2. How to build decision trees?

The procedure of the formation of a decision tree by statistical data is also called construction of a tree. In this paragraph we will get acquainted to some ways of construction of trees and also ways of definition of decision tree quality.

For each specific target of the statistical analysis, there is a large number (frequently even indefinitely) of different variants of decision trees. There is a question: which tree is the best and how to find it? To answer on the first question, we will consider various ways of definition of the parameters describing the quality of a tree. Theoretically, we can consider the expected error of forecasting as the basic parameter. However, this value can be defined only if we know the probabilistic distributive law of the examined variables. In practice however, this law, as a rule, is unknown. Therefore we can estimate quality only approximately, using the set of observations given to us.

2.1 Parameters of the quality of a tree.

Let us assume that there is a decision tree and a sample of objects of size *N*. It is possible to choose two basic kinds of the parameters describing the quality of a tree. The first kind are parameters of accuracy and the second are parameters of complexity of a tree.

Parameters of accuracy of a tree are defined with the help of sample and characterize how good the objects of different classes are divided (in case of a recognition problem), or how high the prediction error is (in case of a regression analysis problem).

The relative number (frequency) of mistakes represents a part of the objects incorrectly referred by a tree to another's class:

$$\hat{P}_{err} = \frac{N_{err}}{N} ,$$

where

$$N_{err} = \sum_{S=1}^{M} \sum_{\substack{i=1 \\ i \neq \hat{Y}(S)}}^{K} N_i^S ,$$

where K is a number of classes.

The relative variance for a regression tree can be calculated by the next formula:

$$d_{om} = \frac{d_{oc}}{d_0},$$

where $d_{oc} = \frac{1}{N} \sum_{S=1}^{M} \sum_{i \in Data^S} (\hat{Y}(S) - y^i)^2$ is a residual variance,

$$d_0 = \frac{1}{N}\sum_{i=1}^{N}(y^i - \overline{y})^2$$

is an initial variance and

$$\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y^i .$$

Parameters of complexity characterize the form of the tree and are not depending on the sample.

For instance, parameters of complexity of a tree are the number of leaves of the tree, the number of its internal nodes and the maximal length of a path from the root to a leaf.

Also it is possible to use the length of an external way which is defined as number of the branches supplementing the tree up to a full tree.

Parameters of complexity and accuracy are interconnected: a more complex tree, as a rule, is more accurate (accuracy will be maximal for the tree where one leaf corresponds to each object).

A less complex tree, with other things being equal, is more preferably. It explains the aspiration to receive a simpler model of the researched phenomenon and to facilitate the subsequent interpretation (explanation of the model). Besides, from theoretical research follows that in case of a small (in comparison with the number of characteristics) sample size a too complex tree becomes unstable, i.e. gives a higher error for new observations.

On the other hand, it is clear that a very simple tree will also not allow to achieve good results of forecasting. Thus, at a choice of the best decision tree, there should be reached a certain «compromise» between parameters of accuracy and complexity.

To get such a compromise variant, it is possible to use, for example, the following criterion of the tree quality: $Q = p + \alpha M$, where p is a parameter of accuracy, α is a given parameter. The best tree corresponds to the minimal value of the given criterion.

The approach, where maximal admissible complexity of a tree is specified, is used also and in the same time the most precise variant is searched.