ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 206, 2007

*Małgorzata Misztal**

A PROPOSAL FOR USING SELECTED TREE-BASED MODELS TO IDENTIFY OPERATIVE RISK SUBGROUPS AMONG PATIENTS UNDERGOING CORONARY ARTERY BYPASS GRAFTING

Abstract. Classification and regression trees are very popular and attractive types of classifiers, widely used to solve decision-making problems in different fields of science.

The study was conducted to identify preoperative risk factors associated with morbidity outcome among patients undergoing isolated Coronary Artery Bypass Grafting (CABG) and to develop some classification rules assigning patients to selected risk subgroups. Prediction rules were established on the basis of the selected tree-structured models. The following tree-based algorithms were used: QUEST, CRUISE, LOTUS and PLUS.

Key words: recursive partitioning method, classification and regression trees, coronary artery disease, coronary artery bypass grafting.

1. INTRODUCTION AND OBJECTIVES

The decision to perform coronary artery bypass grafting (CABG) surgery on a patient is taken under conditions of uncertainty. In that case the benefits of CABG must be balanced against its risk. To estimate this risk we must simultaneously consider many types of information including characteristics of the patient and characteristics of the disease.

The main goal of the study was to identify factors associated with morbidity outcome among patients undergoing CABG and to develop decision rules for the classification of patients into selected risk subgroups. Prediction rules were established on the basis of tree-structured models.

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.

We consider a learning set $U = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$, where $\mathbf{x}$ is the vector of independent variables $\mathbf{x} = [x_1, x_2, ..., x_p]^T$ and y is the response

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

(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. If y is nominal we deal with nonparametric discriminant analysis (classification tree), when y is numerical – with nonparametric regression analysis (regression tree) (see e.g. Breiman *et al.* 1984; Gatnar 2001).

In medical diagnosis tasks vector **x** consists of variables describing patient's symptoms, characteristics of the disease and the state of the patient before and during the treatment. The response variable y, in our study, is the number of the class (risk subgroup) the patient belongs to.

2. MATERIAL AND METHODS

The set of 2568 case records of patients undergoing CABG during 2003–2004 in Poland were analysed. The data from 2003 ($N = 947$) constituted the learning set and from 2004 ($N = 1621$) – the test set.

Only preoperative risk factors were taken into account. 37 predictor variables were evaluated. Three clinical scoring systems: EuroSCORE (Nashef *et al.* 1999), Cleveland Clinic Foundation (Higgins *et al.* 1992) and lately created Łódź Score of Surgical Risk (Domański *et al.* 2003, see: Tab. 1) were also taken into consideration.

Table 1

Łódź Clinical Scoring System

Risk Factors	Score
EF < 40%	3
Emergency case	3
Age ⩾ 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 ⩾ 40 U/L	1
Creatinin level > 1.2 mg/dl	1
Arterial obstruction	1
Left main stenosis > 75%	2
Unstable haemodynamic state	4

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

The outcome after CABG included the following 2 classes:

1) class 0 – with uncomplicated postoperative outcome (629 patients in the learning set and 922 patients in the test set);

2) class 1 – patients with one or more of the following: i) deaths; ii) cardiac complications; iii) central nervous system complications, iv) renal failure, v) respiratory failure, vi) any serious infection (318 cases in the learning sample and 699 cases in the test sample).

The potential association of each of the considered factors with the postoperative outcome was calculated using χ^2 test or Mann–Whitney's test. Factors significant to at least $p < 0.10$ were used to establish classification rules to identify the high-risk subgroup. Statistical analyses were performed with STATISTICA PL Software ver. 6.0.

The following tree–based algorithms were used:

1) QUEST (Quick, Unbiased, Efficient Statistical Trees) described in W.-Y. Loh and Y.-S. Shih (1997) – designed to have unbiased variable selection in the splitting procedures (obtained from: http://www.stat.wisc.edu/~Loh/quest.html);

2) CRUISE (Classification Rule with Unbiased Interaction Selection and Estimation) described in H. Kim and W.-Y. Loh (2001) – with an interaction detection methods in the splitting process (obtained from: http://www.wpi.edu/~hkim/cruise/).

3) LOTUS (Logistic Regression Trees with Unbiased Selection) described in K.-Y. Chan and W.-Y. Loh (2004) – designed to fit a piecewise (multiple or simple) linear logistic regression model by recursively partitioning the data and fitting a different logistic regression in each partition (obtained from: http://www.stat.nus.sg/~kinyee/lotus.html).

