ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 228, 2009

*Dorota Rozmus**

UNBIASED RECURSIVE PARTITIONING ALGORITHM IN REGRESSION TREES

Abstract. Classification and regression trees are very popular tool for prediction. The construction of these models is based on recursive partitioning of multidimensional attribute space into disjoint homogenous regions till gaining the maximum homogeneity from the point of view of the dependent variable value.

The main aim of this research is to apply in regression trees unbiased recursive partitioning algorithm proposed by Hothorn, Hornik and Zeileis (2006), which is based on permutation tests. The research takes into consideration both single and aggregated approach and compare the results with classical method of tree model construction based on exhaustive search algorithm proposed by Breiman et al. (1984).

Key words: recursive partitioning, regression trees, aggregated models (ensembles), prediction.

I. INTRODUCTION

Classification and regression trees are widely used predictors. This method is based on recursive partitioning of multidimensional attribute space (containing set of observations) into disjoint regions (homogenous subsets) till gaining the maximum homogeneity from the point of view of the dependent variable value. This partitioning is made step by step on the base of chosen independent variables value (split points). These values are chosen in such a way that guarantees the improvement of model accuracy. The most often used method for choosing split points is exhaustive search algorithm, where the variable and split values are indicated at the same time. The fundamental problem of exhaustive search procedures have been known for a long time is a selection bias towards covariates with many possible splits or missing values. But it appears that this failure may be omitted by separating the variable choosing phase from the split point choosing phase. This approach applies statistical tests.

* Ph.D., Department of Statistics, The Karol Adamiecki University of Economics, Katowice.

In the literature there were proposed many different methods based on statistical tests. Eg. in 1980 Kaas has proposed CHAID algorithm where χ^2 statistics was applied; in 1988 Loh and Vanichsetakul introduced FACT algorithm where covariates are selected within an analysis of variance framework (ANOVA) or QUEST algorithm introduced by Loh and Shih (1997) where F and χ^2 statistics are used.

The main aim of this article is to present a new algorithm for recursive partitioning based on permutation tests that was introduced by Hothorn, Hornik and Zeileis (2006) and is called unbiased recursive partitioning. In this approach the conditional distribution of test statistics measuring the association between responses and covariates is the basis for an unbiased selection among covariates measured at different scales. Moreover, multiple test procedures are applied to determine whether no significant association between any of the covariates and the response can be stated and the recursion needs to stop.

II. UNBIASED RECURSIVE PARTITIONING ALGORITHM

We focus on regression models describing the conditional distribution of a response variable Y given the status of m covariates $\mathbf{X} = (X_1,...,X_m)$ by means of tree-structured recursive partitioning. Both response variable and covariates may be measured at arbitrary scales. The response Y may be multivariate as well. We assume that the conditional distribution $D(\mathbf{Y}|\mathbf{X})$ of the response Y given the covariates X depends on a function f of the covariates:

$$D(\mathbf{Y}|\mathbf{X}) = D(\mathbf{Y}|X_1,...,X_m) = D(\mathbf{Y}|f(X_1,...,X_m)). \quad (1)$$

A regression model of the relationship is to be fitted based on a learning sample $L_n = \{(\mathbf{Y}_i, X_1,...,X_m); i = 1,...,n\}$, i.e., a random sample of n independent and identically distributed observations, possibly with some covariates X_{ji} missing.

A generic algorithm for recursive binary partitioning for a given learning sample L_n can be formulated using non-negative integer valued case weights $\mathbf{w} = (w_1,...,w_n)$. Each node of a tree is represented by a vector of case weights having non-zero elements when the corresponding observations are elements of the node and are zero otherwise. The following generic algorithm implements unbiased recursive binary partitioning:

1. For case weights **w** test the global null hypothesis of independence between any of the m covariates X and the response Y. Stop if this hypothesis can-

not be rejected. Otherwise select the j^* th covariate X_{j^*} with strongest association to Y.

