ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 194, 2005

*Malgorzata Misztal**

ON THE APPLICATION OF CLASSIFICATION AND REGRESSION TREES IN MEDICAL DIAGNOSIS

Abstract

Decision tree is a graphical presentation of the recursive partitioning the learning set into homogenous subsets considering dependent variable y.

If dependent variable y is nominal we deal with nonparametric discriminant analysis (classification trees), when y is numerical – with nonparametric regression analysis (regression trees).

The aim of the paper is to present some applications of regression and classification trees in medical diagnosis for solving decision – making problems.

Key words: classification and regression trees, medical diagnosis.

I. INTRODUCTION

Decision tree can be described as a tree-like way of representing a collection of hierarchical rules that lead to a class or to a value.

Let us consider a learning set:

$$U = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_N, y_N)\} \tag{1}$$

where $\mathbf{x}$ is the vector of independent variables $\mathbf{x} = [x_1, x_2, ..., x_p]^T$ and y is the response (dependent) variable.

The model building process is based on recursive partitioning the learning set into homogenous subsets $U_1, U_2, ..., U_M$ considering dependent variable y.

Tree-based models are simple, flexible and powerful tools for classification and regression analysis, dealing with different kinds of variables, including missing values and very easy to interpret.

* Ph.D., Chair of Statistical Methods, University of Łódź.

II. MODEL BUILDING PROCESS

We consider an additive model:

$$y = \alpha_0 + \sum_{m=1}^{M} \alpha_m g_m(\mathbf{x}, \beta), \tag{2}$$

where $g_m(\mathbf{x}, \beta)$ are functions of $\mathbf{x}$ with parameters β.

An approximation of (2) can be written as:

$$y = a_0 + \sum_{m=1}^{M} a_m I\{\mathbf{x} \in R_m\}, \tag{3}$$

where R_m $(m = 1, ..., M)$ are disjoint regions in the p-dimensional feature space, a_m are real parameters and $I\{A\}$ is an indicator function:

$$I\{A\} = \begin{cases} 1, & \text{if the proposition inside the brackets is true} \\ 0, & \text{otherwise} \end{cases}. \tag{4}$$

For real-valued dimension of the region R_m, characterized by its upper and lower boundary $x_r^{(d)}$ and $x_r^{(g)}$, we have:

$$I\{\mathbf{x} \in R_m\} = \prod_{r=1}^{p} (x_{mr}^{(d)} \leqslant x_r \leqslant x_{mr}^{(g)}). \tag{5}$$

For each categorical variable x_r we have:

$$I\{\mathbf{x} \in R_m\} = \prod_{r=1}^{p} I\{r_r \in B_{mr}\}, \tag{6}$$

where B_{mr} is a subset of the set of the variable values (see Gatnar, 2001).

If dependent variable y is nominal we deal with nonparametric discriminant analysis (so we have classification trees); when y is numerical – with nonparametric regression analysis (regression trees).

In discriminant analysis the response variable y represents a class label, so that we want to predict the class of an object from the values of its predictor variables.

In regression analysis the response variable y is assumed to depend on the regressors $x_1, x_2, ..., x_p$ through the relationship:

$$y = f(\mathbf{x}, \boldsymbol{\alpha}) + \varepsilon, \tag{7}$$

where ε is a noise component in the model.

The goal of the regression analysis is to find an estimate $\hat{y}$ of y that minimizes a certain loss function (e.g. squared error loss). The estimation of y can be done by recursive partitioning the data and the regressor space using techniques that yield piecewise constant estimate of y:

$$f(\mathbf{x}) = \sum_{m=1}^{M} \theta_m I\{\mathbf{x} \in R_m\}, \tag{8}$$

so that for each $\mathbf{x} \in R_m$:

$$f_m(\mathbf{x}) = \theta_m. \tag{9}$$

The procedure of the partitioning of the sample space requires selecting appropriate variables.

The goal of the variable selection process is to choose those variables that yield the best partition. The quality of partition of a learning set U is measured by the difference between heterogeneity (in terms of y) of U and of resulted subsets $U_1, U_2, ..., U_M$ (see Gatnar, 2001). In other words, we have a reduction of impurity:

$$\Delta H(U; x_r) = H(U) - \sum_{m=1}^{M} H(U_m)p(m), \tag{10}$$

where:

H (*) – is a heterogeneity measure,

$p(m) = \dfrac{N(m)}{N}$ – is the proportion of objects in U_m,

$N(m)$ – is the number of objects in U_m; of course $\sum_{m=1}^{M} N(m) = N$.

