Methods for statistical data analysis with decision trees

Problems of the multivariate statistical analysis

In realizing the statistical analysis, first of all it is necessary to define which objects and for what purpose we want to analyze i.e. to formulate the purpose of statistical research. If the information about objects of the analysis is not collected, it is necessary to define what objects and how to choose these, what characteristics of objects are important to us and how to receive the information about these characteristics.

Statistical set of objects are those objects, the phenomena, events etc. which enter into a circle of interests of the researcher during the solution of some specific target of the analysis. It may be, for example, a set of all enterprises of any branch, or all patients, suffering some illness etc.

A Sample of objects is that part of a statistical set, about which the information is known to the researcher. More often, the sample $o^1,...,o^N$ is formed as a result of random selection of some representatives of the set. Number N of these representatives is called *volume of sample*.

The characteristic of object is on the basis of which it is possible to describe and distinguish objects. For example, it may be the number of employees of the enterprise or age of the patient. Other equivalent names such as parameter, attribute and factor are frequently used. In mathematical statistics the term *variable* is used. To describe the characteristic, it is necessary to put its name and the set of values. There are the following basic types of characteristics: *quantitative, qualitative and ordered.*

With values *of quantitative* characteristics, it is possible to carry out various arithmetic operations: addition, multiplication, division etc. With *qualitative* characteristics it is impossible to carry out such operations, but it is possible to check coincidence of values of these characteristics. For example, there is no sense in trying to divide one profession into another. With *ordered* characteristics it is allowed to compare their values according to the given order. For example, it is possible to compare a grade of various products or to tell, which officer in the military is higher ranked.

For convenience of the analysis, ordered characteristics can be ranked i.e. to give each value a rank (ordered number) according to increasing or decreasing.

The set of characteristics contains various characteristics $X_1,...,X_n$, by which objects are described. A set $X=\{X_1,...,X_n\}$ may contain characteristics of one type and may include characteristics of different types (both quantitative and qualitative).

A set of possible values, which may accept X, is called *space of characteristics*. The set of characteristics may also include dependent characteristics $Y_1,..Y_m$, i.e. such characteristics, so that each depends from other characteristics X. We shall consider a case $m=1$ i.e. where there is one dependent characteristic Y.

We can understand *observation* as two things: process of measurement of characteristics and the result of this process. If the result of the observation of characteristics of some object can change by case, we can speak about random observation. For any object o a set of observation of its

characteristics is a set $x=x(o)=x_1(o),...,x_n(o)$ where $x_j(o)$ designates the value of characteristic X_j for object o.

The set of observations is a set of measurements of characteristics for objects from the sample. This set is usually represented as the data table. *Data* with *N* rows and *n* columns: *Data*=$\{x^i_j\}$, where value x^i_j, is taking place on crossing *i-th* line and *j-th* column and corresponds to observation *j-th* characteristics of *i-th* object: $x^i_j = X_j(o^i)$.

For some reasons, some observations of any characteristics may remain unknown. In this case we can say that the table of the data contains *missed values*. These *missed value* are coded by a special number or by a special symbol.

A time series is a set of observations of the characteristic of one object at the various moments of time $t^1,...,t^N$. A multivariate time series represents a set of observations of several characteristics of one object.

The main goal of the statistical analysis consists in using the given set of observations, to catch the latent statistical regularities in the data, to establish influences between given random characteristics and other characteristics and to construct a model of dependence. The given set of observations is also called training sample (by analogy to process of person training).

The Model of dependence is the mathematical record of how one or several characteristics depend on other characteristics. The model can be described as the formula, the equation or system of the equations, a set of logic statements and graphically as a decision tree. The model can be used for forecasting values of the characteristic on value of other characteristics. Thus, conformity between sets of values of these characteristics (this conformity refers to as *decision function*) is established.

Let us consider the following basic kinds of statements of problems of the statistical analysis.

Regression Analysis (RA). In this kind of the statistical analysis, it is required to find a model of dependence of quantitative characteristic Y from other characteristics $X_1,...,X_n$.

Pattern Recognition Problem (PRP) (synonyms: the discriminant analysis, supervised classification). In this case, the dependent characteristic is qualitative, its values are called classes (patterns). It is necessary to find the model which would allow to predict a class depending on values of other characteristics. We will consider a variant of a recognition problem in which the cost of an error is not taken into account, i.e. it is not important, to which classes instead of true observation is referred; only the fact of an error is important. Sometimes a recognition problem with two classes is considered as a regression analysis problem in which instead of the class, it predicts probability (i.e. the quantitative characteristic).

If the predicted characteristic is ordered then such a problem can also be presented as a regression analysis problem, in this case the ranks turn out by a rounding off of values of the predicted quantitative characteristic.

The time series analysis. It is required to predict the value of any (quantitative or qualitative) characteristics at some future moment of time on values of all characteristics in the past.

The Cluster analysis (synonyms: an automatic grouping, unsupervised classification). In this kind of analysis there is no dependent characteristic; it is necessary to generate groups of objects, so that inside each group of objects, the elements are the closest to each other while on the same time

the elements of various groups have to be farther away from each other.

1. What is a decision tree?

Let us consider the following example of a recognition problem. During a doctor‘s examination of some patients the following characteristics are determined:

X_1 - temperature, X_2 - coughing, X_3 - a reddening throat,

$Y=\{W_1,W_2,W_3,W_4,W_5\}$= {a cold, quinsy, the influenza, a pneumonia, is healthy} - a set from the possible diagnoses, demanding more profound inspection.

It is required to find a model, where Y depends on X. The example (figure 1) illustrates such a model, which can be seen as a decision tree.

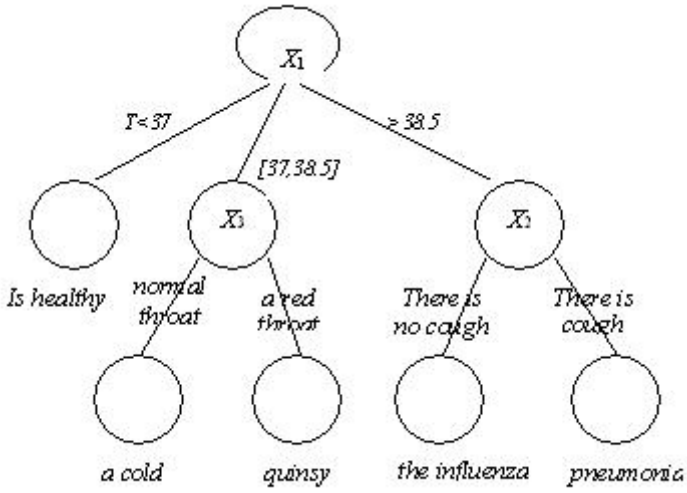


Fig. 1

The ordinary tree consists of one root, branches, nodes (places where branches are divided) and leaves. In the same way the decision tree consists of nodes which stand for circles, the branches stand for segments connecting the nodes. A decision tree is usually drawn from left to right or beginning from the root downwards, so it is easier to draw it. The first node is a root. The end of the chain " root - branch - node-... - node " is called "leaf". From each internal node (i.e. not a leaf) may grow out two or more branches. Each node corresponds with a certain characteristic and the branches correspond with a range of values. These ranges of values must give a partition of the set of values of the given characteristic.

When precisely two branches grow out from an internal node (the tree of such type is called a dichotomic tree), each of these branches can give a true of false statement concerning the given

characteristic as is shown on figure 2.

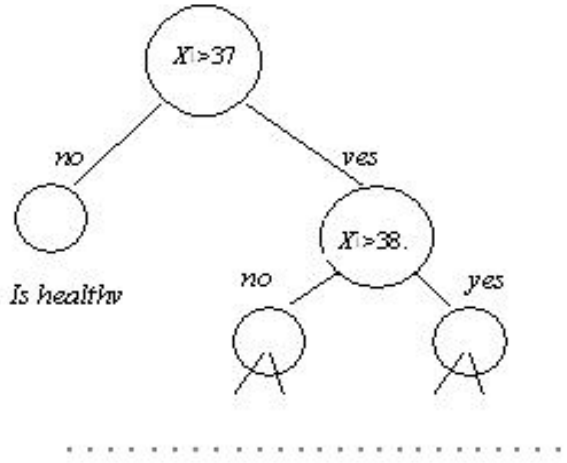


Fig. 2

The value Y is ascribed for each terminal node of a tree (named "leaf"). In case of pattern recognition problem the given value is a certain class, and in a regression analysis case the given value represents a real number. Decision trees for a cluster analysis problem will be considered separately in §4.

For any observation of x, using a decision tree, we can find the predicted value Y. For this purpose we start with a root of a tree, we consider the characteristic, corresponding to a root and we define, to which branch the observed value of the given characteristic corresponds. Then we consider the node in which the given branch comes. We repeat the same operations for this node etc., until we reach a leaf. The value YS ascribed to S-th leaf will be the forecast for x. Thus, the decision tree gives the model T of dependence Y from X: Y=T(X).

Decision trees, which are considered in a regression analysis problem, are called regression trees.

In the given manual we consider the simplest kind of decision trees, described above. There are, however, more complex kinds of trees, in which each internal node corresponds to more complex statements, not one but several characteristics are given. For example, these statements can be defined by a linear combination of quantitative characteristics (for example, expression 10x1+5x2-1>0) i.e. corresponding to various subregions of multivariate space which are split by a hyper plane).

From this point of view, the hyper planes of the considered decision trees are perpendicular to the numerical axes.

The decision tree should be consistent, which means that on the way from the root to a leaf, there should be no mutually excluding variants, for example «$X_1 > 37$» и «$X_1 < 30$».

