Decision Tree for uncertainty measures

M. Tami [1], M. Clausel [2], E. Devijver [1], E. Gaussier [1], J-M. Aubert [3] & M. Chebre [3]

[1] *Univ. Grenoble Alpes, CNRS, Grenoble INP[1], LIG, 38000 Grenoble, France*
[2] *Université de Lorraine, Nancy · IEC - Institut Elie Cartan, 54052 Nancy, France*
[3] *TOTAL S.A., 24 cours Michelet - Défense 10, 92069 Paris La Défense Cedex France*

Résumé. Les méthodes d'ensembles sont très populaires et performantes sur des problématiques de classification ou de prédiction. Elles sont basées sur l'agrégation de plusieurs classifieurs, qui sont des arbres de régression ou de classification. On considère une pondération de leurs prédictions pour prédire une valeur ou une classe d'une nouvelle instance de donnée. Concrètement, un arbre est un estimateur constant par morceaux sur des partitions disjointes de l'espace des variables d'entrées. Ces partitions sont construites par des divisions dyadiques récursives de l'ensemble des variables d'entrées qui minimisent une fonction de risque. En pratique, les observations d'une base de données sont sous-jacentes à des mesures qui peuvent présenter une incertitude. Ce travail propose d'étendre la construction d'un arbre de régression tel que CART à ce type de données. Nous introduisons d'abord la modélisation induite et adaptée à des données incertaines, puis nous présentons des règles de partitionnement et de prédiction pour la construction d'arbres prenant en compte l'incertitude de chaque observation quantitative d'une base de données.

Mots-clés. Apprentissage machine, Méthodes d'ensemble, Arbres de régression, Mesures incertaines, Variables hétérogènes.

Abstract. The ensemble methods are popular machine learning techniques which are powerful when one wants to deal with both classification or prediction problems. A set of classifiers (regression or classification trees) is constructed, and the classification or the prediction of a new data instance is done by tacking a weighted vote. A tree is a piece-wise constant estimator on partitions obtained from the data. These partitions are induced by recursive dyadic split of the set of input variables. For example, CART (Classification And Regression Trees) [1] is an efficient algorithm for the construction of a tree. The goal is to partition the space of input variable values in the most as possible "homogeneous" K disjoint regions. More precisely, each partitioning value has to minimize a risk function. However, in practice, experimental measures can be observed with uncertainty. This work proposes to extend CART algorithm to this kind of data. We present an induced model adapted to uncertainty data and both a prediction and split rule for a tree construction taking into account the uncertainty of each quantitative observation from the data base.

Keywords. Machine learning, Ensemble methods, Regression trees, Uncertainty measures, Heterogeneous variables.

[1]Institute of Engineering Univ. Grenoble Alpes

1 Introduction

When each quantitative observation of a dataset or a part of them has been measured with an uncertainty described by a known probability distribution and a known variance parameter value (but unknown mean parameter), the induced learning set is more informative than a classical learning set used by ensemble methods. To deal with this supplementary information, this work proposes to extend the ensemble methods to this kind of learning set. According to the literature [2, 3, 4], several algorithms for the construction of a tree are available, but they are all based on the minimization of a risk function and can be summarized into two tasks: a **split rule** and a **prediction rule**. For that purpose, each data instance is supposed to belong to a subset of the input variables space. However, in our context of uncertainty, each data instance is associated with a latent instance data. Thus, the construction process of a tree has to take into account the probability that the underlying latent instance belongs to each other disjoint regions. This is made by the knowledge available in the learning set, meaning, probability distributions and their variance parameter values. The induced model and notations are introduced in Section 2. To illustrate the contribution of this work, Section 3 presents the risk function, the prediction rule and the split rule for the CART algorithm and then, a new tree formula, which takes into account the uncertainty, is proposed in Section 4. The risk function formula modified is presented, and the induced split and prediction rules are demonstrated. To conclude, we present a discussion on a set of perspectives of this work in Section 5.