If y is categorical we have for example entropy function:

$$H(U_m) = -\sum_{i=1}^{k} p(i/m)\log_2 p(i/m) \tag{11}$$

or Gini index:

$$H(U_m) = 1 - \sum_{i=1}^{k} p^2(i/m), \tag{12}$$

where $p(i/m) = \dfrac{N_i(m)}{N(m)}$; $N_i(m)$ – is the number of objects from class i in U_m.

In regression trees analysis the most frequently used heterogeneity measure is variance:

$$s^2(U) = \frac{1}{N}\sum_{n=1}^{N}[y_n - f(\mathbf{x})]^2. \tag{13}$$

Using (13) we have:

$$\Delta H(U; x_r) = s^2(U) = -\sum_{m=1}^{M} s^2(U_m)p(m) \tag{14}$$

and

$$\hat{\theta}_m = \frac{1}{N(m)}\sum_{n=1}^{N(m)} y_n. \tag{15}$$

An ideal goal of recursive partitioning is to find a decision tree of minimal size and maximal predictive accuracy. To evaluate the model accuracy a test set is required. If not, cross – validation is used.

The right size of the tree can be chosen using some pruning methods. The usual procedure is to choose optimal splits all the way back down the tree so that as an intermediary step one obtains a tree with as many leaves as observations. Then the tree is pruned back using a function that gives a cost to complexity (see Breiman et al., 1984):

$$\mathrm{R}_\alpha(\mathrm{T}) = \mathrm{R}(\mathrm{T}) + \alpha \times \mathrm{size}(\mathrm{T}), \tag{16}$$

where size(T) is some measure of the complexity of the tree such as its size (= number of leaves), α is the cost-complexity parameter and R(T) is a risk function that penalizes for bad prediction.

The optimal tree is determined for example by the 1-SE rule proposed by Breiman et al., 1984. Then, it is the smallest tree with error not more than one standard error greater than the lowest error tree in the sequence.

Tree-structured models have many practical applications. One of them is medical diagnosis where the learning sample consists of case records containing a description of patient's symptoms and corresponding outcome.

III. APPLICATION: PREDICTION OF DURATION OF ICU STAY AFTER CABG

In the following example we analyse some data from the Department of Cardiothoracic Surgery of Lodz Medical University, where the set of 244 case records of patients undergoing CABG (Coronary Artery Bypass Grafting) during 1997–1998 was collected.

Dependent variable y is the duration of a patient's stay in the Intensive Care Unit (ICU) – in days. For classification trees analysis length of stay is categorized as follows: class 1 – ICU stay 1–4 days; class 2 – ICU stay 5 or more days. Deaths are excluded.

Predictor variables are the following:

1) sex;
2) age in years;
3) BMI – body mass index;
4) BSA – body surface area;
5) diabetes mellitus (yes/no);
6) chronic pulmonary diseases (yes/no);
7) AO – arterial obstruction (yes/no);
8) hyperthyroidism (on medication) (yes/no);
9) diagnosis (stable angina, unstable angina);
10) CCS – Canadian Coronary Score (class I, II, III, IV);
11) history of myocardial infarction;
12) previous cardiac surgery (yes/no);
13) left main stenosis > 75% (yes/no);
14) EF% – left ventricular ejection fraction in %;
15) AspAt – aspartate aminotransferase in U/L;
16) priority of operation (elective, urgent, emergent);
17) intraoperative myocardial infarction or low cardiac output syndrome (yes/no);
18) perfusion time in minutes;
19) aortic clamping time in minutes;
20) reperfusion time in minutes;
21) severity score based on preoperative risk factors (see Tab. 1).

The following algorithms are used:

1) CART (Classification and Regression Trees) by Breiman et al. (1984);
2) QUEST (Quick, Unbiased, Efficient Statistical Trees) by Loh and Shih (1997);
3) GUIDE (Generalized, Unbiased Interaction Detection and Estimation) by Loh (2002).

CART and QUEST are well known algorithms, available for example in STATISTICA PL package.

GUIDE is a new algorithm for building piecewise constant or piecewise linear regression models with univariate splits. It has four useful properties: (i) negligible bias selection, (ii) sensitivity to curvature and local pairwise interactions between regressor variables, (iii) inclusion of categorical predictor variables, including ordinal categorical variables, (iv) choice of three roles for each ordered predictor variable: split selection only, regression modelling only, or both. For more details see Loh (2002).

