ASSESSMENT OF THE TRAINING SAMPLE QUALITY IN DATA MINING TASKS AND THE STRATEGY SELECTION FOR FINDING DECISIVE RULES

Polyakov M.V., Khanin I.G. , Shevchenko G.Y.

maxvp77@gmail.com, khanin.ig@gmail.com, nikk.gena@gmail.com Noosphere Company, Ukraine

An attempt was made in [1] to assess the training sample (TS) quality for constructing a predictive neural network. However, the requirements to TS for such an assessment reduce the effect of its application and do not provide an opportunity to assess the TS quality for more general cases that are much more frequent in practice. Undoubtedly, the TS quality is directly related to the choice of characteristics system for the objects description. Nowadays, almost the only way to assess the TS quality in thegeneral case is the assessment that is determined by the probability of TS sets correct recognition built by the decisive rule (DR). However, such an assessment obtaining requires a training procedure and a significant time consumption [2]. At the same time, it is desirable to know the TS quality, without training itself and indicate the TS limiting capabilities for objects identifiability. It is desirable to obtain a similar assessment for the examination sample (ES). In this case, different strategies for the DR searching are possible depending on the relationship between the values of such assessments for TS and ES. The proposed assessment is calculated directly from the TS (ES) data and characterizes the TS (ES) distinctive ability. The assessment is based on the formula application for calculating the characteristics informativeness that is given in [3] and studies in the field of nonparametric statistical problems [4].

References

1. Oleshko D. Building a quality training sample for predictive neuralnetwork models / D. Oleshko, V. Krisilov, A. Blazhko // Artificial Intelligence. – 2004. – № 3’2004. – P.567-573. (in Russian)

2. Dyukova E.V. Search for object descriptions informative fragmentsin discrete recognition procedures / E.V. Dyukova, N.V. Peskov // Journal ofComputational Mathematics and Mathematical Physics. – 2002. – № 5 (42). – P.741-753. (in Russian)3.

3. Vasilenko Y.A. Analytical method for tests finding / Y.A. Vasilenko,G.Y. Shevchenko // Automation. – 1979. – № 2. – P.3-7. (in Russian)4.

4. Kendall M.D. Statistical inferences and connections / M.D. Kendall,A. St`yuart – M.: Science, 1973. – 899 p. (in Russian)