Comparative Analysis of Random Forest and XGBoost in Classifying Ionospheric Signal Disturbances During Solar Flares

Filip Arnaut (filip.arnaut@ipb.ac.rs), Aleksandra Kolarski, Vladimir Srećković

Institute of Physics Belgrade, Laboratory for astrophysics and physics of ionosphere, University of Belgrade, Pregrevica 118, 11080, Belgrade, Republic of Serbia

This presentation participates in OSPP

Outstanding Student & PhD candidate Presentation contest

Introduction

This study expands upon prior research on anomaly detection in VLF amplitude data resulting from solar flares and other disruptions by comparing the efficacy of the XGBoost (XGB) to the Random Forest (RF) algorithm.

Problem Statement: Although the RF algorithm has been shown to be effective in classifying disturbances in VLF amplitude data (Arnaut et al. 2023), it is necessary to investigate and assess alternative algorithms such as XGBoost in order to potentially improve the predictive power of anomaly detection in this VLF amplitude data.

Previous Solutions: The previous method utilized the RF algorithm to detect anomalies in VLF amplitude data, taking advantage of its simplicity and capacity to prevent overfitting. This approach demonstrated encouraging results in categorizing different disturbances and erroneous datapoints.

Why Current Research: This research aims to evaluate and compare different machine learning methods, specifically XGBoost, with the previously utilized RF algorithm. The objective is to identify any potential benefits or improvements in correctly identifying anomalous VLF amplitude data.

Methods and data

Data: The data employed for this study consists of VLF amplitude measurements obtained during the months of September and October of 2011. The transmitters and receivers utilized were NAA, NAU, NLK, NMP, NML and Oklahoma East and South, Sheridan and Walsenburg. The data processing workflow is displayed in Figure 1 and can be fully seen in Arnaut et al. (2023).

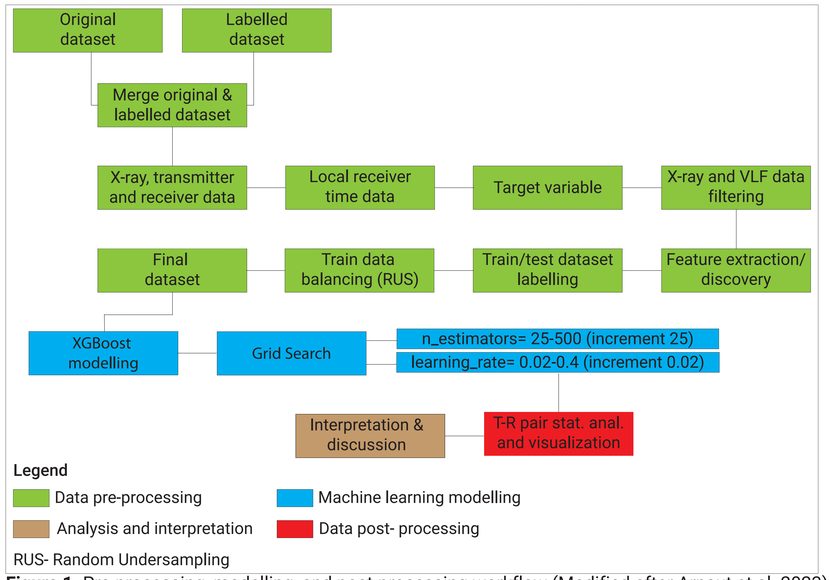

Figure 1. Pre-processing, modelling, and post-processing workflow (Modified after Arnaut et al. 2023)

Hyperparaemter tuning: The XGB model (Chen and Guestrin, 2016) hyperparameter tuning method used in this research was grid search. The number of estimators varied from 25 to 500, with increments of 25. The learning rate ranged from 0.02 to 0.4, with increments of 0.02 (Figure 1).