It is possible to allocate the following features of the decision trees.

Decision trees allow to process both quantitative and qualitative characteristics simultaneously.

A set of logic statements about values of characteristics corresponds to decision trees. Each statement is obtained by passing the way from root to leaf. So, for example, for the tree represented on figure 1 the following list of statements corresponds to:

1. If $X_1 < 37$, *Y="is health"*.

2. If $X_1 \in [37,38.5]$ and *X3="there is no reddening of throat"*, then *Y="to catch cold";*

3. If $X_1 \in [37,38.5]$ and *X3="there is reddening of throat"*, then *Y="angina";*

4. If $X_1 > 38.5$ and *X2="there is no cough"*, then *Y="influenza";*

5. If $X_1 > 38.5$ and *X2="there is cough"*, then *Y="pneumonia";*

Thus, the decision tree represents a logic model of regularities of the researched phenomenon.

The lack of decision trees is the fact that in a case where all characteristics are quantitative, the decision trees may represent sufficiently rough approximation of the optimum solution. For example, the regression tree, which is drawn on figure 3, is piecewise a constant approximation of the regression function. On the other hand, it is possible to compensate this lack by increasing the number of leaves, i.e. by decreasing the length of appropriate "segments" or "steps".

Let us consider a decision tree with M leaves. This decision tree corresponds to the decomposition of the characteristic space into M non-overlapping subregions E1,...,EM , so that subregion ES corresponds to S-th leaf (fig. 4). How is the given subregion formed?

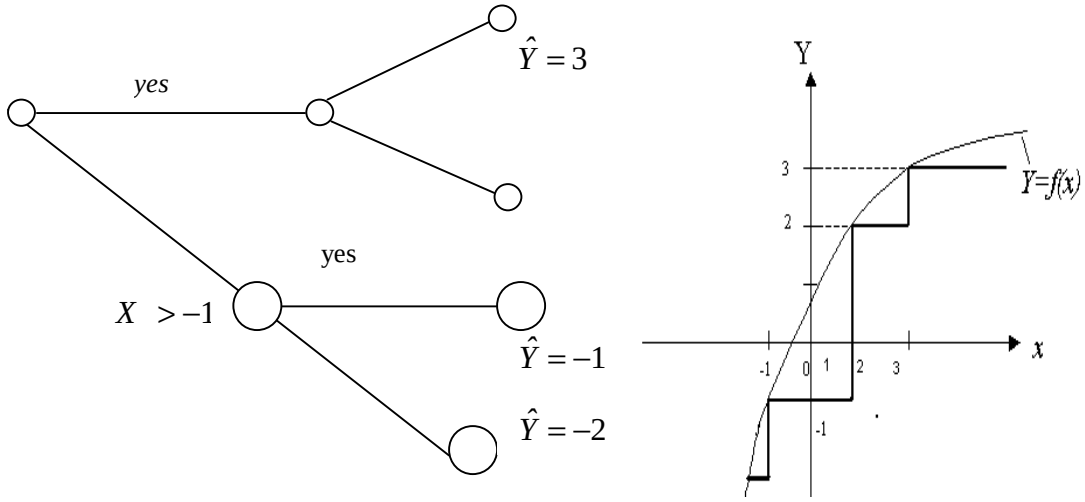


Fig. 3

E^S is defined as Cartesian product $E^S = E_1^S \times E_2^S \times ... \times E_n^S$, where E_j^S - is projection E^S on *j-th* characteristic. E_j^S is obtained at the next way. If the characteristic X_j never situated on the way

from the root to *S-th* leaf, then E_j^S coincides with a range of definitions of the characteristic X_j. Otherwise, E_j^S is equal to intersection of all subregions of the characteristic X_j, which were met on the way from the root to *S-th* leaf.

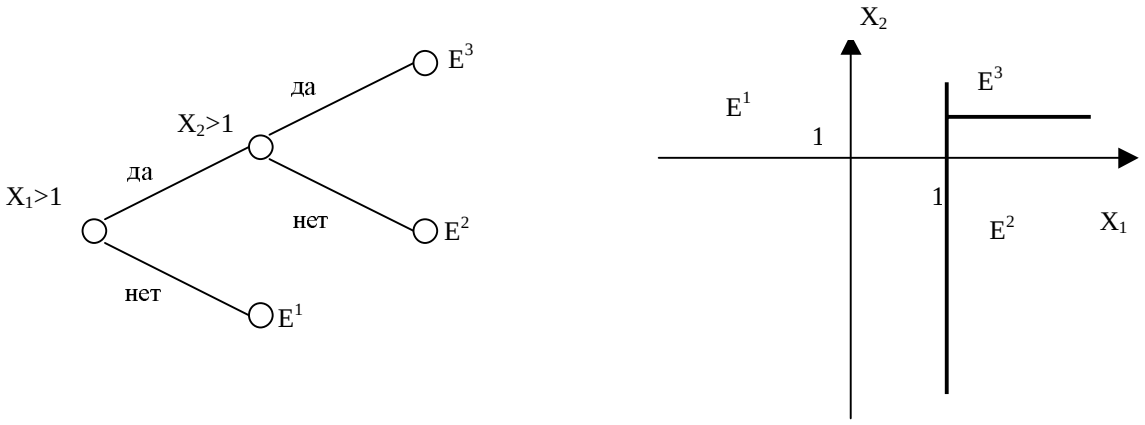


Fig. 4

Let there be some set of experimental observations $Data=(x^i,y^i)$, $i=1,...,N$. Each of these observations belongs to (with respect to X) some of the considered subdomains, i.e. $x^i \in E^S$. We will denote the set of the observations belonging to E^S as $Data^S$, and the number of the observations, we denote as N^S. Let N_i^S denote the number of observations from $Data^S$, belonging to *i-th* class (pattern recognition problem PRP).

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.

2.2 An estimation of quality on a control sample.

Control (test) sample is called sample, which is not used for building a tree, but used for an estimation of quality of the constructed tree. Two parameters are calculated: for the recognition problem the relative number of mistakes and for the regression analysis problem the variance on control sample. Since this sample does not participate in the tree construction, these parameters reflect the «true» unknown error more objectively. The bigger the control sample size, the higher the degree of approximation.

For a recognition problem, under the condition of independency of the observations, the frequency of mistakes belongs to binomial distribution. Therefore, knowing the number of errors on the control sample, it is possible to find a confidence interval to which, with the given probability, the unknown value of misclassification error belongs. In work [5] are given diagrams in which it is possible to define a confidence interval for the given control sample size and number of errors in the control.

2.3 Methods of construction of decision trees.

Existing methods (there are dozens of methods) can be divided into two basic groups. The first group is methods of building of a strict-optimum tree (by the given criteria of quality of a tree) and the second group is methods of construction of an approximately optimum tree.

The problem of searching the optimum variant of a tree can be related to a discrete programming problem or a choice from finite (but very large) numbers of variants. It follows from the fact that for finite training sample the number of variants of branching (see below) for each characteristic is finite.

Three basic kinds of methods in discrete programming are considered: exhausting search, a method of dynamic programming and a branch and bounds method. However, these methods, in the application to decision trees, as a rule, are very laborious, especially for the large number of observations and characteristics. Therefore, we will consider approximate methods: method of consecutive branching, a method of pruning and a recursive method.

Let us consider all basic operations with decision trees. Methods of tree constructing will represent the certain sequence of these operations.

2.3.1 Operation of branching (division).

This operation is the basic operation for tree constructions. We will consider a node of a tree and some characteristic X_j. Let the range of definition of this characteristic be divided on L_j subsets (ways of a choice of such subsets we will consider below). In case of the quantitative characteristic, these subsets represent a set subintervals of splitting, in case of the qualitative characteristic a subsets of values and in case of the ordered characteristic the subsets including the neighbouring values.

Let us associate with each of these subsets a branch of the tree leaving the given (parent) node and going into a new node which is called descendant. Thus, the node "has branched" ("has divided") on L_j new nodes (figure 5).

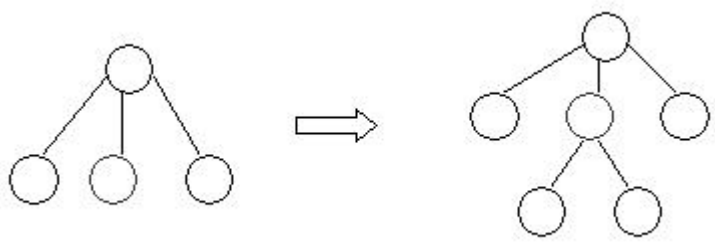


Fig. 5

Notice that for binary trees L_j is always equal to two. If L_j is always equal to three such trees are called *ternary*. If L_j is always equal to four we receive quadratic-trees.

How to obtain the splitting of a range of definition? We take a set of the observations corresponding to the given node and consider values of characteristic X_j for these observations.