2. Choose a set $A^* \subset X_{j^*}$ in order to split X_{j^*} into two disjoint sets A^* and $A^* \setminus X_{j^*}$. The case weights $\mathbf{w}_L$ and $\mathbf{w}_P$ determine the two subgroups with: $\mathbf{w}_{L,i} = w_i I(X_{j^*,i} \in A^*)$ and $\mathbf{w}_{P,i} = w_i I(X_{j^*,i} \notin A^*)$ for $I = 1,\ldots, m$; where $I(\cdot)$ denotes the indicator function.

3. Repeat recursively steps 1 and 2 and modify case weights $\mathbf{w}_L$ and $\mathbf{w}_P$, respectively. The algorithm stops when the global null hypothesis of independence between the response and any of the m covariates cannot be rejected at a pre-specified nominal level α.

The main idea of presented approach is included in step 1 of the generic algorithm. Unified tests for independence are constructed by means of the conditional distribution of multivariate linear statistics in the permutation test framework developed by Strasser and Weber (1999). The determination of the best binary split point of the chosen covariate is performed based on standardized linear statistics within the same framework as well.

At step 1 of the generic algorithm we face an independence problem. We should decide whether there is any dependence between the response variable and any of the m covariates. In each node identified by case weights $\mathbf{w}$, the global hypothesis of independence is formulated in terms of the m partial hypotheses $H_0^j : D(\mathbf{Y} \mid X_j) = D(\mathbf{Y})$ with global null hypothesis: $H_0 = \bigcap_{j=1}^{m} H_0^j$. When we are not able to reject H_0 at a pre-specified level α, we stop the recursion. If the global hypothesis can be rejected, we measure the association between Y and each of the covariates X_j, $j = 1, 2, \ldots, m$, by test statistics. For notational convenience and without loss of generality we assume that the case weights w_i are either zero or one. Let $S(L_n, \mathbf{w})$ denotes the symmetric group of all permutations of the elements of $(1, 2,\ldots,n)$ with corresponding case weights w_i. We measure the association between Y and X_j $j = 1,\ldots, m$, by linear statistics of the form:

$$T_j(L_n, w) = vec\left(\sum_{i=1}^{n} w_i g_i(X_{ji}) h(\mathbf{Y}_i)^T\right) \in \quad ^{\times}, \qquad (2)$$

where: g_j : is a non-random transformation of the covariate X_j, and h is an influence function. The most often chosen forms of g_j and h functions are identity function and for nominal variables taking k = 1, 2, ..., K values - the unit vector of length K with kth element being equal to one. As a result a $p_j \times q$ matrix is converted into a $p_j q$ column vector by column-wise combination using the '*vec*' operator.

The distribution of $T_j(L_n, w)$ under H_0^j depends on the joint distribution of Y and X_j and is unknown under almost all practical circumstances. Assuming that the null hypothesis is true one can dispose of this dependency by fixing the covariates and conditioning on all possible permutations of the response variable values. This principle leads to test procedure known as permutation test. The conditional expectation $\mu_j \in$ $^{\times}$ and covariance $\Sigma_j \in$ $^{\times}$ of $T_j(L_n, w)$ under H_0, given all permutations $\sigma \in S(L_n, \mathbf{w})$ of the responses are derived by Strasser and Weber (1999):

$$\mu_j = \left(\mathbf{T}_j(L_n, \mathbf{w}) \mid S(L_n, \mathbf{w})\right), \tag{3}$$

$$\Sigma_j = \left(\mathbf{T}_j(L_n, \mathbf{w}) \mid S(L_n, \mathbf{w})\right). \tag{4}$$

The conditional expectation μ_j and covariance Σ_j is used to standardize a multivariate linear statistic $T \in$ $^{\times}$ of the form (1) in order to get a scalar. Univariate test statistics c mapping an observed statistic $T \in$ $^{\times}$ into scalar can be calculated as:

$$c_{\max}(\mathbf{t}, \mu, \Sigma) = \max_{k=1,2,\ldots,pq} \left| \frac{(\mathbf{t} - \mu)_k}{\sqrt{(\Sigma)_{kk}}} \right|. \tag{5}$$

