ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 194, 2005

Eugeniusz Gatnar[*]

GRADIENT BOOSTING IN REGRESSION

Abstract

The successful tree-based methodology has one serious disadvantage: lack of stability. That is, regression tree model depends on the training set and even small change in a predictor value could lead to a quite different model. In order to solve this problem single trees are combined into one model. There are three aggregation methods used in classification: bootstrap aggregation (bagging), adaptive resample and combine (boosting) and adaptive bagging (hybrid bagging-boosting procedure).

In the field of regression a variant of boosting, i.e. gradient boosting, can be used. Friedman (1999) proved that boosting is equivalent to a stepwise function approximation in which in each step a regression tree models residuals from last step model.

Key words: tree-based models, regression, boosting.

I. INTRODUCTION

The goal of a regression is to find a function $F^*(x)$ that maps x to y:

$$F^*(\mathbf{x}) : \mathbf{x} \rightarrow y, \tag{1}$$

and minimises the expected value of a specified loss function $L(y, F(x))$ over the joint distribution of all (x, y) values:

$$F^*(\mathbf{x}) = \underset{F(x)}{\arg\min}\, E_{y,x} L(y, F(\mathbf{x})), \tag{2}$$

given a sample (called "training set"):

[*] Professor, Institute of Statistics, University of Economics in Katowice.

$$(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), ..., (y_N, \mathbf{x}_N). \tag{3}$$

The most frequently used loss function for measuring errors between y and $F(x)$ is the squared error:

$$L(y, F(\mathbf{x})) = (y - F(\mathbf{x}))^2. \tag{4}$$

In this paper we consider $F(x)$ having an additive form:

$$F(\mathbf{x}) = \sum_{m=0}^{M} \beta_m f_m(\mathbf{x}, \mathbf{a}_m). \tag{5}$$

where $f_m(\mathbf{x}, \mathbf{a})$ is a simple function of $\mathbf{x}$ with parameters $\mathbf{a}$ (called "base learner"), for example the linear function:

$$f_m(\mathbf{x}, \mathbf{a}_m) = \sum_{l=1}^{L} a_m x_l. \tag{6}$$

When the base learner (6) is a tree, the parameters $\mathbf{a}$ are the spliting variables, split locations and mean values of y in regions R_k.

II. REGRESSION TREES

The tree corresponds to an additive model in the form of:

$$f(\mathbf{x}, \mathbf{a}) = \sum_{k=1}^{K} a_k I(x \in R_k), \tag{7}$$

where R_k are hyper-rectangular disjoint regions in the M-dimensional feature space, a_k denotes real parameters and I is an indicator function (Gatnar, 2001).

Each real-valued dimension of the region R_k is characterised by its upper and lower boundary: $v_{km}^{(d)}$ i $v_{km}^{(g)}$ respectively. Therefore the region induces a product of M indicator functions:

$$I(\mathbf{x} \in R_k) = \prod_{m=1}^{M} I(v_{km}^{(d)} \leqslant x_m \leqslant v_{km}^{(g)}), \tag{8}$$

If x_m is a categorical variable, the region R_k is defined as:

$$I(\mathbf{x} \in R_k) = \prod_{m=1}^{M} I(x_m \in B_{km}), \tag{9}$$

where B_m is a subset of the set of the variable values.

The parameter estimation formula depends on the way how the homogenity of the region R_k is measured. In the simplest case, when variance is used, the best estimate is the mean of all y values in R_k:

$$a_k = \frac{1}{N(k)} \sum_{i=1}^{N(k)} y_i, \tag{10}$$

where $N(k)$ is the number of objects from training set belong to region R_k.

Tree-based regression models are represented by step functions (Fig. 1).