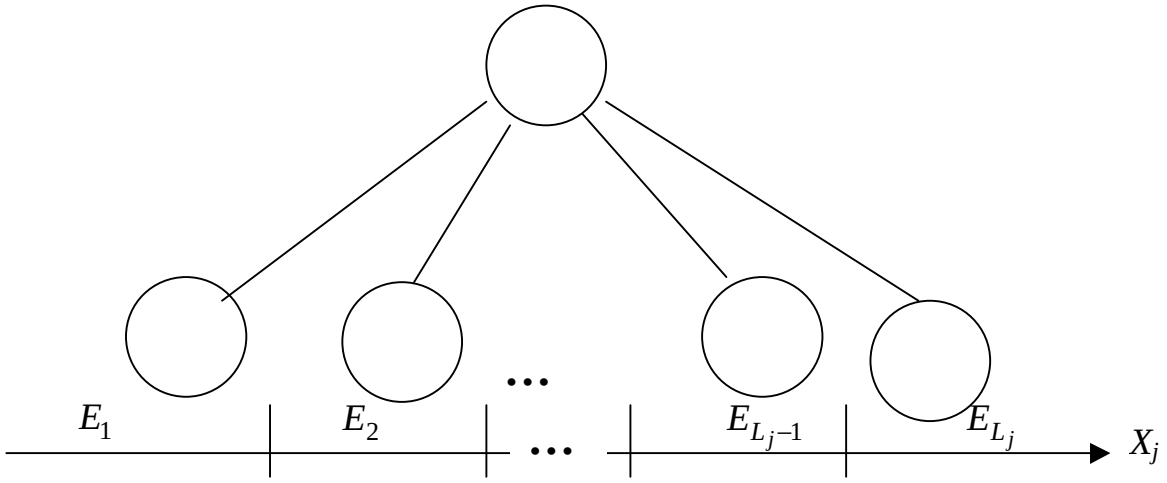


Fig. 6

Consider a quantitative characteristic. In this case, boundaries are in the middle of intervals between

the neighbor values and splitting carried out on these boundaries (figure 6).

For example on figure 7 (values of the characteristic for observations are designated through ⊗), in case of a binary tree, it is possible to consider the following variants of splitting: $X_j < 0.5$ or $X_j >= 0.5$, $X_j < 1.5$ or $X_j >= 1.5$. If the characteristic is qualitative, then the variants of splitting are values of the characteristic, for example if X_j means a country, the following splitting can be received: $X_j \in$ {Canada, Mexico, USA} or $X_j \in$ {Argentina, Brazil}.

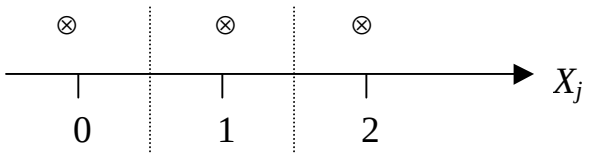


Fig. 7

In case of the large number of values, the number of possible variants of splitting becomes too big, therefore for acceleration of process of tree construction one considers not all variants, but only some of them (for example, such as X_j ="Canada» or $X_j \neq$ "Canada»).

In case of the ordered characteristic, variants of splitting consist of the ordered values, for example if X_j is a military rank, division can be such: $X_j \in$ [the soldier – the first sergeant] or $X_j \in$ [the lieutenant - the major].

For qualitative or ordered characteristics, it can happen (when sample size of observations is small) that the set of values of characteristics for the observations appropriate to the node, are only a part of all range of definition of this characteristic. In this case, it is necessary to attribute the rest values to a new branch. Now at forecasting the object of the control sample, which had such value, we can define to which branches it belongs. For example, it is possible to attribute the given values to the branch, corresponding with the greatest number of observations.

2.3.2 Operation of the definition of degree of promise for branching node (rule of a stopping).

Let us consider a dangling node of a tree, i.e. the node which is not branched, but it is not clear, whether this node will be a leaf or whether we need further branching. We will consider the subset of observations appropriate to the given node. We will divide the nodes into two cases. First, if these observations are homogeneous, i.e. basically belong to the same class (a pattern recognition problem, RP), or if the variance of *Y* for them is small enough (a regression analysis problem, RA). The variant when the values of characteristic are equal for all observations corresponds also to this case. Second, if the number of observations is not enough.

The node, unpromising for the further branching, is called leaf.

For definition of degree of promise, it is possible to set the following parameters: an allowable error for node (PR problem), an allowable variance (RA problem) and a threshold on the quantity of observations.

2.3.3 Operation "to attribute a solution to a leaf ".

Let us consider a leaf of a tree. A subset of observations *Data* corresponds to this leaf. During the solution of a pattern recognition problem, the class with maximal quantity of observations from *Data* is assigned to a leaf, in comparison with other classes. At the solution of a regression analysis problem, the solution attributed to a leaf, is equal to the average value of dependent characteristic *Y* for observation from *Data*.

2.3.4. Operation of "growth" of node.

This operation represents a sequence of operations of branching for each of the new nodes of a tree. As a result of this operation, the given node is replaced with some sub tree (i.e. a part of a full tree which also looks like a tree (figure 8)).

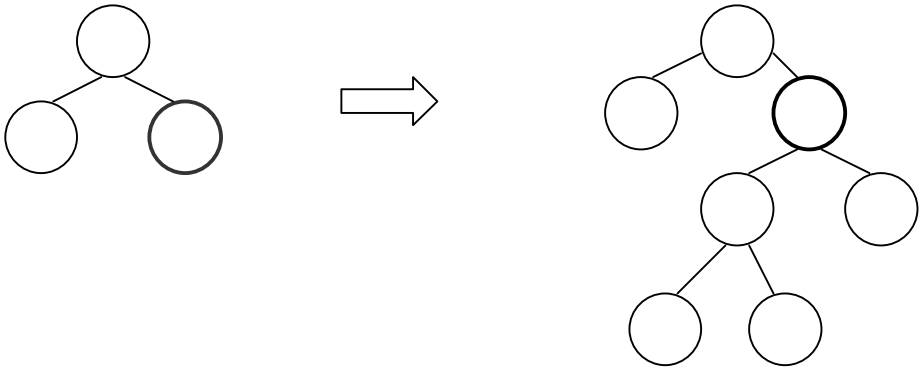


Fig. 8

The complexity of the sub tree is limited to a parameter. One of the ways of growth will be described below more detailed.

2.3.5. Operation of pruning.

This operation is the opposite of operations of growth, i.e. for the given node appropriate sub tree for which this node is a root completely cuts (figure 9). The node is then called a leaf.

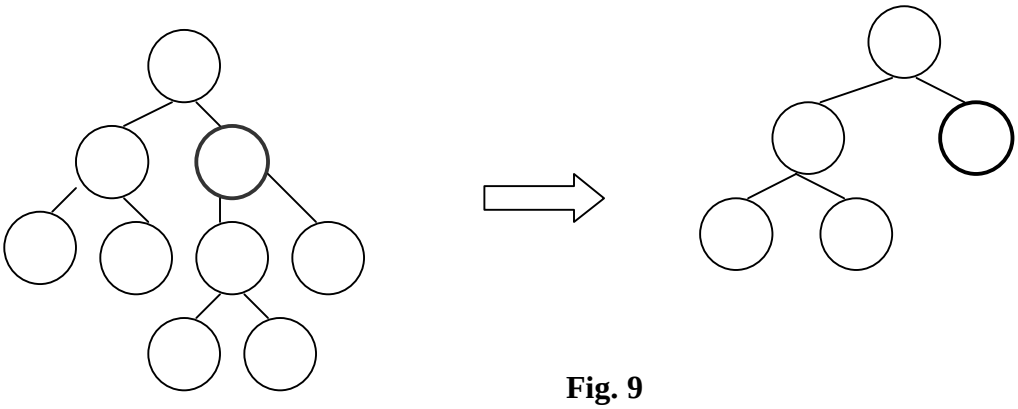


Fig. 9

2.3.6. Operation of "aggregation" of nodes or ("join").

Let the node be divided on *L* new nodes. We will take any pair of these nodes and we will unite them in one node and connect it with the parental node (figure 10). Thus subsets of the values to the appropriate aggregation nodes are united.

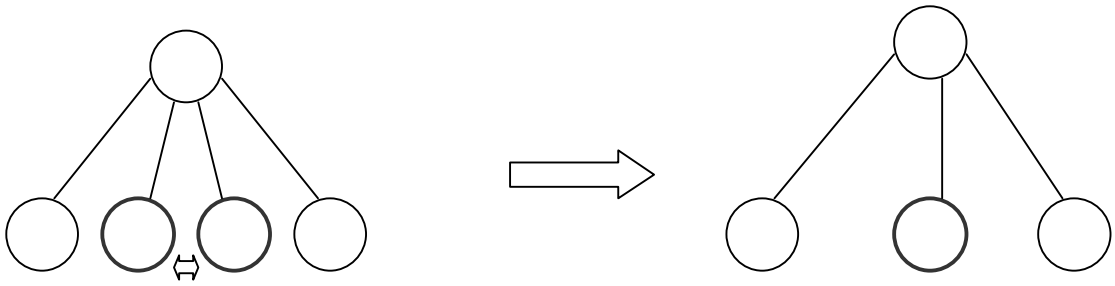


Fig.10

The aggregation of nodes, which correspond to the neighbor subintervals or subsets of values of quantitative and ordered characteristics is allowed.

2.4 Criteria of branching quality.

It is necessary to have a criterion, which will allow to compare all various variants of node branching and to choose the best of them.

The frequency of mistakes (PR problem) or a relative variance (RA problem) can be considered as such criterion.

Let the node be divided on *L* new nodes.

Let the number of observations appropriate to *l-th* new node be N_l,

$Data_l$ will be a set of these observations,

$\hat{Y}(l)$ will be the decision, attributed to *l-th* node and

N_l^{ω} will be a number of observations of ω–th class, which corresponds to *l-th* node (PR problem).

