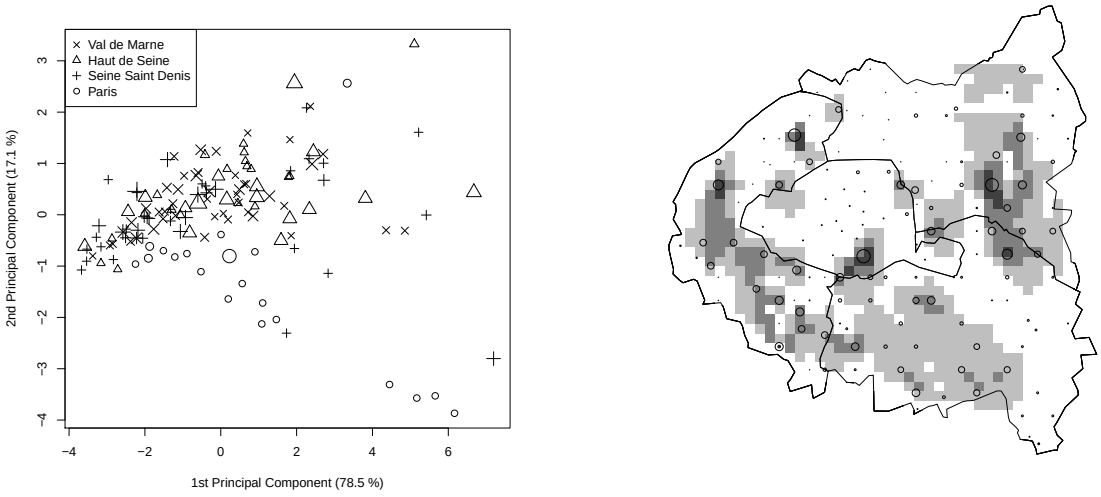

Figure 2: *Left*: Plane of the two first principal components: Cities are represented by symbols proportional to influence index I_4. *Right*: The 143 cities are represented by a circle proportional to the influence index I_4 and a spatial interpolation is performed using 4 grey levels

References

[1] Bar-Hen, A., Mariadassou, M., Poursat, M.-A. and Vandenkoornhuyse, Ph. (2008).*Influence Function for Robust Phylogenetic Reconstructions*. Molecular Biology and Evolution, 25(5), 869-873.

[2] Bel, L., Allard, D., Laurent, J.M., Cheddadi, R. and Bar-Hen, A. (2009). *CART al-*

gorithm for spatial data: application to environmental and ecological data. Computat. Stat. and Data Anal., 53(8), 3082-3093.

[3] Bousquet, O., Elisseeff, A. (2002). *Stability and generalization.* J. Machine Learning Res. 2, 499–526.

[4] Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. *Classification and regression trees.* Chapman & Hall (1984).

[5] Briand, B., Ducharme, G. R., Parache, V. and Mercat-Rommens, C. (2009). *A similarity measure to assess the stability of classification trees*, Comput. Stat. Data Anal., 53(4), 1208–1217.

[6] Campbell, N.A. (1978). *The influence function as an aid in outlier detection in discriminant analysis.*, Appl. Statist., 27, 251–258.

[7] Chèze, N. and Poggi, J.M. (2006). *Outlier detection by boosting regression trees.* Journal of Statistical Research of Iran (JSRI), 3, 1–21.

[8] Critchley, F. and Vitiello, C. (1991). *The influence of observations on misclassification probability estimates in linear discriminant analysis.*, Biometrika, 78, 677–690.

[9] Croux, C. and Joossens, K. (2005). *Influence of observations on the misclassification probability in quadratic discriminant analysis.*, Journal of Multivariate Analysis, 96(2), 384–403.

[10] Croux, C., Filzmoser, P., and Joossens, K. (2008). *Classification Efficiencies for Robust Linear Discriminant Analysis*,Statistica Sinica, 18(2), 581–599

[11] Croux, C., Haesbroeck, G., and Joossens, K. (2008). *Logistic Discrimination using Robust Estimators: an influence function approach*, The Canadian Journal of Statistics, 36(1), 157–174.

[12] A. Cuevas and J. Romo (1995). *On the estimation of the influence curve.* The Canadian Journal of Statistics, vol.23, 1–9.

[13] Gey, S. and Poggi, J.M. (2006). *Boosting and instability for regression trees.* Comput. Stat. Data Anal., 50(2), 533-550.

[14] R. D. Gill *Non- and semi-parametric maximum likelihood estimators and the von Mises method (part. 1).* Scand. J. Statist., 16, 97–128 (1989).

[15] Hampel, F. R. (1988). *The influence curve and its role in robust estimation.* J. Amer. Statist. Assoc., 69.

[16] Hastie, T.J., Tibshirani, R.J. and Friedman, J.H. (2009). *The elements of statistical learning: data mining, inference and prediction.* Third edition, Springer, New-York.

[17] Huber, P. J. (1981). *"Robust Statistics"*, Wiley & Sons.

[18] Miglio, R. and Soffritti, G. (2004). *The comparison between classification trees through proximity measures.* Comput. Stat. Data Anal., 45(3), 577–593.

[19] Miller, R. G. (1974). *The jackknife - a review.* Biometrika, 61, 1-15.

[20] R Development Core Team *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0 (2009). URL http://www.R-project.org/.

[21] Rousseeuw, P. (1984). *Least median of squares regression*, J. Amer. Statist. Assoc., 79, 871-880.

[22] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection.* Wiley, Interscience, New York.

[23] Youness, G. and Saporta, G. (2009). *Comparing partitions of two sets of units based on the same variables.* Advances in Data Analysis and Classification, 4(1), 53-64.

[24] Venables, W. N., and Ripley, B.D. (2002). *Modern Applied Statistics with S.* Fourth Edition, Springer.

[25] Verboven, S. and Hubert, M. (2005). *LIBRA: a MATLAB library for robust analysis*, Chemometrics and Intelligent Laboratory Systems, 75, 127-136.

[26] Zhang, H. and Singer, B. H. (2010). *Recursive Partitioning and Applications*, 2^{nd} edition, Springer.