Table 1. Clinical Severity Scoring System

Risk Factors	Score
Left ventricular ejection fraction $< 40\%$	3
Emergency case	3
Age $\geqslant 60$	1 (+ 1 point per 5 years)
Hyperthyroidism (on medication)	2
Diabetes mellitus	2
Previous cardiac surgery	2
Chronic pulmonary diseases	2
Unstable angina	2
BSA $< 1{,}75\ m^2$	2
AspAt $\geqslant 40$ U/L	1
Creatinin level $> 1{,}2$ mg/dl	1
Arterial obstruction	1
Left main stenosis $> 75\%$	2
Unstable hemodynamic state	4

Source: Elaborated by Department of Cardiothoracic Surgery of Łódź Medical University and Chair of Statistical Methods, University of Łódź.

The results of using selected tree building methods are shown in Figures 1–4.

Figure 1 shows the piecewise constant GUIDE tree. 0-SE tree has 8 terminal nodes. 5 variables appear in the splits: priority of operation, body surface area, age, severity score and sex. The number in each terminal node is the sample mean of the ICU stay. Predicted mean squared error is 5.245.

The piecewise constant GUIDE tree from quantile (median) regression model is shown in Figure 2. It is shorter than the least squares tree in Figure 1. 0-SE tree has 3 terminal nodes. The number in each terminal node is the sample median of the ICU stay. The splitting variables are: priority of operation and body surface area. Predicted mean squared error equals 5.893.

Figures 3 and 4 show trees obtained using the classification trees algorithms (QUEST and CART respectively).

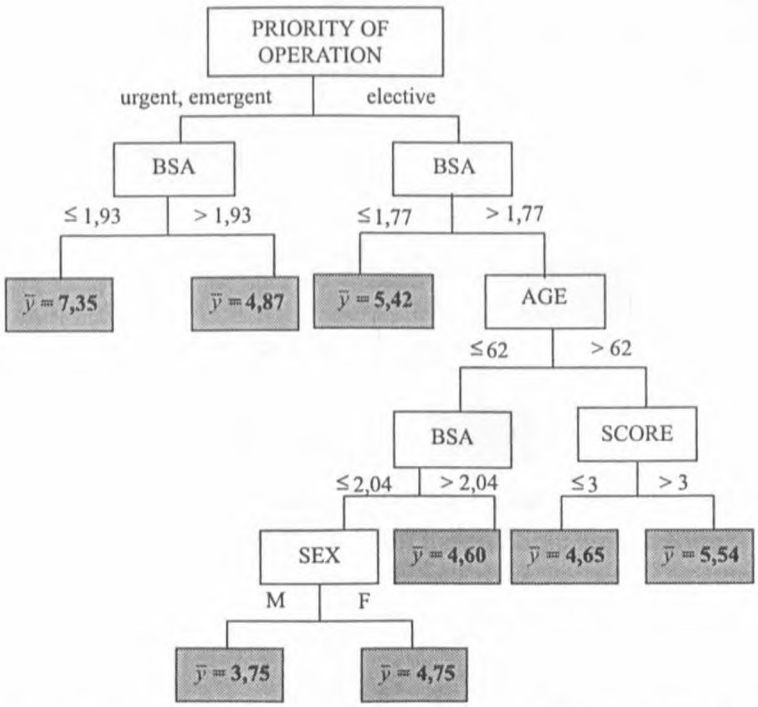

Figure 1. GUIDE regression tree – least squares regression, piecewise constant model
Source: Author's calculations

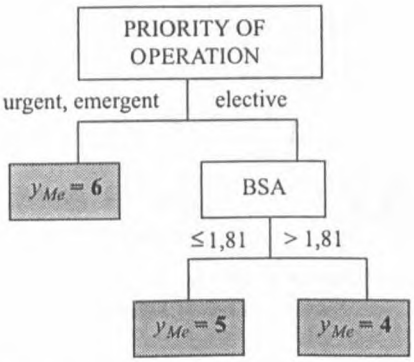

Figure 2. GUIDE regression tree – median regression, piecewise constant model
Source: Author's calculations