The general number of observations in the initial node will be *N*. Formulas for the calculation of the criteria are similar to which were used at defining the decision tree quality (because the variant of branching also represents a tree):

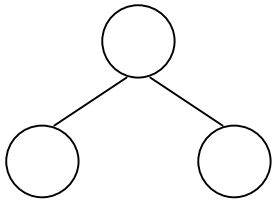
$$N_{err} = \sum_{l=1}^{L} \sum_{\substack{\omega=1 \\ \omega \neq \hat{Y}(S)}}^{K} N_l^{\omega} \text{ (PR problem);}$$

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

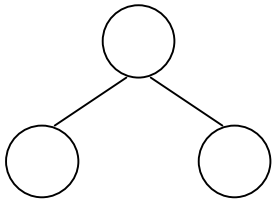
where

$$d_{oc} = \frac{1}{N}\sum_{l=1}^{L}\sum_{i \in Data_l} (\hat{Y}(l) - y^i)^2 \,,\; d_0 = \frac{1}{N}\sum_{i=1}^{N}(y^i - \bar{y})^2 \,,\; \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y^i \,.$$

For the PR problem more precisely methods of definition exist, as the objects of different classes are divided (in literature the term *"impurity" is used*, which can be interpreted as a degree of "pollution" of observations by objects from another classes). For example, we consider two variants of division (figure 11).



$N_1^1=0, \quad N_2^1=20,$
$N_1^2=20, \quad N_2^2=10,$
$N_{err}=10$



$N_1^1=5, \quad N_2^1=25,$
$N_1^2=15, \quad N_2^2=5,$
$N_{err}=10$

Fig. 11

The numbers of mistakes for these variants coincide, however it is clear, that the first variant is more preferable, since one of the new nodes does not need more branching because all objects in it are correctly referred to the same class.

To take into account similar cases, for the definition of division quality, it is possible to use entropy criterion or Gini's criterion.

The entropy criterion of splitting is defined by the formula:

$$H(L) = \sum_{l=1}^{L}\frac{N_l}{N}\sum_{\omega=1}^{K} -\frac{N_l^\omega}{N_l} log \frac{N_l^\omega}{N_l} = \frac{1}{N}\left(\sum_{l=1}^{L} N_l \, log \, N_l - \sum_{l=1}^{L}\sum_{\omega=1}^{K} N_l^\omega \, log \, N_l^\omega\right)$$

The lesser the entropy value is, the more information contains in the variant of division. Let the entropy for the initial node denoted as

$$H(0) = \sum_{\omega=1}^{K}\frac{N^\omega}{N} log \frac{N^\omega}{N} \,,$$

where N^ω means the number of observations of ω-th class.

It is possible to use for given branching the value *gain=H(L)-H(0)* as a measure of "usefulness" or "gain".

Note the next properties of entropy criterion:

1) If the number of classes is constant and frequencies of various classes converge to each other, then value *H* is increased.
2) If various classes are equiprobable and the number of classes is increasing, then *H* is

increasing logarithmically (i.e. proportionally to $\log_2 K$).

Some researches recommend that it is better to use L as the basis of *log*.

Gini's criterion for splitting is calculated by the following formula:

$$G(L)=\sum_{l=1}^{L}\frac{N_l}{N}\left(1-\sum_{\omega=1}^{K}\left(\frac{N_l^{\omega}}{N_l}\right)^2\right)=1-\frac{1}{N}\sum_{l=1}^{L}\sum_{\omega=1}^{K}\frac{\left(N_l^{\omega}\right)^2}{N_l}$$

The smallest value of this parameter corresponds to the best division of objects.

For the definition of quality, one can also use a parameter of branching *"gain"*, which is defined as a difference between the value of the given criterion for the initial node and the value of the variant of its division.

2.5 Method of sequential branching.

The given method represents a procedure of step-by-step branching at which on each step the best variant of division gets out.

Let the possible allowable complexity of a tree (for example, number of leaves M_{max}) be given. The method consists of the following steps (figure 12).

1) To divide the root into the given number of new nodes, using all variants of division by each characteristic X_i from $X_1,...,X_n$ by turns. The best variant of division by the given criterion of quality is saved.

2) To check up the degree of promise of branching for each of new child nodes. If a node becomes a leaf, we give to it the appropriate decision.

3) Each node (not leaf) is divided into new nodes similarly to item 1.

If branching of the given node does not improve the quality of a tree (or quality is improved, but it is less than the given threshold) branching is not made; the node becomes leaf and a decision is attributed to it.

After that, steps 2,3 are repeating until there will be no more perspective nodes for branching, or the maximal possible complexity of a tree will not be achieved.

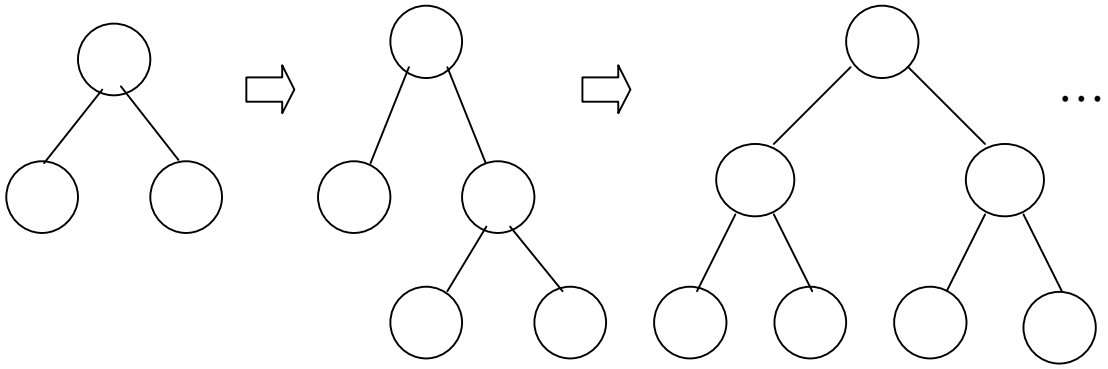


Fig. 12

The described method is the simplest and the fastest in execution. However, the problem of the given method is that when there are large allowable complexity and small sample size, the received tree, as a rule, is excessively 'optimistic' ('overtrained'). In other words, the prediction error determined on training sample will be much smaller than the 'true' error. This effect arises because observations are random and a tree, which depends on them, catches also random laws.

One more problem appears when characteristics have complex dependence between each

other. The received decision, as a rule, will not coincide with optimum. It happens because on each step one characteristic is considered only, without taking into account its interrelations with others.

2.6 Method of pruning.

As it was already mentioned, the elementary method of tree construction described above, can give a very "optimistic" decision. In this method, the estimation of the tree quality will be carried out on the same training sample on which the decision tree is formed.

For more objective estimation and elimination of random laws, it is necessary to use a sample which did not participate in tree construction.

In a pruning method, the training sample is divided in two parts. The first part is used for tree construction by a method of consecutive branching. Parameters of a stop are set in such way to provide the maximal possible accuracy of the received decision. The number of leaves of the tree can be very large.

The second part of the sample serves for pruning ("simplification") of the received tree. For this purpose, the following steps are carried out.

1) All internal nodes of the tree are considered one by one.

2) Operation of pruning for considered node is carried out.

3) Using the second part of the sample, we estimate the prediction error for the truncated variant of the tree.

The variant with minimal error is a result (figure 13).

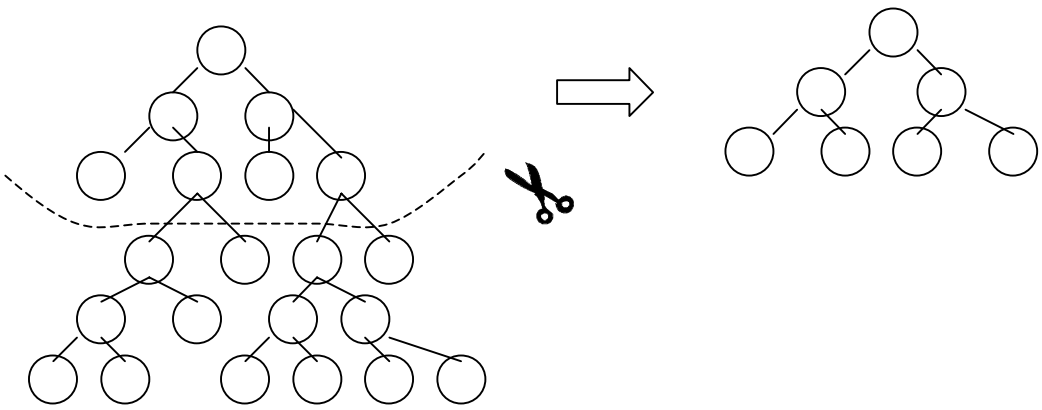


Fig. 13

The described method gives more objective estimation of quality, however, if the initial tree is far from optimum, then the truncated variant will not be ideal also.

2.7 A recursive method.

In case of complex dependence between characteristics for decision tree construction, one can apply more complex methods then a method of sequential branching.

The general scheme of a method is similar to the scheme which was used in a method of sequential branching with the difference, that instead of operation of division, more complex

operation of growth is used. Let us consider the basic steps of this method.

1) To carry out an operation of growth for the root node.
2) To check the degree of promise of branching for each of the new child nodes.

If a node becomes a leaf, then we give to it the appropriate decision.

3) To carry out an operation of growth for each child node;

After that we repeat steps 2 and 3, until there will be no more perspective nodes for branching, or maximal possible complexity of a tree will not be achieved.

Let us describe the operation of growth. In this operation, the next criterion of quality of the tree is used:

$Q=N_{err}/N+\alpha M/N$ (for PR problem) or

$Q=d/d0+\alpha M/N$ (for RA problem),

where α is some given parameter.

The operation consists of the following steps.