Feature discovery (VLF amplitude and X-ray data): The model incorporates various statistical features, including rolling window statistics such as mean, median, and standard deviation for different window lengths. It also includes lagged signal features for different lag lengths, as well as first and second differences.

Results and discussion

Figure 2 illustrates the comparison between the true data labels and the data labels obtained from the RF and XGB models. All three graphs depict a single solar flare of C2.4 magnitude, accurately classified by the RF model. In addition, the XGB model accurately identified the signal interruption, or erroneous signal, while the RF model classified those data points as normal. Both the RF and XGB models accurately identified and labeled the signal interruption. It is important to mention that the XGB model classified all data points as anomalous signal until the return to normal signal. On the other hand, the RF model showed some ability to differentiate, therefore, in this particular example, the RF model yielded superior results.

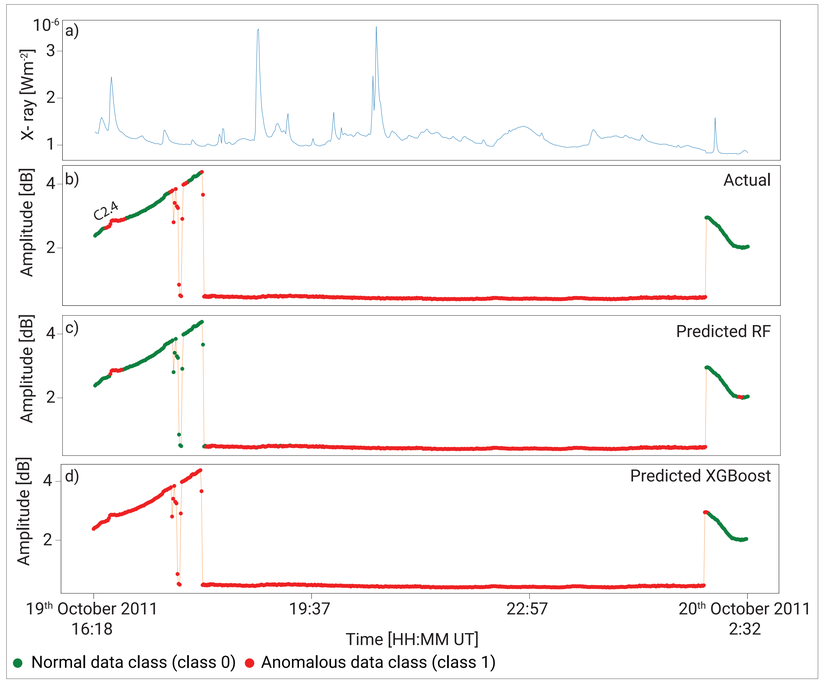

Figure 2. Transmitter- receiver pair NPM Walsenburg from 19th October, 2011 16:18 UT to 20th October, 2011 02:32 UT; (a) X-ray irradiance data; (b) Actual class labels; (c) Predicted random forest (RF) class labels; (d) Predicted XGBoost class labels (Expanded after Arnaut et al. 2023).

The provided example in Figure 3 depicts the NPM OklahomaEast transmitter-receiver pair's activity from 15:47 UT on October 21, 2011, to 00:07 UT on October 22, 2011. The example serves as a useful reference for comparing models, the VLF amplitude signal contains outlier data points, the effects of solar flares, and erroneous signal values. Regarding the outlier data points, the RF algorithm did not categorize them as anomalous, whereas the XGB algorithm did. Both the RF and XGB models accurately identified the impact of the solar flare. However, it is worth noting that the XGB model classified a greater number of data points as anomalous in that specific part of the signal. The XGB model accurately identified the erroneous measurements with rapid signal variations and classified them as anomalous. During the later portion of the signal, the XGB model exhibited false anomalous labels.

When examining the F1-Score for the anomalous class (class 1), it is evident that, on average, the XGB model shows a 7% improvement. However, both models have instances where they perform better than the other. For the NPM-Walsenburg transmitter-receiver pair, the RF model showed an improvement of approximately 0.16 in F1-Score. However, for the NPM-OklahomaEast, the XGB model showed a greater F1-Score by about 0.15.