In step 1 of the generic algorithm we select the covariate with minimum p-value, i.e., the covariate X_{j^*} with $j^* = \arg\min_{j=1,2,\ldots,m} P_j$, where:

$$P_j = \left(c\left(\mathbf{T}_j(L_n, \mathbf{w}), \mu_j, \Sigma_j\right) \geq c\left(\mathbf{t}_j, \mu_j, \Sigma_j\right) \mid S(L_n, \mathbf{w})\right) \tag{6}$$

Once we have selected a covariate in step 1 of the algorithm, the split point can be established by utilizing the permutation test framework described above to find the optimal binary split in one selected covariate X_{j^*} in step 2 of the generic algorithm. The accuracy of a split is evaluated by two-sample linear statistics which are special cases of the linear statistic (1). For all possible subsets A of the sample space X_{j^*} the linear statistic:

$$\mathbf{T}_{j^*}^{A}(L_n, \mathbf{w}) = vec\left(\sum w_i I(X_{j^* i} \in A) h(\mathbf{Y}_i, (\mathbf{Y}_1, ..., \mathbf{Y}_n))^T \right) \in \tag{7}$$

measures the discrepancy between the samples $\{\mathbf{Y}_i \mid w_i > 0 \text{ and } X_{ji} \in A, i = 1, ..., n\}$; and $\{\mathbf{Y}_i \mid w_i > 0 \text{ and } X_{ji} \notin A, i = 1, ..., n\}$. The conditional expectation $\mu_{j^*}^{A}$ and covariance $\Sigma_{j^*}^{A}$ can be computed according to (3) and (4). The split A^* with a test statistic maximized over all possible subsets A is established as:

$$A^* = \arg\max_{A} c\left(\mathbf{t}_{j^*}^{A}, \mu_{j^*}^{A}, \Sigma_{j^*}^{A}\right). \tag{8}$$

III. EMPIRICAL RESULTS

In my empirical part of this research I wanted to compare results with using regression trees built by *rpart* method that uses exhaustive search algorithm for construction of the model and *ctree* method based on permutation tests. The trees were applied in single and aggregated approach. In order to get an aggregated model (ensemble) in first step we build many different single models and then we combine them by means of some aggregation operator. In regression the most popular operator is taking mean of all theoretical values of dependent variable. As ensembles I used *bagging* (Breiman, 1996) and *random forest* (Breiman, 2001) algorithms. Among used there are six real and three artificial data sets (Friedman 1–3) that are widely used in comparative researches.

Table 1. Results for *rpart* and *ctree* method in regression

Method / Data set	Coefficient of determination R^2 (in %)					
	single model		random forest		bagging	
	rpart	Ctree	rpart	ctree	rpart	ctree
Boston	75.4	78.2	88.9	81.6	81.1	82.4
Real estate	74.1	75.1	74.9	59.4	72.9	76.3
Budgets	57.9	58.4	66.4	56.6	59.5	64.3
Housing	45.6	48.3	65.3	60.7	54.2	58.7
BudgetItaly	89.6	76.1	95.4	78.7	88.5	92.6
Electricity	79.9	80.7	94.6	83.6	89.3	89.7
Friedman1	61.7	60.4	75.7	81.6	81.1	72.0
Friedman2	79.8	80.0	79.2	49.3	84.3	85.2
Friedman3	62.7	62.8	73.8	56.3	72.0	73.6

Source: own computations.

Looking at the results for single models we can see that generally *ctree* method gives slightly better results for almost every data set. Higher differences in R^2 value we can observe in aggregated approach. For *random forest* it is better to use the classical *rpart* trees than *ctree*. Moreover applying trees based on permutation tests in this method of aggregation can causes that the results will be even worse than for single models. From the results it appears also that in the case of *bagging* method it is better to aggregate trees based on algorithm proposed by Hothorn, Hornik and Zeileis.

