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).

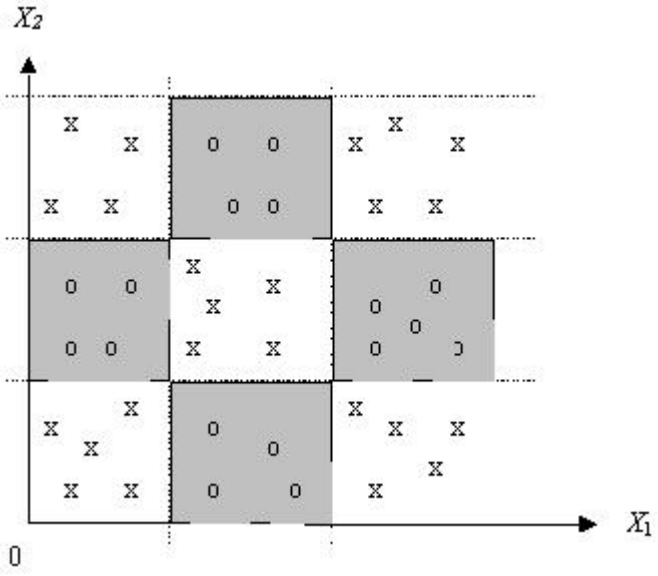

Fig. 17

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.