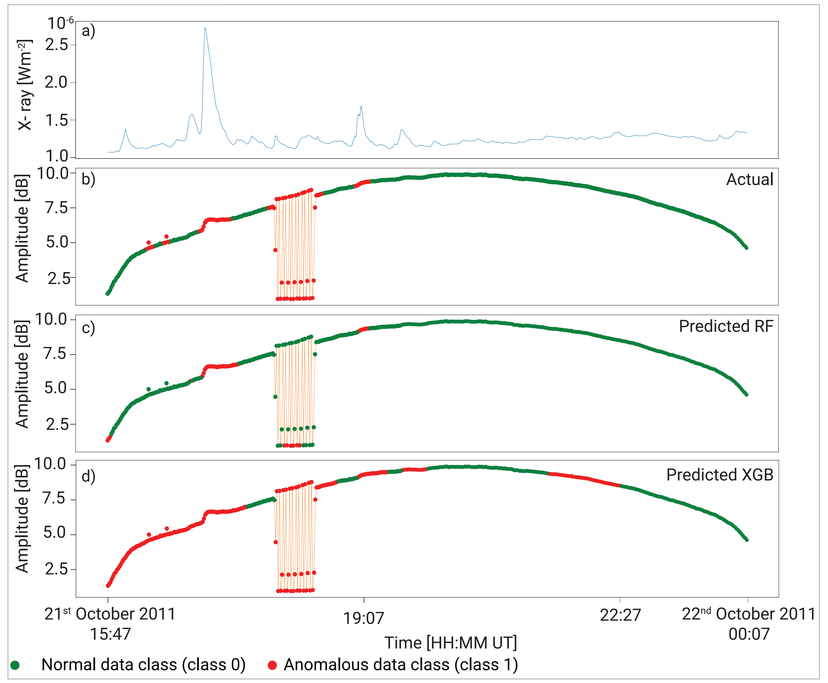

Figure 3. Transmitter- receiver pair NPM OklahomaEast from 21st October, 2011 15:47 UT to 22nd October, 2011 00:07 UT; (a) X-ray irradiance data; (b) Actual class labels; (c) Predicted random forest (RF) class labels; (d) Predicted XGBoost class labels.

Possible additional research could involve developing a multi-class system that identifies anomalous signals, which may consist of erroneous measurements (as depicted in Figures 2 and 3), outlier data points (as shown in Figure 3), and the impacts of solar flares. The proposed system has the potential to improve the model by firstly providing more detailed information about the specific type of anomalous VLF signal being classified, and secondly by generating more accurate outputs for the researcher. This system requires additional data, and the process of labeling the data is laborious and time-consuming.

Conclusion

Overall, the comparison between the RF and XGB models is ambiguous. Both models have instances where one is superior to the other. Further research is necessary to fully optimize the method, which has benefits in automatically classifying VLF amplitude anomalous signals caused by SF effects, erroneous measurements, and other factors.

Acknowledgement

This work was funded by the Institute of Physics Belgrade, University of Belgrade, through a grant by the Ministry of Science, Technological Development and Innovations of the Republic of Serbia.

VLF data are provided by the WALDO database (https://waldo.world, accessed on 1 January 2023), operated jointly by the Georgia Institute of Technology and the University of Colorado Denver, using data collected from those institutions as well as Stanford University, and has been supported by various US government grants from the NSF, NASA, and the Department of Defense.

References

Arnaut, F., Kolarski, A. and Srećković, V.A., 2023. Random Forest Classification and Ionospheric Response to Solar Flares: Analysis and Validation. Universe, 9(10), p.436.

Chen, T. and Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). New York, NY, USA: Association for Computing Machinery, pp.785-794. Available at: https://doi.org/10.1145/2939672.2939785.