2 Model and notations

Let $(\mathbf{X}, Y)$ be a concatenation of variables taking values in $\mathcal{X} \times \mathbb{R}$, where $\mathbf{X} = (X^1, \cdots, X^p)$, $\mathcal{X} = \mathbb{R}^p$ is the input space[2], and $\mathbb{R}$ is the output space. We assume we have access to a set $(\boldsymbol{x}_i, y_i)_{1 \leq i \leq n}$ of size n of independent past cases of $(\mathbf{X}, Y)$ taken from a universe Ω and a set of probability distributions with their respective variance parameter values. The learning set $\mathcal{L}_n$ is then defined by

$$\mathcal{L}_n = \left\{ \left((x_i^1, \cdots, x_i^p, y_i)_{1 \leq i \leq n}, (\mathbb{P}^j, \sigma_j)_{1 \leq j \leq p}, \mathbb{P}^Y, \sigma_Y \right) \right\} \tag{1}$$

where $\mathbb{P}^j$ is the probability distribution associated to the input variable X^j for $j \in \{1, \cdots, p\}$ and σ_j^2 is its associated variance (the corresponding mean is unknown), and similarly $\mathbb{P}^Y$ is the probability distribution associated to the output variable Y_i and σ_Y^2 is its associated variance (and the corresponding mean is unknown).

[2]In order to avoid heavy formulas, we present the reduced case $\mathcal{X} = \mathbb{R}^p$ in the sequel, but, with no loss of generality, we can extend this work to $\mathcal{X} = \mathbb{R}^{p-q} \times \mathcal{Q}^q$ where $\mathcal{Q}$ is a finite set of values. Thus, categorical input variables can be managed by this approach. Then, we can also extend this case to some quantitative input observations without uncertainty. For this purpose, one just has to manage these input both categorical and without uncertainty observations classically such as in CART.

Each observation is independent and identically distributed from the other. Then, probability distributions defined by $(\mathbb{P}^j)_{1\leq j\leq p}$, $\mathbb{P}^Y$ and $(\sigma_j^2)_{1\leq j\leq p}$, σ_Y^2 do not depend on the index i of the observations and are fixed through the learning set $\mathcal{L}_n$.

In the uncertainty measure case, we assume that, for every $i \in \{1, \cdots, n\}$, each observed value x_i^j is coming from a random variable X^j, for $j \in \{1, \cdots, p\}$, which corresponds to a latent variable U^j with a measurement error ϵ^j: formally, each observation (input and output) is defined such as, for $1 \leq j \leq p$,

$$\begin{cases} Y = f(\mathbf{X}) + \epsilon, \\ X^j = U^j + \epsilon^j \\ Y = U^Y + \epsilon^Y, \end{cases} \tag{2}$$

where $\epsilon^j \sim \mathbb{P}^j_{\sigma_j^2}$, $\epsilon_i^Y \sim \mathbb{P}^Y_{\sigma_Y^2}$, ϵ given $\mathbf{X}$ is a zero mean measurement error with a variance parameter σ^2 and f is the unknown regression function. The joint probability distribution of $(\mathbf{X}, Y)$ is unknown, then our goal is to learn the function $f : \mathcal{X} \mapsto \mathbb{R}$ from $\mathcal{L}_n$. This supervised learning task can be realized by the construction of a regression tree T, providing an estimator $\hat{f}$ and then a prediction $\hat{y} = \hat{f}(\boldsymbol{x})$ for a new observation $\boldsymbol{x}$. The estimator $\hat{f}$ is constructed by minimizing the empirical quadratic risk:

$$F_n(f, \hat{f}, \mathcal{L}_m) = \frac{1}{n} \sum_{k=1}^{K} \sum_{\{1\leq i\leq m:(\boldsymbol{x}_i, y_i)\in R_k\}} (y_i - \hat{f}(\boldsymbol{x}_i))^2. \tag{3}$$

In this work, we propose to extend the popular CART construction (restricted to the regression case) to the uncertainty measures case. To tackle this issue, we take into account the supplementary knowledge given by $((\mathbb{P}^j, \sigma_j)_{1\leq j\leq p})$ in the learning set. The knowledge given by $(\mathbb{P}^Y, \sigma_Y)$ is not taken into account in this project, but is part of our perspectives.

