5. Software description for decision tree construction.

Now, dozens of computer programs for construction of decision trees are known. The difference between these programs lies into in a type of solved problems, in used methods, in a level of the service given to users. Many of these programs are available on the Internet for free or share ware access. The most popular programs are the systems CART (used for pattern recognition problem and regression analysis problem) and system C4.5 (for pattern recognition problem).

One can acquaint with the widely known program system CART destined for construction of decision trees in pattern recognition and regression analysis problems on the Internet site http://www.salford-systems.com/.

The institute of mathematics of the Siberian Branch of the Russian Academy of Science developed the system LASTAN, destined for the solution of recognition problem and regression analysis problems. At the present moment a new version of this system which allow to solve a cluster analysis problem and the time series analysis problem being developed

The recursive method (see paragraph 2.7) of construction of a decision tree in system LASTAN is realized (*the parameter* α for simplicity, is equal to unit). In the given version the following restrictions are used:
The maximal number of objects - 1000,
The maximal number of variables - 50,
The maximal number of classes - 10.

To carry out the analysis of the data table after start program, it is necessary to take the following steps:
1. Open your data table file (File|Open). If such file does not exist, choose File|New (it is possible to use the correspondent icons for the convenience).
The given file must be textual and written in the following format, for example:
(the text between symbols # # means the comments and is ignored by the program).

```
  16  # number of objects #
  16  #number of variables  #
  N  #code of unknown value #
 #vector of types of variables: 1 - quantitative; 0 (M) - qualitative, where M - number of values  #
 1 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1
 #names of variables:  #
Y X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X15
#data table; the numbers must be separated by a space or comma; decimal
values are separated by dot  #

 0 1 2 3 2 4 3 4 1 3 1 3 3 2 4 2
 2 2 2 4 4 4 3 4 3 2 3 3 0 3 4 4
 0 4 4 4 3 4 3 2 3 3 2 3 0 2 5 4
 4 1 1 1 1 1 3 2 3 3 1 1 7 1 3 0
14 3 4 3 2 3 3 4 1 4 1 1 3 1 4 3
15 3 3 4 3 3 3 4 1 1 1 4 3 0 5 3
```

You can edit this file by using a build editor.

There is also a possibility to use the constructed decision tree for forecasting new

observations. For this purpose it is necessary, after the above mentioned data table to write key word ***Predict***, to specify the number of objects for forecasting, then the data table for forecasting (in which, instead of values of a predicted variable symbols of unknown value SHOULD be specified) should follow. For example:

predict 20

N 4 1 4 3 4 3 3 3 1 1 1 4 0 4 3

N 4 3 4 3 3 3 1 1 1 4 3 3 0 5 3

N 1 1 1 1 1 3 2 3 3 1 1 7 1 3 0

N 3 4 3 2 3 3 4 1 4 1 1 3 1 4 3

N 1 4 2 4 2 3 4 3 1 4 1 3 2 2 4

_____

2. Assign parameters of algorithm (Run|Parameters)

XXX Recursive complexity:

Defines an amount of variables, falling into the combination considered by the algorithm in each step (search depth). When this parameter is increasing, both quality of the decision, time and required memory are increasing also.

The integer part of the parameter assigns the guaranteed size (r) combinations of variables.

The fractional part of the parameter sets a share of the most informative variables (for which the expected error is minimal), selected for searching combinations of the size r+1, r+2 etc. (from the variables chosen at the previous stage).

 Homogeneity :

The node is not subjected to the further branching if objects, which correspond it, are homogeneous. The homogeneous objects are the objects for which the relative variance for a predicting variable (or a percentage of recognition error, in case of forecasting of a qualitative variable) is less than the given parameter.

 MinObjInNode :

(the minimal number of objects in the node, concerning their general number): Defines a minimal possible number of objects in the node, which needs further branching (i.e. no branching is produced from this node if the node has less objects than specified).

 VariantsNumber:

Desired number of variants of a decision tree. The variants are build by the way of consequent excluding the most informative variable (corresponding to the root of tree) in the preceding variant. If new observations should be forecasted, the obtained decision trees are used for the voting procedure.

 FeatureToPredict:

The number of predicted variable in data table;

 FoldsCrossValidate:

Parameter *L* in the method *of L-fold* cross validation. In this method, the initial sample is divided randomly on *L* parts of approximately equal size. Then each part serially acts as a test sample, whereas other parts are united into learning sample. As a factor of quality of a method, the averaged test samples error is calculated. In case of a regression tree, the error (standard deviation) is normalized to the standard deviation of the initial sample.

1. Start the program of decision tree design (Run|Grow Tree)

4. Results are automatically saved in file 'output.txt'. For each variant of a tree, the cross-validation error is resulted.