

# **The results of solving tasks of medical diagnosis for dental patients on the basis of chemometrics methods**

*Delakova Ekaterina*

Strategically important tasks for modern science are improving the quality of medical diagnosis, using of new information technologies in the work with the results of biomedical and clinical research, choosing optimal methods for the analysis of multivariate and various medical data.

The data of biomedical researches are the multidimensional heterogeneous attributes, which have different qualitative and quantitative characteristics, varying in different ranges. These include laboratory patient tests, various indices of somatic health status or data device-computer systems. The results of solving tasks of medical diagnosis (classification and clustering tasks) based on chemometrics methods and comparative analysis of their effectiveness in relation to other information technologies are offered in this report.

The input data are the results of dental examination of patients of different age on morphofunctional state of the oral mucosa and lips (the status of each patient is characterized by about 60 parameters). These data were obtained on the basis of Pavlov First Saint Petersburg State Medical University in which researches in the field of evaluation of the effectiveness of methods of multivariate medical information data analysis are being conducted. The central chemometrics method - the principal component analysis (PCA) is used for studying the data structures. PCA - modeling revealed scorecard with effective dimensionality containing 7 principal components. Also we have identified the parameter which have the greatest influence on the separation of data to classes.

The Classification problem –is referring the patient to one of the pre-formed classes for baseline survey data. This task is solved with the help of formal analogies independent modeling classes method. The teaching model was build and this model proved its efficiency in testing: when using three principal components, the key measure of model quality (residual dispersion) was 2%. The Recognition procedure is carried out, the condition of the patients is rightly assigned to the Class " presence of the disease " modeling error was 4%. The comparative analysis of the effectiveness of classification models is carried out. As an alternative models the neural network technology, the algorithm KNN and the Bayes classifier were taken.

The method of cross-validation and methodology ROC-curves designing were used to assess the quality of each classification model. According to the results of cross-validation conventional indicators of statistical evidence [1] were calculated. There were sensitivity, specificity, likelihood ratio and predictive values for the presence or absence of disease. After comparing the obtained results SIMCA method showed the highest efficiency: sensitivity value (Se) was 90%, specificity value (Sp) was 92%, predictive value (PPV, NPV) of 97% and 94%, respectively. Index AUC [2] was almost 90%, which also corresponds to a high quality model.

The obtained results showed high efficiency of chemometrics methods for solving problems of analysis, classification and clustering for the work with the data of biomedical researches.

1. Olive Jean Dunn, Virginia A. Clark: BASIC STATISTICS A Primer for the Biomedical Sciences // 2000 A JOHN WILEY & SONS, INC., PUBLICATION.
2. Fawcett T. ROC Graphs: Notes and Practical Considerations for Researchers // 2004 Kluwer Academic Publishers.