3 Regression decision tree: classical case

In this section, the main steps to construct a regression decision tree without uncertainty measures are described. The learning set in this case is defined by $\mathcal{L}_n^* = \left\{(x_i^1, \cdots, x_i^p, y_i)_{1\leq i\leq n}\right\}$. Binary tree is considered. The set $\mathcal{X}$ is associated to the root. Then, a node t represents a subset of input observations, which belong to a certain region, the corresponding learning set is denoted by $\mathcal{L}_{n|t}^*$ and similarly the corresponding estimator is denoted $\hat{f}_{|t}$. The construction of the tree is done by induction: we subdivide every node t (or the corresponding region) into two nodes t_L and t_R (or two subregions) with respect to a split which needs to be defined. In the end, we get a partition of the input space $\mathcal{X}$ into K regions $(R_k)_{1\leq k\leq K}$, and then, for $1 \leq k \leq K$, a prediction γ_k is assigned to each region R_k. In this work, the regions $(R_k)_{1\leq k\leq K}$ are generated by splits (hyper-planes from $\mathcal{X}$) with split zones (boundaries of R_k) parallel to the axes formed by the input variables: we focus on hyperrectangles. Thus, a tree is defined by its parameters $\Theta = \{(R_k, \gamma_k)_{1\leq k\leq K}\}$, and we estimate the function f by $\hat{f}(\boldsymbol{x}) = T(\boldsymbol{x}; \Theta) := \sum_{k=1}^{K} \gamma_k \mathbb{1}_{\{\boldsymbol{x}\in R_k\}}$.

Parameters are estimated by minimizing the quadratic risk (3). It consists of a minimization problem over two multivariate variables, and we decompose this optimization problem into two steps: the **split rule** and the **predictive rule** described in the next subsections. Those steps are iterated until a stopping rule (to be defined). It leads to construct a maximal tree, which should be pruned.

3.1 Predictive rule

For $1 \leq k \leq K$, a constant γ_k has to be assigned to each region R_k such that the predictive rule for a new observation $\boldsymbol{x}$ becomes: $\boldsymbol{x} \in R_k \Rightarrow \hat{f}(\boldsymbol{x}) = \gamma_k$. Each constant is computed by minimizing the empirical quadratic risk defined in (3): for all $1 \leq k \leq K$,

$$\hat{\gamma}_k = \underset{\gamma_k \in \mathbb{R}}{\operatorname{argmin}} \left\{ F_n(f, \gamma_k \mathbb{1}_{\{\boldsymbol{x}_i \in R_k\}}, \mathcal{L}_n^*) \right\} = \frac{1}{|\{i : \boldsymbol{x}_i \in R_k\}|} \sum_{i:\boldsymbol{x}_i \in R_k} y_i. \tag{4}$$

3.2 Split rule

To define the regions, we have to define the variable with respect to which we are splitting, and the splitting value. To select the best variable to split, we compute the best split for every variable, and then, among all those splits, we define the variable which minimizes the quadratic risk. As the construction is done by induction, we fix a node t which corresponds to a region $\mathcal{R}$, and we split it into two disjoint child nodes t_L and t_R corresponding to the regions $\mathcal{R} \cap \{X^j < \hat{s}_t^j\}$ and $\mathcal{R} \cap \{X^j \geq \hat{s}_t^j\}$.