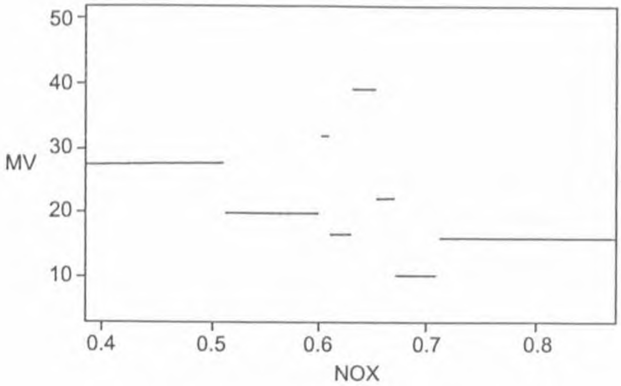

Figure 1. Example of a step function

Because their lack of smoothness could be sometimes a disadvantage, Friedman (1991) proposed to use splines (in the MARS procedure) to solve this problem.

III. BOOSTING

The successful tree-based methodology has one undesirable feature: lack of stability. That is a regression tree model depends on the training set and even small change in a predictor value could lead to a quite different model.

To solve this problem in the field of classification single trees are combined into one model and then averaged. There are three aggregation methods developed so far:

1) bootstrap aggregation (bagging), developed by Breiman (1996),

2) adaptive resample and combine (boosting), proposed by Freund and Shapire (1996),

3) adaptive bagging, proposed by Breiman (1999).

Boosting is seen as the most successful and powerful idea in statistical learning (Hastie et al., 2001). It was developed by Freund and Shapire (1996) originally for classification problems, to produce the most accurate model as a committee of many "weak" classifiers.

Given a set of training data (3) and classifier $f_m(\mathbf{x}, \mathbf{a})$ producing values from the set $\{-1, +1\}$, the algorithm *Ada.Boost* trains the classifier on modified training sample, giving higher weights to cases that are currently misclassified. This repeats for a sequence of weighted samples and the result is a linear combination of the classifiers from each stage.

The algorithm works as follows:

1. Start with equal weights for each case:

$$\underset{i=1,\ldots,N}{\forall} w_i = \frac{1}{N}, \tag{11}$$

2. Repeat for $m = 1$ to M:

a) fit the classifier: $f_m(\mathbf{x}, \mathbf{a})$ to the training data using weights w_i,

b) compute the classification error:

$$e_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq f_m(\mathbf{x}_i, \mathbf{a}))}{\sum_{i=1}^{N} w_i}, \tag{12}$$

c) compute the classifier weight:

$$\beta_m = \log\left(\frac{1 - e_m}{e_m}\right), \tag{13}$$

d) set weights for cases:

$$w_i \leftarrow w_i \cdot e^{\beta_m I(y_i \neq f_m(\mathbf{x}_i, \mathbf{a}))}, \tag{14}$$

3. Final classifier is:

$$F(\mathbf{x}) = \operatorname{sgn}\left(\sum_{m=0}^{M} \beta_m f_m(\mathbf{x}_i, \mathbf{a})\right). \tag{15}$$

In the step 2d) cases misclassified by $f_m(\mathbf{x}, \mathbf{a})$ have their weights increased and then they form the classifier $f_{m+1}(\mathbf{x}, \mathbf{a})$.

IV. GRADIENT BOOSTING

Friedman (1999) developed a variant of boosting, i.e. "gradient boosting" of trees which produces highly robust mod~ls, especially appropriate for imperfect data. He proved that boosting is equivalent to forward stepwise modelling, that is sequentially adding new functions to the expansion:

$$F(\mathbf{x}) = f_0(\mathbf{x}_i) + \beta_1 f_1(\mathbf{x}_i) + \beta_2 f_2(\mathbf{x}_i) + \dots \tag{16}$$

