Alternative Evaluation Metrics for Machine Learning Model Selection in Ionospheric VLF Amplitude Data Exclusion

Filip Arnaut and Aleksandra Kolarski

Institute of Physics Belgrade, University of Belgrade, Pregrevica 118, 11080 Belgrade, Serbia
E-mail: filip.arnaut@ipb.ac.rs

When applying machine learning (ML) methods to classify ionospheric VLF amplitude data for data exclusion, there is a significant imbalance in the ML task. Specifically, the proportion of non-anomalous data to anomalous data is 85- 15% in our example. Commonly used ML evaluation metrics include accuracy, precision, F-measure, recall (Powers, 2020) among others. Standard evaluation metrics for imbalanced ML tasks can yield subpar results, requiring careful interpretation in relation to the distribution of the test dataset. This communication attests to the selection of the Random Forest (RF) model and discusses the inclusion of additional evaluation metrics, including Youden's J statistic, Markedness, General Performance Score (GPS) (De Diego et al. 2022), and Unified Performance Measure (UPM) (Redondo et al. 2020). According to Youden's J statistic, Markedness, GPS, and UPM, the previously selected model with 100 trees is the best overall model, with values of 0.692, 0.673, 0.776, and 0.833, respectively. Furthermore, the interpretation suggested that there was little difference between the models, which is supported by the additional evaluation metrics (the biggest discrepancy can be seen in Markedness at 2.1%). However, the model with 100 trees had the highest evaluation metric values and the fewest (hyper)parameters, making it the most preferable option. Additional evaluation metrics should be incorporated in further research on the utilization of ML methods for automating data exclusion, which will provide a more comprehensive understanding of the model.

References

De Diego, I. M., Redondo, A. R., Fernández, R. R., Navarro, J. and Moguerza, J. M., 2022. General Performance Score for classification problems. *Applied Intelligence*, *52*(10), pp.12049-12063.

Powers, D. M., 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies,* 2(1), pp. 37-63.

Redondo, A. R., Navarro, J., Fernández, R. R., de Diego, I. M., Moguerza, J. M. and Fernández-Muñoz, J. J., 2020. Unified performance measure for binary classification problems. In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 104-112). Cham: Springer International Publishing.