1) To fix the next characteristic X_j (j=1,...,n).
2) Initial branching (initial tree construction). One can understand initial branching as the dividing of the node on maximal possible (determined by sample) number of new nodes (figure 14). In case of the quantitative characteristic, this number is equal to the number of subintervals of splitting (i.e. the new node is matched to each subinterval), in case of the qualitative or ordered characteristic one value of the characteristic is matched to each node.

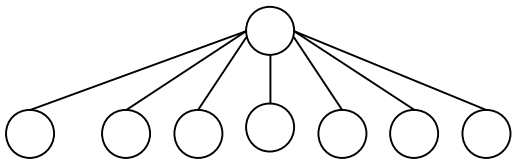


Fig. 14

3) To check up degree of promise of branching for all child nodes of the initial tree.
4) To carry out an operation of growth for perspective nodes of initial tree recursively (figure 15). The maximal recursion depth is limited to the given threshold *R*.

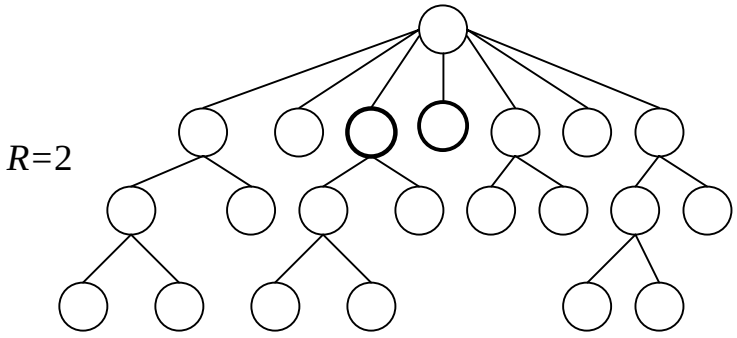


Fig. 15

5) To calculate the criterion of quality of the received tree.
6) To sort out all pairs of nodes of the initial tree allowable for aggregation.
7) To carry out for each pair the recursively operation of growth of the incorporated node and to calculate the quality of the received variant of a tree.
8) To remember the variant of a tree with the best criterion of quality (figure 16).

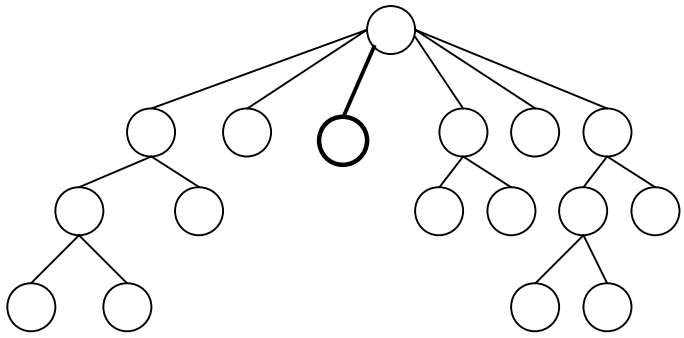


Figure 16

9) To repeat steps 6-8 for new pairs of incorporated nodes of initial tree. Every time to save the best variant. To repeat steps until all nodes will be united in one.
10) To pass to the next characteristic and to repeat steps 1-9 until all characteristics will be considered.

The most time-consuming in this algorithm is the fourth step. At each recursive reference to the operation of growth, steps 1-11 will be carried out. In other words, the optimum sub tree is forming for the appropriate subset of observations and this operation may call itself some times. By increasing the parameter *R*, it is possible to increase depth of a recursive nesting which allows finding more complex dependencies between characteristics (but in this case, time and a required memory is increased).

One more feature of this method is that the number of the branches leaving each node beforehand is not fixed and their optimum number is searched for.

The parameter α can be chosen equal to 1. By increasing the given parameter, the tree will become simpler (i.e. contain smaller number of leaves) and by reducing the tree will become more complex.

2.8 Quality estimation and methods comparison.

During quality estimation of the method of decision tree constructions, it is necessary to take into account required computer resources (time and memory). At the same time, the most important parameter of quality is the forecasting error.

As was shown above, the most objective way of definition of an error is the way based on using the control sample. This way can be applied at processing the large databases consisting of hundreds or thousands observations. However, during the analysis of the not so large sample, the dividing of sample on training and control can result in undesirable consequences, because the part of the information, which might be used for tree construction, is lost. Furthermore, the quality estimation will depend on a way of splitting of sample on training and control sample. For

decreasing the influence of this dependence, it is possible to use the methods based on repeated recurrence of splitting procedure and the averaging of received quality estimations.

One-leave-out method. In this method, each object of sample is taken off from it by turns. With the rest of the sample a tree is constructed, then this tree is used to forecast the given object. The predicted value is compared with the observed one, and then the object comes back into the initial sample. The percent of mistakes (in case of the PR problem) or an average square error (in case of the RA problem) shows the quality of a method.

This method is rather time consuming, since it is necessary to construct N decision trees (N is sample size).

L-fold cross-validation method. In this method, initial sample is divided on L random parts, which have approximately equal size. Then, by turns, we consider each part as a control sample and the rest parts are united in the train sample. Parameter of the quality of the investigated method is the error, averaged on control samples. The given method is less time consuming than one-leave-out method and with the reduction of parameter L comes closer to this method.

During the comparison of various methods of decision trees construction, it is very important by which data these trees are constructed. The ways of getting these data can be divided into two groups. The first group includes the real data intended for the decision of a concrete applied problem. For convenience of the comparison of various methods, these data are stored in the special databases in Internet. For example, UCI Machine Learning Database Repository
< http://www.ics.uci.edu / ~ mlearn/MLRepository.html .

The second group includes the data, which were artificially generated according to an algorithm. In this case, the data structure in space of characteristics is known beforehand. This information allows us to define precisely the quality of each method depend upon distribution family, training sample size and number of characteristics. For example, let us consider the data, which have a chessboard structure (figure 17). The first class (x) corresponds to the white cells and the second class (0) to the black cells. In addition to the characteristics X_1 and X_2, 'noise' characteristics X_3 and X_4 are available (each class has equal distribution on X_3 and X_4).

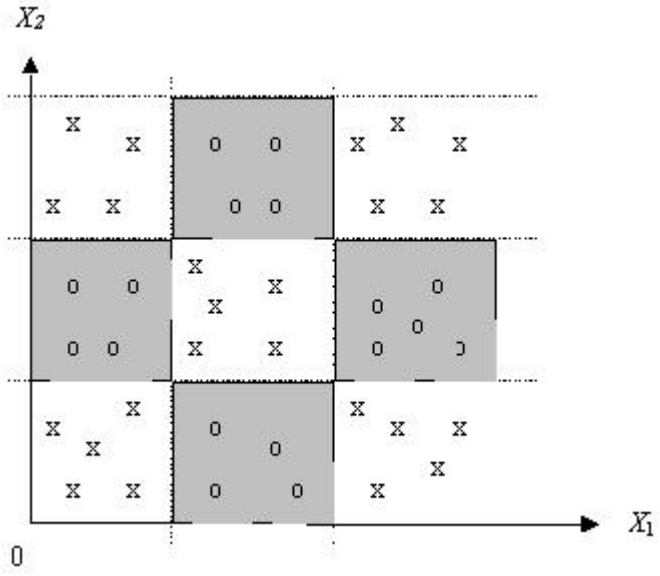


Fig. 17

The comparison of the methods described above has shown that only the recursive method is able to construct a decision tree, which can correctly determine the conceived date structure.

In general, numerous comparisons of existing methods show, that there is no universal method which equally well works on any data.

3. From decision trees to decision forest.

A decision forest is a set of several decision trees. These trees can be formed by various methods (or by one method, but with various parameters of work), by different sub-samples of observations over one and the same phenomenon, by use of different characteristics. Such many-sided consideration of a problem, as a rule, gives the improvement of quality of forecasting and a better understanding of laws of the researched phenomenon.

Let us consider a set of trees and an observation x. Each tree gives a forecast for x. How to find the general (collective) decision for the predicted variable Y?

The simplest way to obtain the collective forecast, which gives the given decision forest, is voting method (PR problem) or method of averaging (RA problem).

Using a voting method, a class attributed to observation x is a class which the majority of trees prefer. In the regression analysis problem, the predicted value is a mean of forecasts of all trees. We will consider, for example, a set of regression trees, which are shown at figure 18. Consider observation x = (3,4,8). The collective decision will be equal to $Y(x)$=(10.2+6.3+11.2)/3=9.233.