Using steepest-descent method from numerical minimisation, the negative gradient:

$$g_m(\mathbf{x}) = \left[\frac{\partial E_y(L(y, F(\mathbf{x}))|\mathbf{x})}{\partial F(\mathbf{x})}\right]_{F(\mathbf{x}) = F_{m-1}(\mathbf{x})} \tag{17}$$

define the "steepest-descent" direction:

$$f_m(\mathbf{x}) = -\lambda_m g_m(\mathbf{x}) \tag{18}$$

and:

$$F_{m-1}(\mathbf{x}) = \sum_{i=0}^{m-1} f_i(\mathbf{x}). \tag{19}$$

The weights λ_m in (18) are estimated as:

$$\lambda_m = \arg\min_{\lambda} E_{y,\mathbf{x}}\, L(y, F_{m-1}(x) + \lambda \cdot f_m(\mathbf{x})), \tag{20}$$

and the approximation updated:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \lambda_m \cdot f_m(\mathbf{x}). \tag{21}$$

For squared error loss function (4), or its minor modification:

$$L(y, F(\mathbf{x})) = \frac{1}{2}(y - F(\mathbf{x}))^2, \tag{22}$$

the negative gradient is just the residual:

$$-g_m = y_i - F_{m-1}(\mathbf{x}). \tag{23}$$

The gradient boosting is a stepwise function approximation in which each step models residuals from last step model.

If the base learner $f_m(\mathbf{x}, \mathbf{a})$ is a regression tree (7), then the boosted tree is induced according to the procedure:

1. Initialise:

$$F_0(\mathbf{x}) = \bar{y}. \tag{24}$$

2. For $m = 1$ to M:

a) repeat for each $i = 1, ..., N$:

$$u_i = y_i - F_{m-1}(\mathbf{x}). \tag{25}$$

b) grow regression tree for the residuals u_i finding homogeneous regions R_{jm},

c) for $j = 1, ..., J_m$ compute:

$$a_{jm} = \arg\min_{\mathbf{a}} \sum_{\mathbf{x}_j \in R_{jm}} (y_i - (F_{m-1}(\mathbf{x}_i) + a))^2. \tag{26}$$

d) modify:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \sum_{j=1}^{J_m} a_{jm} I(\mathbf{x} \in R_{jm}). \tag{27}$$

3. The final model:

$$F^*(\mathbf{x}) = F_M(\mathbf{x}). \tag{28}$$

The tree is grown to group observations into homogenous subsets. Once we have the subsets our update quantities for each subset are computed in a separate step.

V. EXAMPLE

Consider Boston Housing data set (Harrison and Rubinfeld, 1978). The data consisted of 14 variables measured for each of 506 census tracts in the Boston area. The dependent variable is MV – median of neighborhood home value and independent variables are: CRIM – crime rate, RM – average number of rooms, LSTAT – percent lower-status population, etc.

Average value of MV is $ 22.533. We start model $F_0(x)$ with the mean (24) and construct residuals. The residuals ~re computed with two-node tree and the tree separates positive from negative residuals.

Then we update the model, obtain new residuals and repeat the process (e.g. twice). Estimated function consists of three parts and is shown in Figure 2.

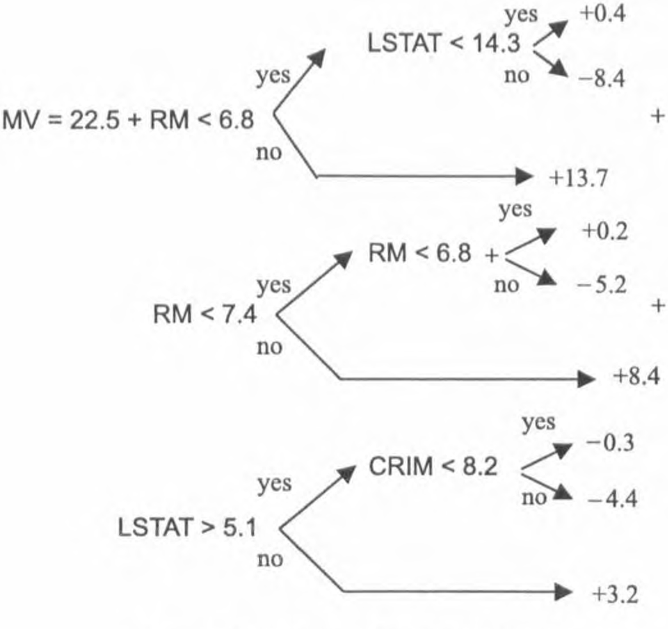

Figure 2. Boosting tree for Boston data

As we can see (Fig. 2) only three independent variables were selected to the model[1]: RM, LSTAT and CRIM.