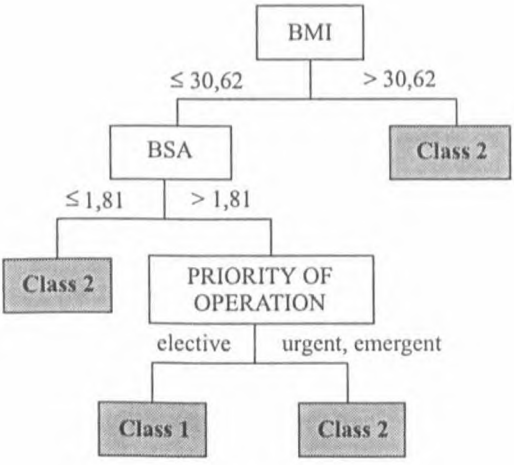

Figure 3. QUEST classification tree
Source: Author's calculations

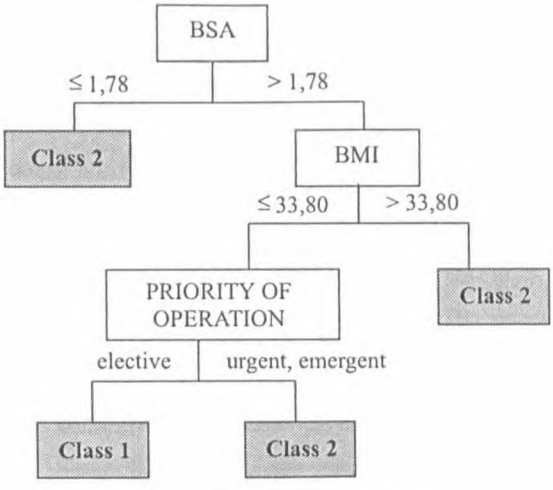

Figure 4. CART classification tree
Source: Author's calculations

Both: CART and QUEST 0-SE trees have 4 terminal nodes. Trees are quite similar. Three predictor variables may be sufficient to predict the response: body surface area, body mass index and priority of operation. Cross-validation estimate of misclassification error is 0.4303 for the QUEST tree and 0.3934 for the CART tree.

IV. CONCLUSIONS

Prediction of the length of stay in the Intensive Care Unit after cardiac surgery is not easy. Prolonged stay in the ICU increases the overall costs of cardiac surgery and may also limit the number of operations performed. Therefore, the ability to accurately predict the length of stay in the ICU seems to be very important (see Michalopoulos et al., 1996).

In order to predict the duration of the patient's stay in the ICU we propose to use tree-structured methods. Advantages of tree-based models are, among others:

1) no requirement of knowledge of the variable distribution;

2) dealing with different types of variables – both: categorical and continuous, including missing values;

3) robustness to outliers;

4) direct and intuitive way of interpretation;

5) reduction of the cost of the research by selecting only some important variables for splitting nodes, so that each new object can be described by a few risk factors.

Interpretability of a tree structure decreases with increase in its complexity so that in further researches we will pay more attention to fitting piecewise linear models which tend to possess better prediction accuracy.

REFERENCES

Breiman L., Friedman J., Olshen R., Stone C. (1984), *Classification and Regression Trees*, CRC Press, London.

Domański Cz., Pruska K., Wagner W. (1998), *Wnioskowanie statystyczne przy nieklasycznych założeniach*, Wyd. UŁ, Łódź.

Gatnar E. (2001), *Nieparametryczna metoda dyskryminacji i regresji*, PWN, Warszawa.

Loh W.-Y. (2002), Regression trees with unbiased variable selection and interaction detection, *Statistica Sinica*, **12**, 361–386.

Loh W.-Y., Shih Y.-S. (1997), Split selection methods for classification trees, *Statistica Sinica*, 7, 815–840.

Michalopoulos A., Tzelepis G., Pavlides G., Kriars J., Dafni U., Geroulanos S. (1996), Determinants of duration of ICU stay after coronary artery bypass graft surgery, *British Journal of Anaesthesia*, 77, 208–212.

Małgorzata Misztal

O ZASTOSOWANIU DRZEW KLASYFIKACYJNYCH I REGRESYJNYCH W DIAGNOSTYCE MEDYCZNEJ

Streszczenie

Drzewo decyzyjne jest graficzną prezentacją metody rekurencyjnego podziału. Metoda ta polega na stopniowym podziale zbioru obiektów na rozłączne podzbiory aż do momentu uzyskania ich jednorodności ze względu na wyróżnioną cechę y.

Gdy y jest zmienną nominalną, mamy do czynienia z nieparametryczną analizą dyskryminacji (drzewa klasyfikacyjne), gdy zaś jest zmienną ilościową – z nieparametryczną analizą regresji (drzewa regresyjne).

W referacie przedstawiono możliwości zastosowań drzew regresyjnych i klasyfikacyjnych do rozwiązywania problemów o charakterze decyzyjnym w diagnostyce medycznej.