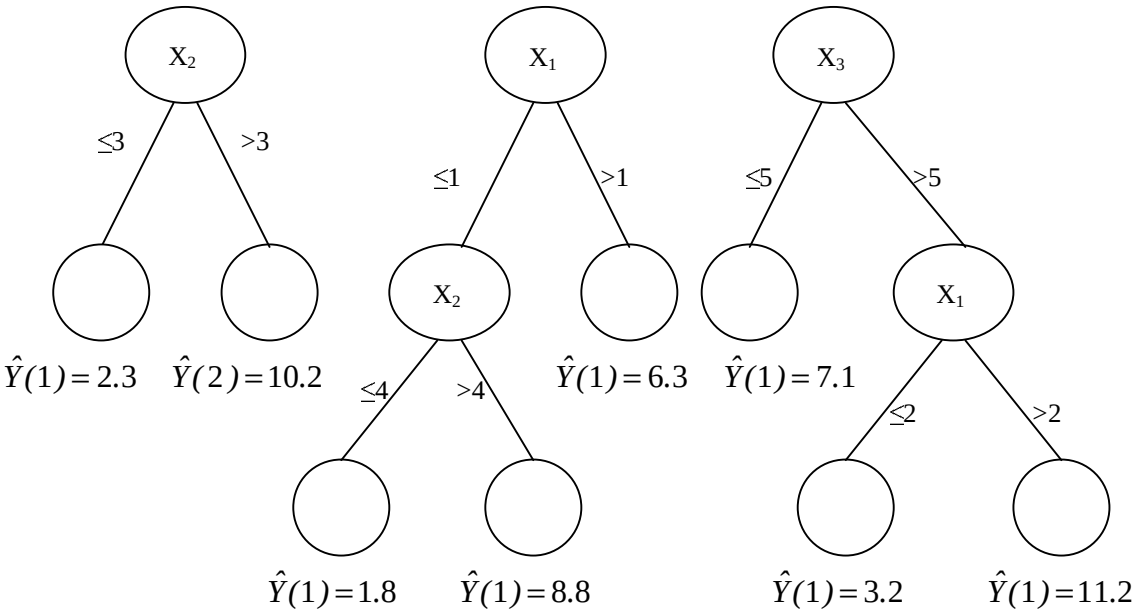


Fig. 18

Next to simple averaging or voting with equal contributions of each vote, it is possible to use a procedure in which the number of mistakes accomplished by each tree on training sample, is taken into account. The fewer mistakes, the greater weight the appropriate voice has.

For estimating the quality of the constructed decision forest, it is possible to use similar ways, which were used for an individual tree. Thus, both control sample and one-leave-out or cross-validation can be used.

Let us consider some ways of construction of a decision forest more detailed.

3.1 Consecutive exception of characteristics.

The given method consists of several stages. On each stage, we build a tree with the help of the full set of observations, but we use a different set of characteristics. On the first stage, all available characteristics are used. The characteristic, which corresponds to a root of the constructed tree, is the most informative, because firstly the further movement in a tree depends on it, and secondly, when branching, the full set of observations is used. On the next stage, when the second tree is forming, all characteristics are used, except for the above mentioned. It is done with the purpose to receive the variant of a tree which is the most distinguished from the previous, i.e. to find out the new set of laws. On the following stages, the characteristics appropriate to the roots of already constructed trees are consecutively excluded. Since the most informative characteristics are excluded, the quality of the trees, as a rule, only become worse from stage to stage.

The algorithm finishes work as soon as the given number of trees will be constructed, or the parameter of quality will reach the given minimal allowable value.

3.2 Using of various sub-samples.

The basic idea of the given method is to use various parts of initial training sample for tree constructions (for the construction of each tree the whole set of characteristics is used). Thus, the quality of the final (collective) decisions, as a rule, is improved. This property can be explained by fact, that the forecasts, given by random (unstable) laws on various sub-samples, at the end have no deciding vote. At the same time, really steady laws on various sub-samples are confirmed only.

Let us consider three methods of sub-samples forming: *bootstrap aggregation method, L-fold method and boosting method.*

In *the bootstrap aggregation method (‘bagging’)* for construction of the new tree, the sub-sample is formed by random independent selection of objects from initial sample. The probability of selection is identical to each object. The volume of sub-sample is set beforehand (for example, 70 % from initial). After the construction of a tree by way of analysis of given sub-sample, the selected observations return into initial sample and the process repeats the given number of times. Thus, each object can repeatedly get into analyzed sub-sample, but also some objects of initial sample can never be included into analyzed sub-samples.

To achieve guaranteed inclusion, it is possible to use *L-fold method*, which uses the same principle as the method of *L*-fold cross-validation described above (paragraph 2.8). The sample divided by case on *L* parts of approximately equal size, then each part is orderedly thrown out, and the rests are united in sub-sample, which is used to construct the next decision tree.

The boosting method is based on the following adaptation idea. At the beginning, the first decision tree is constructed on all objects of initial sample. As a rule, for a part of objects the forecast given by a tree will differ from observed values. During construction of the second tree, more attention is given to those objects which have a bigger error, with the purpose to reduce it. The constructed tree also will give an error for some objects, and the third tree should be constructed to reduce this error. The given procedure repeats the given number of times or until the error is bigger than the given acceptable value.

Errors can be seen the following way. A probability of choosing an object from initial sample is attributed to it. During the construction of the first tree, as well as in the bootstrap aggregation procedure, value of this probability is the same for all objects. On the following stages, the probability of the selection of each object changes. In PR problem, incorrectly classified objects receive an increment of probability on the given size. In RA problem, an increment of probability is proportional to the square of an error. Thus, for the construction of the next tree, a sub-sample of the given size is formed. The objects are chosen according to the current distribution of probabilities.

As researches show, the last described method allows to reduce an error of the collective forecast more than others do.

4. Decision trees in cluster and time series analysis problems.

In the previous paragraphs, regression analysis and pattern recognition problems were considered. Besides these problems, there are also other kinds of problems of the multivariate statistical analysis in practice. In this paragraph, we will consider cluster analysis and multivariate series analysis problems, which can also be decided with the help of decision trees. Thus, all positive features of the methods based on the decision trees (an opportunity of processing both quantitative and qualitative information, simplicity of interpretation) also attribute to these new problems.

4.1. The cluster analysis with the using of decision trees.

The cluster analysis problem (taxonomy, automatic grouping of objects according to similarity properties, unsupervised classification) consists in the following steps. Using results of observations, it is required to divide initial sets of objects on *K* groups (clusters) so that objects inside each group would be the much alike in some sense, while the objects of different groups will be as more as possible "different". It is necessary to understand the structure of the data better. For this purpose, we replace the large number of initial objects into a small number of groups of similar objects (figure 19).

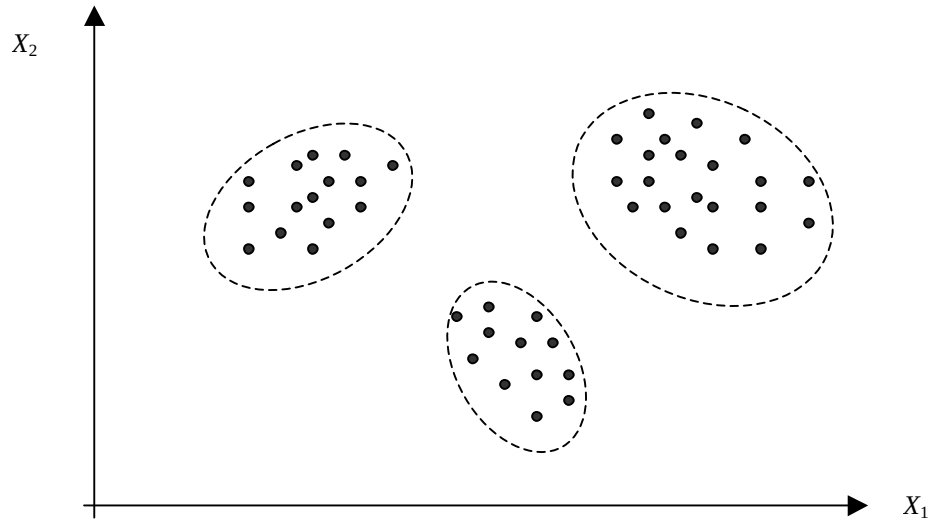


Fig. 19

Thus, it is required to find out such clusters of objects in space of characteristics, which will in the best way satisfy to a criterion of a grouping quality. It is supposed that the characteristics, describing objects, may be both quantitative and qualitative. Various methods of the cluster analysis differ in the ways of understanding of similarity, criterion of quality and ways of finding groups.

Let us solve a problem, using decision trees. At first we define a criterion of quality of the grouping. As already was marked in paragraph 1.1, the decision tree with *M* leaves splits space of characteristics into *M* non overlapping subareas $E^1,...,E^M$. This splitting space corresponds to the splitting of the set of observations *Data* into M subsets $Data^1,...,Data^M$.

Thus, the number of leaves in a tree coincides with the number of groups of objects: $K=M$. We will consider a group of objects $Data^i$.

The *description* of this subset will be the following conjunction of statements: $S(Data^i, \tilde{E}^i)$=«$X_1 \in \tilde{E}_1^i$» And «$X_2 \in \tilde{E}_2^i$» And... And «$X_j \in \tilde{E}_j^i$» And... And «$X_n \in \tilde{E}_n^i$» , where $\tilde{E}_j^i$ is interval $\tilde{E}_j^i = [\min_{Data^i}\{x_j\}, \max_{Data^i}\{x_j\}]$ in case of quantitative characteristic X_j or set of accepted values $\tilde{E}_j^i = \{x_j \mid x_j \in Data^i\}$ in case of the qualitative characteristic.

A characteristic subspace $\tilde{E}^i$, corresponding to the group's description, we call a taxon (plural taxa). In the example in figure 20, the plane is divided with the help of a decision tree on three subareas.