4) PLUS (Polytomous Logistic Regression Trees with Unbiased Split) described in T.-S. Lim (2000) – which combines a polytomous logistic regression with tree-based models (obtained from: http://www/recursive-partitioning.com/plus).

We used the learning set to develop some decision rules for the classification and the test set to evaluate the model accuracy.

3. RESULTS

The following 19 risk factors were significantly associated with morbidity outcome:

- sex ($p < 0.01$);
- diabetes mellitus ($p < 0.05$);

- age ($p < 0.001$);
- BSA (body surface area, $p < 0.001$);
- BMI (body mass index, $p < 0.05$);
- unstable angina ($p < 0.001$);
- recent (< 90 days) myocardial infraction ($p < 0.001$);
- mitral regurgitation ($p < 0.01$);
- EF (left ventricular ejection fraction, $p < 0.001$);
- anticoagulation and/or antiplatelet treatment ($p < 0.05$);
- cleveland Clinic Foundation Score ($p < 0.001$);
- Łódź Clinical Scoring System ($p < 0.001$);
- carotid arteries arteriosclerosis – symptomatic TIA ($p < 0.01$);
- preoperative hematocrit level ($p < 0.10$);
- critical preoperative state (at least one of: preoperative cardiac massage, preoperative intubation, preoperative intra-aortic balloon; $p < 0.01$);
- unstable haemodynamic state ($p < 0.001$);
- priority of operation ($p < 0,001$);
- EuroSCORE ($p < 0.001$);
- type of operation (in extracorporeal circulation – ECC or off-pump operation – without ECC; $p < 0.10$).

Risk factors mentioned above were employed in tree-structured analysis. All the results are shown in Fig. 1–4. Tables 2–3 present details of terminal node models of logistic regression trees for LOTUS and PLUS algorithms.

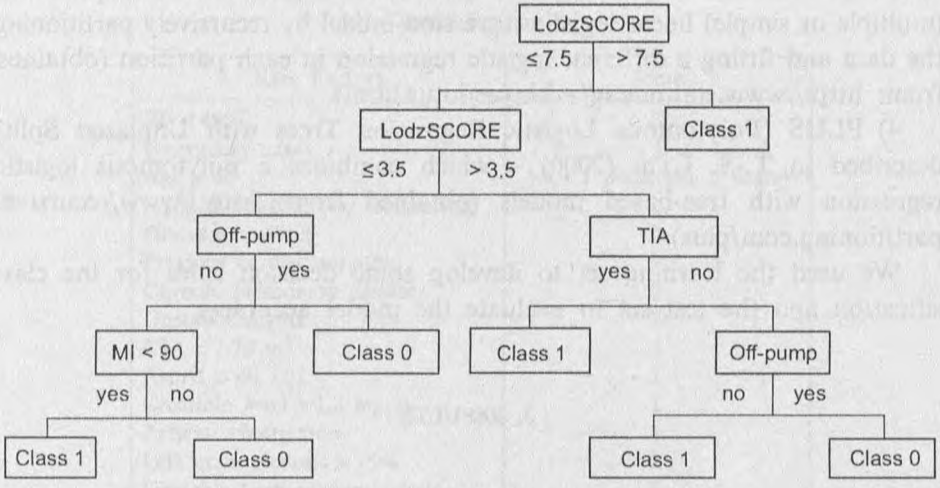

Fig. 1. QUEST classification tree

Source: own elaboration.