Fix the variable index $1 \leq j \leq p$. The split $\hat{s}_t^j$ is chosen to minimize the impurity in both t_L and t_R:

$$\hat{s}_t^j = \underset{\{(t_L, t_R) : t = t_L \cup t_R\}}{\operatorname{argmin}} \left[F_n(f, \hat{f}_{|t_L}, \mathcal{L}_{n|t_L}^*) + F_n(f, \hat{f}_{|t_R}, \mathcal{L}_{n|t_R}^*) \right] \tag{5}$$

To speed up the computation, we are looking for splits in the set of observed values: we are testing every possible values in the finite set $\mathcal{S}^j = \{(x_i^j)_{1 \leq i \leq n}\}$. Then, we select the index j minimizing the quadratic risk, where the split has been defined beforehand.

4 Regression decision tree with uncertainty measures

In this section, uncertainty measures are taken into account through the learning set defined in (1) to construct a tree, in order to be as close as possible to the expected tree that would be constructed on the unknown set $\{((U_i^j)_{1 \leq j \leq p}, U_i^Y)_{1 \leq i \leq n}\}$. Some intuition to understand the differences between the construction of the two trees is first given. For an observation $\boldsymbol{x}_i$ belonging to a region R_k, the underlying latent observation of the variable $\mathbf{U}_i$ can be in a different region $R_{k'}$. Indeed, the measurement error can be large enough to modify the region-belonging. Thus, given the value $\boldsymbol{x}_i$, the construction of the estimator $\hat{f}$ has to take into

account the probability of $\mathbf{U}_i \in R_{k'}$ for each region $k' \in \{1, \cdots, K\}$. We are then interested in $\mathbb{P}(\mathbf{U}_i \in R_k | \mathbf{X}_i = \boldsymbol{x}_i)$. In this work, we assume that the variables are independent, then this probability is decomposed as follows:

$$P_{i,k} := \mathbb{P}(\mathbf{U}_i \in R_k | \mathbf{X}_i = \boldsymbol{x}_i) = \prod_{j=1}^{p} \mathbb{P}(U_i^j \in R_k^j | X_i^j = x_i^j).$$

This probability can be computed in some specific cases, e.g. when $\mathbb{P}^j$ is normal and X is normally distributed. In general, assuming a parametric family, the distribution of X^j can be estimated from the observations; one could then obtain the distribution of $U^j|X^j$ using Fourier transforms (an explicit formula may be obtained when the marginal and conditional distributions are related). We plan to rely on numerical approximations to compute the above probabilities in the general case. We assume in the following that we can compute those probabilities. P denotes the corresponding matrix of size $n \times K$, with coefficients $P_{i,k}$. Thus, to estimate the function f, we propose a new tree formula $\hat{f}(\boldsymbol{x}) = T(\boldsymbol{x}; \Theta) := \sum_{k=1}^{K} \gamma_k \mathbb{P}(\mathbf{U} \in R_k | \mathbf{X} = \boldsymbol{x})$ defined by its parameters $\Theta = \{(R_k, \gamma_k)_{1 \leq k \leq K}\}$.

4.1 Predictive rule

For $1 \leq k \leq K$, constant γ_k has to be assigned to each region R_k such that the predictive rule for a new observation $\boldsymbol{x}$ becomes: $\mathbf{X} = \boldsymbol{x} \Rightarrow \hat{f}(\boldsymbol{x}) = \sum_{k=1}^{K} \gamma_k \mathbb{P}(\mathbf{U} \in R_k | \mathbf{X} = \boldsymbol{x})$. Each constant is computed by minimizing the empirical quadratic risk defined in (3): for all $1 \leq k \leq K$,

$$\hat{\gamma}_k = \underset{\gamma_k \in \mathbb{R}}{\operatorname{argmin}} \{F_n(f, \gamma_k \mathbb{P}(\mathbf{U}_i \in R_k | \mathbf{X}_i = \boldsymbol{x}_i), \mathcal{L}_n^*)\} = \left(P^T P\right)^{-1} P^T \boldsymbol{y}, \tag{6}$$