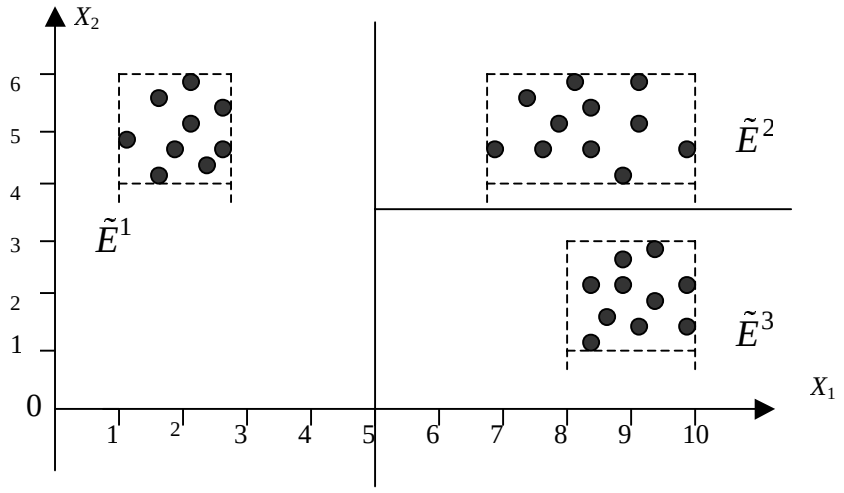


Fig. 20

S(Data^1, $\tilde{E}^1$)=«X$_1$ ∈ *[1,3]*» *And* «X$_2$ ∈ *[4,6]*»;
S(Data^2, $\tilde{E}^2$)=«X$_1$ ∈ *[7,10]*» *And* «X$_2$ ∈ *[4,6]*»;
S(Data^3, $\tilde{E}^3$)=«X$_1$ ∈ *[8,10]*» *And* «X$_2$ ∈ *[1,3]*».

It is important to note, although in a decision tree the part of characteristics can be absent, in the description of each group all available characteristics must participate.

Relative capacity (volume) of taxon is the next value

$$\delta^i = \prod_{j=1}^{n} \frac{|\tilde{E}_j^i|}{|D_j|} \quad ,$$

where symbol $|\tilde{E}_j^i|$ designates the length of an interval (in case of the quantitative characteristic) or capacity (number of values) of appropriate subset $\tilde{E}_j^i$ (in case of the qualitative characteristic); $|D_j|$ is the length of an interval between the minimal and maximal values of characteristic X_j for all objects from initial sample (for the quantitative characteristic) or the general number of values of this characteristic (for the qualitative characteristic).

When the number of clusters is known, the criterion of quality of a grouping is the amount of the relative volume of taxa:

$$q = \sum_{i=1}^{K} \delta^i$$

The grouping with minimal value of the criterion is called *optimum grouping*.

If the number of clusters is not given beforehand, it is possible to understand the next value as the criterion of quality,

$$Q = q + \alpha K ,$$

where α >0 is a given parameter.

When minimizing this criterion, we receive on the one hand taxa of the minimal size and on the other hand aspire to reduce the number of taxa. Notice, that in a case when all characteristics are

quantitative, minimization of criterion means minimization of the total volume of multivariate parallelepipeds, which contain the groups.

For the construction of a tree, the method of consecutive branching described in paragraph 2.3.3 can be used. On each step of this method, a group of the objects corresponding to the leaf of the tree is divided into two new subgroups.

Division occurs with a glance on criterion of quality of a grouping, i.e. the total volume of received taxa should be minimal. The node will be divided if the volume of the appropriate taxon is more than a given value. The division proceeds until there is at least one node for splitting or the current number of groups is less than the given number.

Additional to this method, the recursive method described in paragraph 2.3.4 can also be used. For this method, the second variant of quality criterion of grouping *Q*, for which the number of groups is not given beforehand, is used. All steps of algorithm of tree construction remain without changes, only the second quality criterion is used. Notice, that during the construction of initial tree, the large number of small volume taxa are being formed. These taxa are united into one or several taxa after the mending operation to improve criterion of quality of a grouping.

4.2. Decisions trees and multivariate time series analysis.

In this section we will consider methods for the solvution of problems of the analysis and forecasting of multivariate heterogeneous time series.

In many practical problems, it is required to predict values of characteristics of an object on the basis of the analysis of their values at the previous moments of time.

Now the theory is developed and the large number of various methods of the analysis of multivariate numerical sequences is created. However, application of these results for the decision of a considered problem in case of heterogeneous characteristics is impossible (since for qualitative characteristics, arithmetic operations on set of their values are not defined).

Using decisions trees, we can solve the specified problems.

Let for the description of a object of research the set of stochastic characteristics $X(t)=(X_1(t),...,X_n(t))$ be used. The values of the characteristics are changed in the run of time. The characteristics can be both of the quantitative and qualitative type.

Let characteristics be measured at the consecutive moments of time $t^1,..., t^\mu,...$. For definiteness we will assume that measurements will be carried out through equal intervals of time. We will designate through $x_j(t^\mu)=X_j(t^\mu)$ the value of characteristic X_j at the moment of time t^μ. Thus, we have *n*-dimensional heterogeneous time series $x_j(t^\mu)$, $j=1,...,n$, $\mu=1,2,...$.

Let us choose one predicted characteristic X_{j_0}, $1 \le j_0 \le n$.

We designate, for convenience, this characteristic through *Y*. The characteristic *Y* can be both quantitative, and qualitative type.

Let us consider the moment of time t^μ, and also a set of the previous moments of time $t^{\mu-1}$, $t^{\mu-2},..., t^{\mu-l}$, where *l* is a given size ("deep of history"), $1 \le l < \mu$.

We suppose that conditional distribution $Y(t^\mu)$, when all previous values $X(t)$ are given, depends only on values of series in *l* previous moments of time.

Besides, we suppose, that this dependence is the same for any value μ. The given assumption means, that the statistical properties of series determining dependence are stationary.

For any moment of time t^μ , it is possible to form a set $v^\mu = (X_j(t^{\mu-i}))$, $i=1,...,l$, $j=1,...,n$, representing the time series in *l* previous moments of time. We will call a set v^μ background of

length l for the moment t^{μ}.

It is required to construct a model of dependence of characteristic Y from its background for any moment of time. The model allows to predict the value of characteristic Y at the future moment of time on values of characteristics for l last moments. In other words, the given model, using background, represents decision function for forecasting.

Depending on the type of characteristic Y, we will consider various types of forecasting:

1. Y is the qualitative characteristic.

In analogy to a usual PR problem, we will call a problem of the given type a problem of recognition of dynamic object. The analyzed object can change its class in the run of time.

2. Y is the quantitative characteristic.

In this case, we have a forecasting problem of quantitative characteristic of object.

We will represent a decision function for forecasting time series on its background as a decision tree. This decision tree differs from the described in §2 trees in statements concerning a characteristics X_j in some i-th moment of time back are checked. For convenience, we will designate these characteristics, with a glance to background, through X_j^i (figure 21). Thus, X_j^i means characteristic X_j in i-th previous moment of time (concerning a present situation).

Let there be a set of measurements of characteristics $X=(X_1,...,X_n)$ at the moment of time $t^1,...,t^N$ and value l is also given. Thus, we have a multivariate heterogeneous time series of length N. We generate set of all histories of length l for the moments of time $t^{l+1},...,t^N$: $A= v^{\mu}$, $\mu=l+1,...,N$.

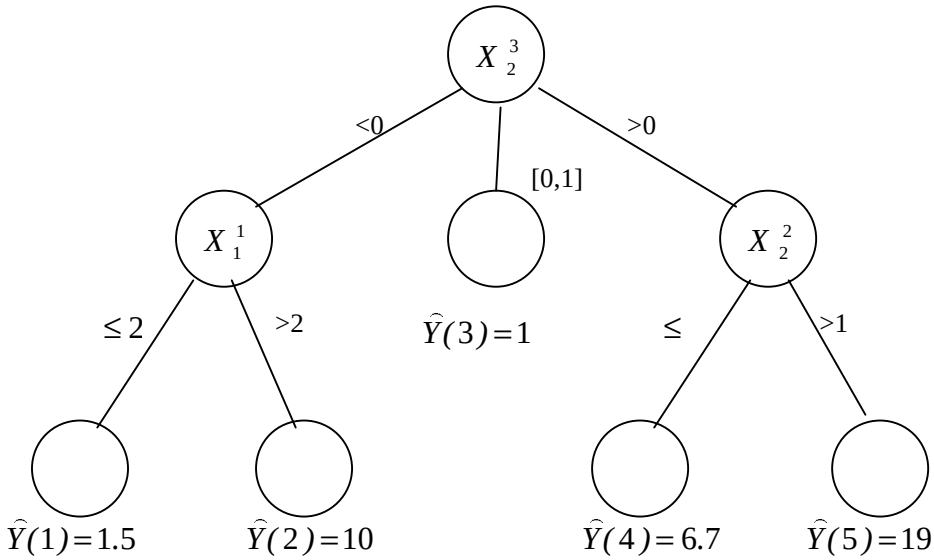


Fig. 21

For any given decision tree for forecasting by background, it is possible to define its quality similarly to how it was done for the usual tree: we will designate through $\hat{Y}(t^{\mu})$ predicted value Y received with the help of a tree by background v^{μ}. The criterion of quality will be

$$Q = \frac{1}{N-l} \sum_{\mu=l+1}^{N} h(\mu),$$

where $h(\mu) = \begin{cases} 0, & if \quad Y(t^{\mu}) = \hat{Y}(t^{\mu}) \\ 1, & otherwise \end{cases}$

for a recognition problem of dynamic object and

$$h(\mu) = (Y(t^{\mu}) - \hat{Y}(t^{\mu}))^2$$

for a forecasting problem of the quantitative characteristic.

The given series is used for learning to forecasting.

Let there be a series $x(t^{\mu})$ of length N_c, $\mu=N+1,...,N+N_c$. It is then possible to compare the predicted values $\hat{Y}(t^{\mu})$ **,** received as a result of training with "true" values $Y(t^{\mu})$ and to define an error of the forecast. We will say in this case, that the given series is used for control of qualities of forecasting.

How to construct a decision tree for forecasting by background on an available time series? Some ways are described below. The initial problem of construction of a decision tree is divided into some more simple pattern recognition or regression analysis problems, depending on type of predicted characteristic *Y*.