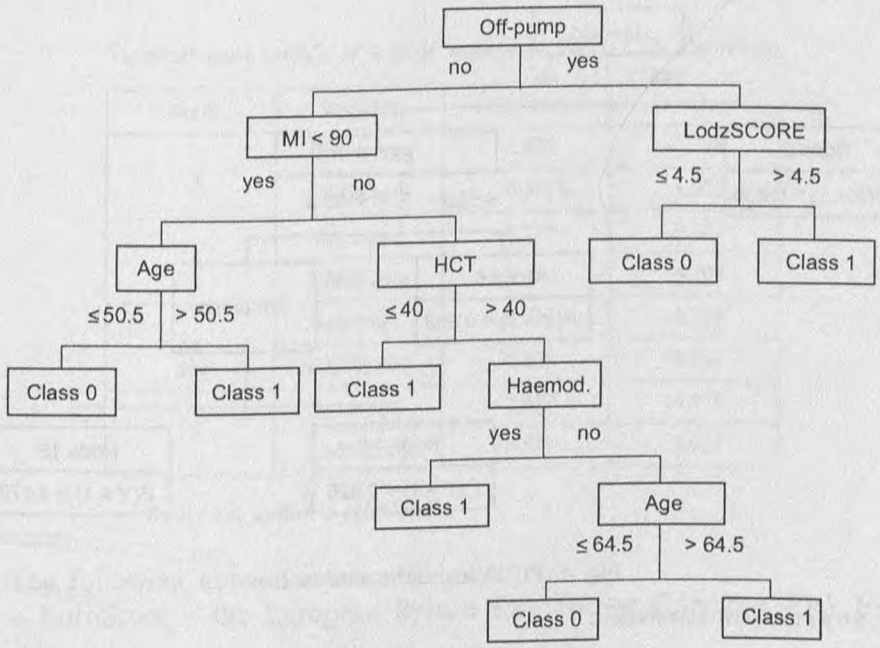

Fig. 2. CRUISE classification tree

Source: own elaboration.

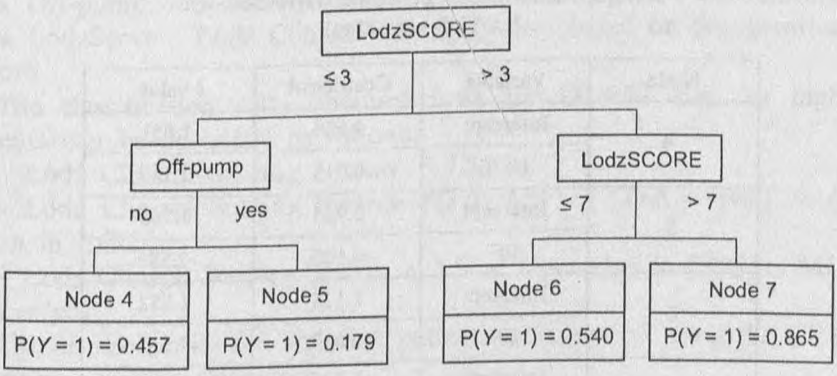

Fig. 3. LOTUS logistic regression tree (best simple linear logistic models in terminal nodes)

Source: own elaboration.

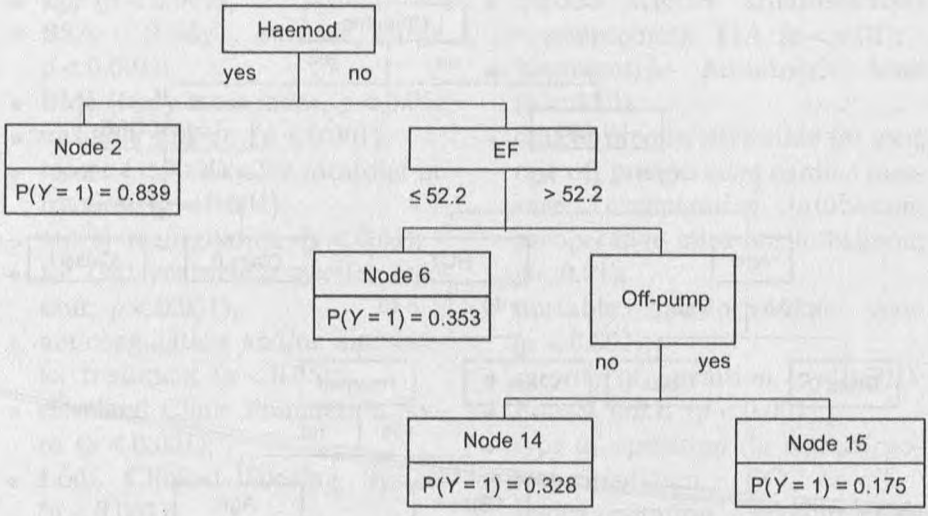

Fig. 4. PLUS logistic regression tree

Source: own elaboration.

Table 2

Terminal node models of logistic regression tree (LOTUS algorithm)

Node	Variable	Coefficient	*t*-value
4	Intercept	1.628	1.051
	HCT	–0.043	–1.167
5	Intercept	3.735	1.748
	EF	–0.105	–2.386
6	Intercept	1.178	1.852
	EF	–0.020	–1.638
7	Intercept	–1.433	–0.487
	Age	0.048	1.112

Source: author's calculations.