The resulted boosting tree is better model, in terms of goodness-of-fit, than other regression models. In Table 1 we present a comparison of the value of R^2 for two data sets: Boston Housing and California Housing[2].

[1] We used the MART system implemented in the S-Plus environment.

[2] The California Housing data set is available from Statlib repository. It was analysed by Pace and Barry (1997) and consists of data from 20460 neighborhoods (1990 census block groups) in California.

Table 1. Comparison of R2 for two data sets

Data set	Single regression tree	Gradient boosting tree
Boston Housing	0.67	0.84
California Housing	0.70	0.86

VI. CONCLUSIONS

There are several advantages of using the method of gradient boosting in nonparametric regression. Boosting regression trees can cope with outliers, are invariant to monotone transformations of variables, and can handle missing values. They also automatically select variables to the model and perform regression very fast.

The regression model obtained in the form of a boosting tree is also extremely easy to interpret and to use for prediction.

REFERENCES

Breiman L. (1996), Bagging predictors, *Machine Learning*, **24**, 123–140.

Breiman L. (1999), Using adaptive bagging to debias regressions, *Technical Report*, **547**, Statistics Department, University of California, Berkeley.

Breiman L., Friedman, J., Olshen, R., Stone, C. (1984), *Classification and Regression Trees*, Wadsworth, Belmont, CA.

Freund Y., Schapire, R.E. (1997), A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences*, **55**, 119–139.

Friedman J.H. (1991), Multivariate adaptive regression splines, *Annals of Statistics*, **19**, 1–141.

Friedman J.H. (1999), *Greedy Function Approximation: a Gradient Boosting Machine*, Statistics Department, Stanford University, Stanford.

Gatnar E. (2001), *Nieparametryczna metoda dyploryminacji i regresji* (Nonparametric method for discrimination and regression; in Polish) PWN, Warszawa.

Harrison D., Rubinfeld, D.L. (1978), Hedonic prices and the demand for clean air, *Journal of Environmental Economics and Management*, **8**, 81–102.

Hastie T., Tibshirani, R., Friedman, J. (2001), *The Elements of Statistical Learning*, Springer, New York.

Pace R.K., Barry, R. (1997), Sparse spatial autoregressions, *Statistics and Probability Letters*, **33**, 291–297.

Eugeniusz Gatnar

GRADIENTOWA ODMIANA METODY *BOOSTING* W ANALIZIE REGRESJI

Streszczenie

Szeroko stosowane w praktyce metody nieparametryczne wykorzystujące tzw. drzewa regresyjne mają jedną istotną wadę. Otóż wykazują one niestabilność, która oznacza, że niewielka zmiana wartości cech obiektów w zbiorze uczącym może prowadzić do powstania zupełnie innego modelu. Oczywiście wpływa to negatywnie na ich trafność prognostyczną. Tę wadę można jednak wyeliminować, dokonując agregacji kilku indywidualnych modeli w jeden.

Znane są trzy metody agregacji modeli i wszystkie opierają się na losowaniu ze zwracaniem obiektów ze zbioru uczącego do kolejnych prób uczących: agregacja bootstrapowa (*boosting*), losowanie adaptacyjne (*bagging*) oraz metoda hybrydowa, łącząca elementy obu poprzednich.

W analizie regresji szczególnie warto zastosować gradientową, sekwencyjną, odmianę metody boosting. W istocie polega ona wykorzystaniu drzew regresyjnych w kolejnych krokach do modelowania reszt dla modelu uzyskanego w poprzednim kroku.