AMEFSS 2022

Improving the Accuracy of the Machine Learning Predictive Models for Analyzing Medical Datasets

Ivan Ivanov, Borislava Toleva
Faculty of Economics and Business Administration
Sofia University "St. Kl. Ohridski"
i_ivanov@feb.uni-sofia.bg, bvrigazova@gmail.com

In this article, we will examine the problem of classifying big data sets on specific medical sets of observations. These sets are characterized by a relatively small number of observations and the fact that they are imbalanced, which means that one class of observations contains a very small number of observations compared to the other class of observations. In the general case, the task of classifying imbalanced sets of observations is relevant and very intensively studied [1-4] and many others. For such tasks, specific models, methods, and algorithms for building a classification model must be applied. The application of generally accepted models for classification problems usually does not predict all classes of the set equally well. The issue of forecasting all classes well enough is important since the "small" class of observations describes important cases in practice. In medicine, these are sick patients for whom it is important to predict the development of their disease, for which purpose the classification model is used.

Pima Indians Diabetes (PID) (*https://networkrepository.com/pima-indians-diabetes.php*), Haberman Breast Cancer (*https://archive.ics.uci.edu/ml/datasets /haberman%27s +survival*), and Chronic Kidney Disease (*https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease*).

The proposed algorithm shows high efficiency on these sets with medical data. The results are not inferior to the analyzes of other authors in recent years and at the same time compete with them. The algorithm is characterized by its naturalness, fast and easy programming. The models predict more accurately and better recognize the "weaker" class in the multitude of observations, and this is an important advantage.

Acknowledgements.
This work was supported by the Sofia University St. Kl. Ohridski project for 2022.

References

1.Ivan Ivanov, Borislava Toleva, Nikolay Netov, Modified Training Models Based on Medical Data Sets, ICEST'2022,Ohrid, North Macedonia, 2022.
2.Ivan Ivanov, Borislava Toleva, Algorithms for classification analysis for imbalanced medical datasets, International Scientific Conference „Management and Engineering '22" (ISCME), Sozopol, Bulgaria.