Table 3

Terminal node models of logistic regression tree (PLUS algorithm)

Node	Variable	Coefficient	*t*-value
2	Intercept	–2.653	–1.318
	ŁódźScore	0.5615	1.928
6	Intercept	–0.713	–3.107
	ŁódźScore	0.190	4.302
14	Intercept	–0.513	–2.399
	EuroScore	0.208	3.314
15	Intercept	–3.468	–4.470
	ŁódźScore	0.589	3.927

Source: author's calculations.

The following abbreviations were used:

- EuroScore – the European System for Cardiac Operative Risk Evaluation;
- EF – left ventricular ejection fraction;
- Haemod. – unstable haemodynamic state;
- HCT – preoperative hematocrit level;
- TIA – carotid arteries arteriosclerosis – symptomatic TIA;
- MI < 90 – recent (< 90 days) MI;
- Off-pump – operation without ECC (extracorporeal circulation);
- LodzScore – Łódź Clinical Scoring System based on preoperative risk factors.

The classification rules obtained from the QUEST tree for high-risk patients can be described as follows:

– Łódź Clinical Scoring System > 7.5;

– Łódź Clinical Scoring System $\in$(3.5; 7.5] $\wedge$[(TIA = 'yes') $\vee$ (Operation in ECC)];

– Łódź Clinical Scoring System $\leqslant$ 3.5 $\wedge$ Operation in ECC $\wedge$ MI < 90 days = 'yes'.

The decision rules for high-risk patients, constructed using the CRUISE algorithm, are the following:

– operation without ECC $\wedge$ Łódź Clinical Scoring System > 4.5;

– operation in ECC $\wedge$ MI < 90 days = 'yes' > age 50.5 years;

– operation in ECC (MI < 90 days = 'no' $\wedge$ [(preoperative hematocrit level $\leqslant$ 40%) $\vee$ unstable haemodynamic state $\vee$ (age > 64.5 years)].

Trees obtained from LOTUS and PLUS algorithms have 4 terminal nodes. Best simple linear logistic regression models are fitted in every terminal node. Some more details and the interpretation of the parameters are expounded in M. Misztal (2005). Logistic regression trees are shorter than classification trees.

The results of the application of the selected tree-based models for the learning and test sets are summarized in Tab. 4–5.

Table 4

Classification matrix based on the learning sample

Method	Actual risk group	Predicted group		% of correct classifications	10-fold *CV*-error rate
		class 0	class 1		
QUEST	class 0	418	211	66.45	37.17
	class 1	121	197	61.95	
CRUISE	class 0	486	139	77.76	26.14
	class 1	64	254	79.87	
LOTUS	class 0	405	224	64.39	38.50
	class 1	117	201	63.21	
PLUS	class 0	409	220	65.02	37.90
	class 1	103	215	67.61	

Source: author's calculations.

Table 5

Classification matrix based on the test sample

Method	Actual risk group	Predicted group		% of correct classifications
		class 0	class 1	
QUEST	class 0	700	222	75.92
	class 1	273	426	60.94
CRUISE	class 0	542	380	58.79
	class 1	200	499	71.39
LOTUS	class 0	695	227	75.38
	class 1	314	386	55.14
PLUS	class 0	616	306	66.81
	class 1	200	499	71.39

Source: author's calculations.

The results obtained for the test sample are usually worse than for the training set. On the basis of the results showed in Tab. 4 and 5 for further analyses we can recommend decision rules constructed using QUEST and PLUS trees.

4. CONCLUSIONS

According to L. Breiman *et al.* (1984) there are at least two main objectives of a classification task: 1) to get as accurate prediction as possible on unseen data and 2) to gain understanding and insight into the predictive structure of the data.

The results obtained from classification and logistic regression trees are not very good in terms of accuracy. One of the reasons can be that we have focused only on preoperative risk factors and have not taken into account events that can affect outcome after CABG during the intraoperative and immediate postoperative period.

However, the results are better than those obtained from classical multivariate statistical analysis (multiple logistic regression model: 62.5% of correct classifications for class 0 and 59.75% for class 1; discriminant analysis: 68.7 and 55.03% for class 0 and class 1 respectively; considering the learning sample. The results for the test set are even worse).

