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).

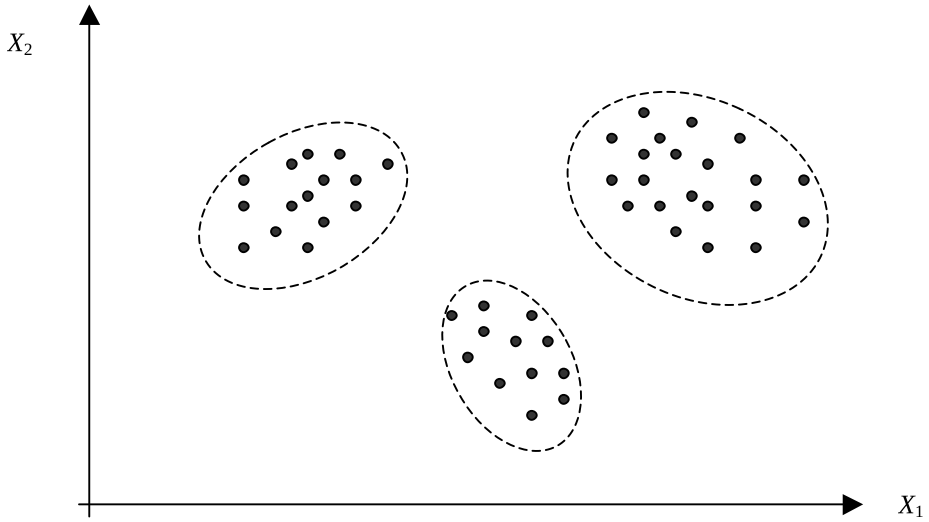

Fig. 19

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.

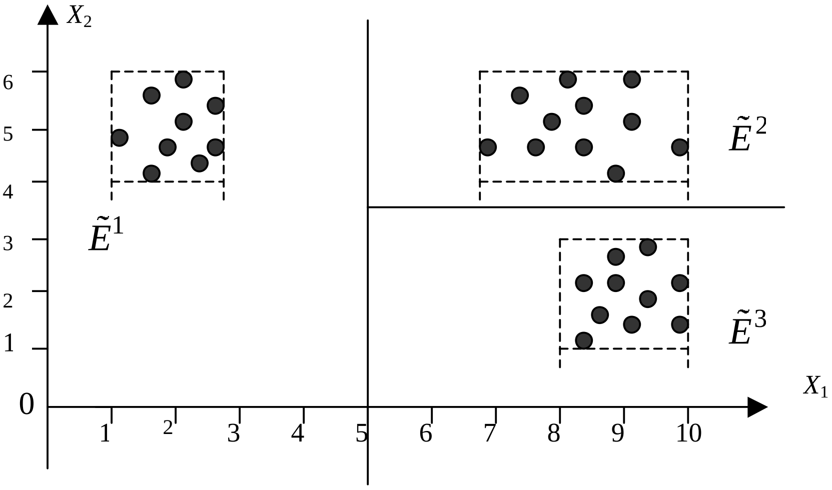

Fig. 20

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.