We will present set v^{μ} as the *table* $v^{\mu} = (X_j(t^{\mu-i}), i=1,...,l, j=1,...,n$ containing *l* rows and *n* columns. Then, the initial information for forecasting is the set of tables v^{μ}, together with the values of predicted characteristic *Y* specified for each table $Y(t^{\mu}), \mu=l+1,...,N$.

It is possible to present set $A = v^{l+1},..., v^{N}$ as the three-dimensional table of dimension $l \times n \times (N-l)$ to which the vector $(y^{l+1},...,y^{N})$ corresponds (figure 22). However, available methods of recognition or regression analysis with using of decisions trees use bi-dimensional tables as input information.

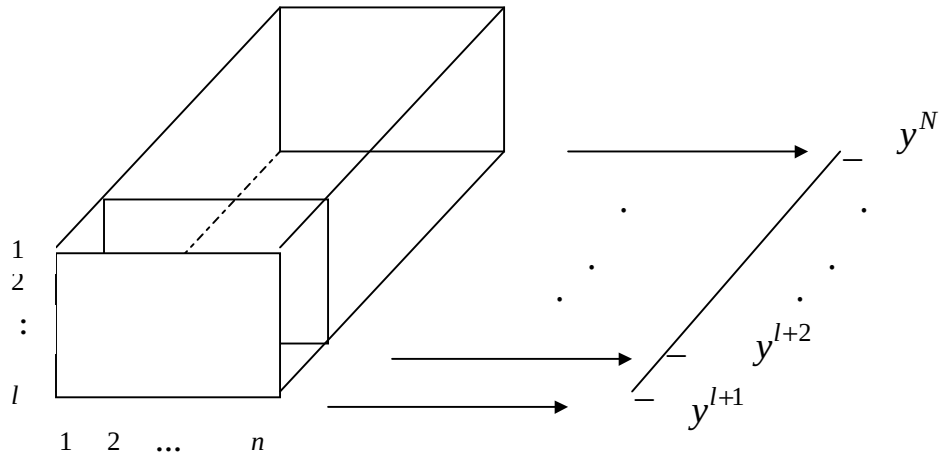


Fig. 22

There are various ways to use given methods for the analysis of three-dimensional data tables.

1. Each table v^{μ} is represented as a line of the appropriate values of characteristics $X_1^1, X_2^1, ..., X_n^1, X_1^2, X_2^2, ..., X_n^l$, $\mu=l+1,...,N$ (in other words, the table is stretched in line).

As a result, we receive the two-dimensional table of dimension l × n × (N–l), for which the decision function as a decision tree is constructed. For this purpose one of the methods described in §2 can be used.

As researches show, the given way is very simple in realization, however the received decisions, based on conditions such as: the number of characteristics is large, the background is long and the length of rows is small, can be unstable, i.e. will give the large error on a control series. It is known, that the effect of instability appears when the sample size is small and the number of characteristics is large.

2. The initial problem is solved in two stages. At the first stage are considered *l* two-dimensional tables of a kind $(X_j(t^{\mu-i}),Y(t^{\mu}))$ where $j \in \{1,2,...,n\}, \mu \in \{l+1,l+2,...,N\}, i \in \{1,2,...,l\}$

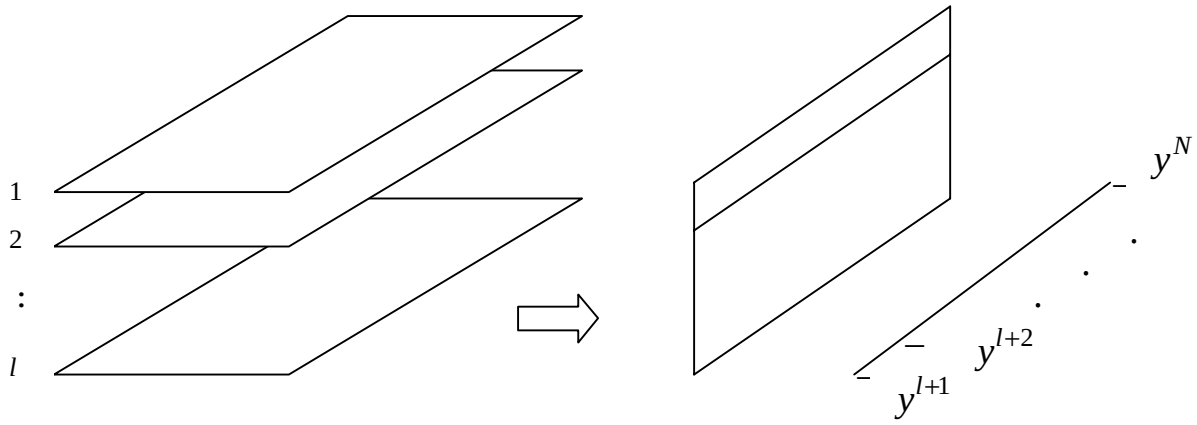


Fig. 23

We construct *l* various decisions trees for a prediction of value of *Y* on each of the given tables. Each of the tables, with the help of the constructed tree, transfer into an one-dimensional row of symbols (each symbol is coded by the number of the corresponding leaf of a tree). Thus, we get a two-dimensional table for which at the second stage the decision tree is again constructed. Then each symbol transforms back into the appropriate chain of statements.

3. The problem is solved in some stages.

1) We consider *l* tables $X_j(t^{\mu-i}),Y(t^{\mu})$, where

$j \in \{1,2,...,n\}, \mu \in \{l+1,l+2,...,N\}, i \in \{1,2,...,l\}$ and then *n* tables $X_j(t^{\mu-i}),Y(t^{\mu})$, where $j \in \{1,2,...,n\}, \mu \in \{l+1,l+2,...,N\}, i \in \{1,2,...,l\}$. Thus, we have *l* horizontal and *n* vertical cuts of the initial table (figure 24).

For each received two-dimensional table, the decision tree is constructed. In a result we receive a set of trees $T_1,T_2,...,T_{l+n}$.

We will denote the best of these trees as T^*.

2) All terminal nodes of a tree T^* are considered.

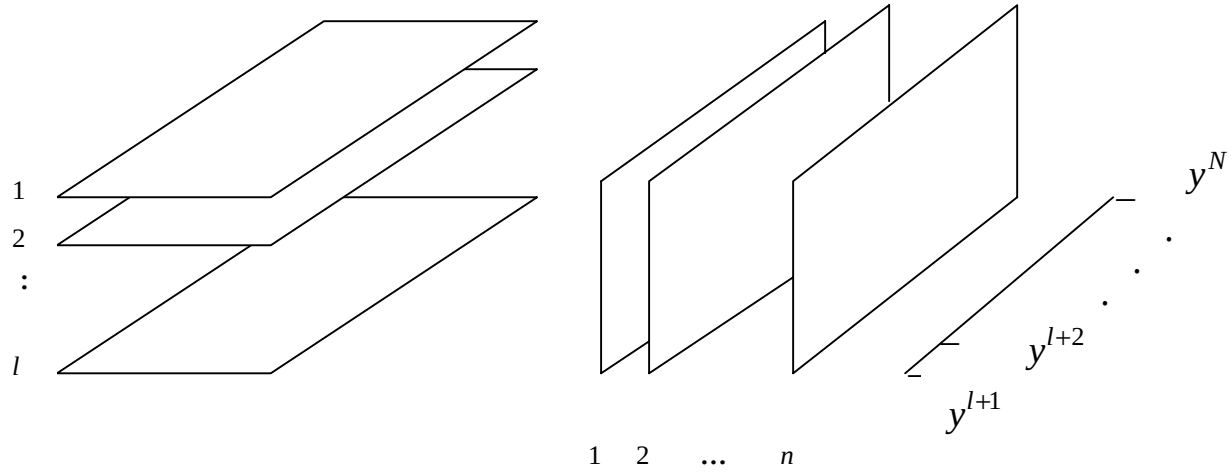


Fig. 24

The trees for which the error of forecasting exceeds the given size are selected. For each of these nodes the appropriate set of backgrounds $A' \subset A$ is formed and then we repeat the process of tree building for A' with the first step.

3) The process stops when the received tree will satisfy to a termination condition (i.e. when the number of leaves will be equal to given size M_{max} or when the error of forecasting will be less than the given value).

The given way differs from previous one in step-by-step tree building.

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.

3. 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.

REFERENCES

1) Breiman, L., Friedman, J., Olshen, R. and Stone, C. *Classification and Regression Trees* - CRC Press 1984.
2) Quinlan, J. R. *C4.5: Programs for Machine Learning* - Morgan Kaufmann 1993.

3) Lbov, G. and Berikov, V. *Recognition of a Dynamic Object and Prediction of Quantitative Characteristics in the Class of Logical Functions*, Pattern Recognition and Image Analysis. Vol 7, N 4, 1997, pp. 407-413.
4) Safavian, S.R. and Landgrebe, D.A. *A Survey of Decision Tree Classifier Methodology*, SMC(21), No. 3, May 1991, pp. 660-674.
5) R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, 1972.
6) Gelfand,S.B., Ravishankar, C.S., and *Delp, E.J. An Iterative Growing and Pruning Algorithm for Classification Tree Design*, PAMI(13), No. 2, February 1991, pp. 163-174.
7) Esposito, F., Malerba, D., *Semeraro, G., A Comparative-Analysis of Methods for Pruning Decision Trees*, PAMI(19), No. 5, May 1997, pp. 476-491.
8) Ho, T.K., *The Random Subspace Method for Constructing Decision Forests*, PAMI(20), No. 8, August 1998, pp. 832-844.