On the other hand, there are some other advantages of tree-based models over many traditional statistical methods:

1) no requirement of knowledge of the variable distribution;

2) dealing with: large data sets, high dimensionality, mixed data types, missing values, and outliers;

3) direct and intuitive way of interpretation (a hierarchy of questions is asked and the final decision depends on the answers to all the previous questions; the predicted classification of each patient as a class 0 or class 1 member can be made from a few simple "if-then" logical conditions);

4) 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 and

5) ability to make sense of the data.

Recursive partitioning can be recommended as a supplement to classical statistical methods such as discriminant analysis or logistic regression. It identifies subgroups with different risk and also may uncover relationships between variables in different parts of the measurement space that may be overlooked in the traditional analysis.

REFERENCES

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

Chan K.-Y., Loh W.-Y. (2004), *LOTUS: An Algorithm for Building Accurate and Comprehensible Logistic Regression Trees*, "Journal of Computational and Graphical Statistics", **13**, Issue 4, 826–852.

Domański Cz., Iwaszkiewicz-Zasłonka A., Jaszewski R., Zasłonka J. (red.) (2003), *Zastosowanie metod statystycznych w badaniach pacjentów z chorobą niedokrwienną serca leczonych operacyjnie*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

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

Higgins T. L., Estafanous F. G., Loop F. D., Beck G. J., Blum J. M., Paranandi L. (1992), *Stratification of Morbidity and Mortality Outcome by Preoperative Risk Factors in Coronary Artery Bypass Patients. A Clinical Severity Score*, JAMA (May 6), **267**, 17, 2344–2348.

Kim H., Loh W.-Y. (2001), *Classification Trees with Unbiased Multiway Splits*, "Journal of the American Statistical Association", **96**, 598–604.

Lim T.-S. (2000), *Polytomous Logistic Regression Trees*, Department of Statistics, University of Wisconsin, Madison, PhD Thesis.

Loh W.-Y., Shih Y.-S. (1997), *Split Selection Methods for Classification Trees*, "Statistica Sinica", **7**, 815–840.

Misztal M. (2005), *Wykorzystanie metody rekurencyjnego podziału do identyfikacji grup ryzyka operacyjnego pacjentów z chorobą wieńcową*, [in:] *Klasyfikacja i analiza danych – teoria i zastosowania*, K. Jajuga, M. Walesiak (red.), "Taksonomia", **12**, (Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1076, Wydawnictwo AE we Wrocławiu), 330–338.

Nashef S. A., Roques F., Michel P., Gauducheau E., Lemeshow S., Lemshow S., Salamon R. (1999), *European System for Cardiac Operative Risk Evaluation (EuroSCORE)*, "European Journal of Cardiothoracic Surgery", **16**, 9–13.

Małgorzata Misztal

PROPOZYCJA WYKORZYSTANIA WYBRANYCH MODELI DRZEW KLASYFIKACYJNYCH I REGRESYJNYCH DO IDENTYFIKACJI GRUP RYZYKA OPERACYJNEGO PACJENTÓW Z CHOROBĄ WIEŃCOWĄ LECZONYCH OPERACYJNIE

Drzewa klasyfikacyjne i regresyjne należą do bardzo popularnych metod klasyfikacji, przede wszystkim ze względu na prostotę interpretacji i przejrzystą formę wizualizacji wyników. Stąd też są one szeroko wykorzystywane do rozwiązywania problemów decyzyjnych w różnych dziedzinach nauki.

Celem prowadzonych badań była identyfikacja przedoperacyjnych czynników ryzyka, związanych z wystąpieniem powikłań śród- i pooperacyjnych wśród pacjentów z chorobą wieńcową, leczonych w sposób operacyjny.

Dodatkowo podjęto próbę zdefiniowania reguł decyzyjnych, które mogłyby umożliwić przydzielenie pacjenta do jednej z wyróżnionych grup ryzyka operacyjnego na podstawie opisujących go cech przedoperacyjnych.

Reguły klasyfikacyjne budowano wykorzystując metodę rekurencyjnego podziału. W analizie uwzględniono algorytmy QUEST i CRUJSE, tworzące drzewa klasyfikacyjne oraz algorytmy LOTUS i PLUS, łączące rekurencyjny podział przestrzeni cech z analizą regresji logistycznej.