IV. SUMMARY

To sum up we can say that proposed unbiased recursive partitioning algorithm the best results gives when is used for construction of base models for *bagging* aggregation method. In the case of single models or *random forest* aggregation method better results we gain with models constructed in a traditional way that is with using exhaustive search algorithm.

It is also worth to say, that the main advantage of the algorithm proposed by Hothorn, Hornik and Zeileis is higher objectiveness in choosing the most important variables. So this algorithm should be used in situations when we are interested mainly in finding variables with the highest discrimination power.

REFERENCES

Breiman L. (1996), *Bagging predictors*, „Machine Learning", 26 (2), 123–140.
Breiman L. (2001), *Random forests*, „Machine Learning" 45, 5–32.
Breiman L., Friedman J., Olshen R., Stone C. (1984), *Classification and regression trees*, CRC Press, London.
Hothorn T., Hornik K., Zeileis A. (2006), Unbiased Recursive Partitioning: A Conditional Inference Framework, *Journal of Computational and Graphical Statistics*, 15 (3), 651–674.
Kass G. (1980), An Exploratory Technique for Investigating Large Quantities of Categorical Data, *Applied Statistics*, 29 (2), 119–127.
Loh W. Y., Shih Y. S. (1997), Split Selection Methods for Classification Trees, *Statistica Sinica*, 7, 815–840.
Loh W.Y., Vanichsetakul N. (1988), Tree-Structured Classification via Generalized Discriminant Analysis, *Journal of the American Statistical Association*, 83, 715–725.
Strasser H., Weber C., (1999), On the Asymptotic Theory of Permutation Statistics, *Mathematical Methods of Statistics*, 8, 220–250.

Dorota Rozmus

ZASTOSOWANIE NIEOBCIĄŻONEJ METODY REKURENCYJNEGO PODZIAŁU W METODZIE DRZEW REGRESYJNYCH

Drzewa klasyfikacyjne i regresyjne są bardzo popularnym narzędziem predykcji. Budowa takiego modelu polega na stopniowym podziale wielowymiarowej przestrzeni cech na rozłączne obszary aż do uzyskania maksymalnej ich homogeniczności ze względu na wartość zmiennej objaśnianej *y*. Podział ten kontynuowany jest w kolejnych krokach, w oparciu o wartości wybranych zmiennych objaśniających. Istnieje wiele możliwych sposobów wyboru tych zmiennych, a jednym z najpopularniejszych jest algorytm wyczerpującego przeszukiwania (ang. *exhaustive search*) opracowany przez Breimana (Breimana et al., 1984).

Zaproponowany przez Hothorna, Hornika i Zeileisa, (2006) sposób doboru zmiennych znany pod nazwą nieobciążonej metody rekurencyjnego podziału (ang. *unbiased recurcive partitioning*) opierający się na zastosowaniu testów permutacyjnych miał na celu ominięcie podstawowej wady tradycyjnego podejścia, jakim jest tendencja do wyboru zmiennych dających wiele potencjalnych możliwości podziału.

Okazuje się, że w przypadku dyskryminacji to nowatorskie podejście prowadzi do uzyskania modeli zapewniających bardzo zbliżone wyniki klasyfikacji jak podejście tradycyjne, a w podejściu wielomodelowym może doprowadzić do pogorszenia poprawności klasyfikacji.

Zasadniczym celem referatu jest przedstawienie wyników badań, które mają na celu porównanie dokładności predykcji na podstawie drzew regresyjnych, które doboru zmiennych objaśniających dokonują za pomocą algorytmu wyczerpującego przeszukiwania oraz za pomocą podejścia bazującego na testach permutacyjnych. Ponadto porównane zostaną wyniki predykcji modeli zagregowanych, w których modelami składowymi będą te dwa rodzaje drzew regresyjnych.