4.2 Split Rule

In the case of the learning set (1) and the model (2), the better split $\hat{s}_t^j$ for a node t minimizes the impurity in both child nodes t_L and t_R:

$$\begin{aligned} \hat{s}_t^j &= \underset{\{(t_L, t_R): t = t_L \cup t_R\}}{\operatorname{argmin}} \left[F_n(f, \hat{f}_{|t_L}, \mathcal{L}_{n|t_L}) + F_n(f, \hat{f}_{|t_R}, \mathcal{L}_{n|t_R}) \right] \\ &= \underset{\{(t_L, t_R): t = t_L \cup t_R\}}{\operatorname{argmin}} \left[\frac{1}{n} \sum_{k=1}^{K} \sum_{\{i: \boldsymbol{x}_i \in t_L\}} \left(y_i - P_i (P^T P)^{-1} P^T \cdot \boldsymbol{y} \right)^2 \right. \\ &\qquad \left. + \frac{1}{n} \sum_{k=1}^{K} \sum_{\{i: \boldsymbol{x}_i \in t_R\}} \left(y_i - P_i (P^T P)^{-1} P^T \cdot \boldsymbol{y} \right)^2 \right] \end{aligned} \tag{7}$$

Then, it minimizes the residual error in each child nodes t_L and t_R.

5 Discussion

From this proposed construction process, we are planning to move to the random forest [5, 4] or to Gradient Boosted Trees (GBT) [7, 8] to improve the performances. Further, we plan to extend Quantile Regression Forests [6] to manage the uncertainty measures from the output data and to take into account the information $(\mathbb{P}^Y, \sigma_Y)$ available in the learning set (1). In the end, this approach will be extended to quantile regression GBT [9, 10, 11] to benefit from both aspects. To this end, we plan to define a pruning step to be available to select the better constructed tree among the sub-trees available via the maximal tree. Thus, the obtained set of classifiers will be better and their weighted vote will be better again. We plan also to extend this work to the classification trees construction.

6 Acknowledgments

The authors acknowledge A. HAUDIBERT-AYET TOTAL CTG and S. JANAQI for usefull discussions and Total, that partially funded this research, as well as the Grenoble Alpes Data Institute, supported by the French National Research Agency under the 305 "Investissements d'avenir" program (ANR-15-IDEX-02), and the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01).

References

[1] Breiman, L. and Friedman, J. and Stone, C. J. and Olshen, R. A. (1984). Classification and regression trees. CRC press

[2] Genuer, R. and Poggi, J.-M. (2016). Arbres CART et Forêts aléatoires, Importance et sélection de variables. arXiv preprint 1610.08203

[3] Gey, S. (2002). Bornes de risque, détection de ruptures, boosting: trois thèmes statistiques autour de CART en régression. Thèse soutenue à l'université Paris-Sud.

[4] Louppe, G. (2014). Understanding random forests: From theory to practice. Thèse soutenue à l'université de Liège.

[5] Breiman, L. (2001). Random forests. Machine learning 45(1):5–32.

[6] Meinshausen, N. (2006). Quantile regression forests. Journal of Machine Learning Research, 7:983–999.

[7] Ridgeway, G. (2007). Generalized Boosted Models: A guide to the gbm package.

[8] Freund, Y. and Schapire, R. and Abe, N. (1999). A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence 14(5):771–780.

[9] Fenske, N. and Kneib, T. and Hothorn, T. (2011). Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. Journal of the American Statistical Association 106(494):494–510

[10] Kriegler, B. and Berk, R. (2007). Boosting the quantile distribution: A cost-sensitive statistical learning procedure. Preprint.

[11] Kriegler, B. and Berk, R. (2010). Small area estimation of the homeless in Los Angeles: An application of cost-sensitive stochastic gradient boosting. The Annals of Applied Statistics 4(3):1234–1255