[https://doi.org/10.69646/aob241203]
[Invited Lecture]

Daytime and nighttime VLF signal classification utilizing machine learning methods

Filip Arnaut[1*]

[1]Institute of Physics Belgrade, University of Belgrade, Pregrevica 118, Belgrade, Republic of Serbia

*Correspondence: filip.arnaut@ipb.ac.rs

Abstract: The automatic classification of ionospheric very low frequency (VLF) signals is a current research endeavor aimed at creating a machine-learning (ML) methodology capable of differentiating among various influences on VLF signals, including solar flares, VLF receiver malfunctions, nighttime VLF signals, and other factors. This communication discusses the enhancement in ML classification of daytime and nighttime ionospheric VLF signals, including the different methodologies, data processing, and various processes that demonstrated improvement over prior research.

Keywords: Machine learning, data-driven modelling, anomaly detection, geophysics.

Introduction

The automatic classification of very low frequency (VLF) signal disturbances and characteristics through machine learning (ML) techniques is a subject previously presented (Arnaut et al., 2023; Arnaut, Kolarski, 2023; Arnaut et al. 2024) and remains an area of active research. VLF signals generally capture the impacts of solar flares, instrument malfunctions, erroneous measurements and other, as well as the variation between daytime and nighttime signals.

Typically, the nighttime signal exceeds the daytime signal, with the exception of the terminator, which is characterized by a minor decline in the signal prior to a pronounced ascent, ultimately stabilizing at a level exceeding that of the daytime signal. The ongoing research focuses on the automatic classification of diverse effects on the VLF signal, aiming to establish a method that minimizes error in this classification process. This research demonstrates the advancements in the classification of daytime and nighttime VLF signals, the modifications implemented during this period, and the resultant outcomes.

Methods and data

The primary distinction in the methodology between prior research and the current study is that data labeling in the latter was conducted on an individual case basis. Daytime and nighttime conditions were previously established through conditional labeling based on the local receiver time. The transition from imprecise conditional labeling to manual data labeling enhanced data quality, thereby improving the predictive capabilities of models and facilitating a more accurate differentiation between daytime and nighttime conditions.

Secondly, the feature list was expanded to include weighted moving averages, in addition to the previously employed statistical features, whereby data points nearer to the instance under analysis are assigned a greater weight coefficient. Furthermore, the class balancing approach of random undersampling was replaced with the Synthetic Minority Oversampling Technique (SMOTE), which ensures that no data is discarded, as occurs in random undersampling. The random forest model was ultimately replaced by the extreme gradient boosting (XGB) model, utilizing the random search hyperparameter tuning method for tuning of the number of estimators and the learning rate.

The model's predictions underwent cluster analysis, as the nighttime signal extends over a longer duration; clusters of relatively few consecutive nighttime labels were reclassified using the cluster analysis. The outcomes for both raw model predictions and predictions reclassified through cluster analysis are presented in this study.

Results and discussion

As previously stated, the training dataset was balanced using the SMOTE technique, which ensured that no data points were omitted, while the minority class (specifically, the nighttime signal class) was oversampled. The original distribution in the training dataset was 70-30, favoring the daytime signal; therefore, SMOTE was employed to oversample the minority class.

The hyperparameter tuning method chosen was random search for both the number of estimators (ranging from 100 to 1000 in increments of 20) and the learning rate (ranging from 0.01 to 0.2 in increments of 0.01). The most effective model exhibited 840 estimators and a high learning rate of 0.2. The initial model exhibited accuracy, precision, and F1-score values of 0.79, 0.71, and 0.78, respectively, while the AUC value was 0.8, indicating an acceptable capacity of the model to differentiate between the classes.

Figure 1 illustrates a dataset comprising 1000 data points, specifically in minute intervals, pertaining to the NLK-Sheridan transmitter-receiver pair, where the nighttime signal is represented (true class labels are shown in the upper panel and predicted labels in the middle panel). The model's raw output exhibited relatively satisfactory classifications, demonstrating moderate predictive capability to differentiate between nighttime and daytime VLF conditions in the provided example. The classification of the daytime-to-nighttime terminator and the segment of the signal immediately following the nighttime-to-daytime terminator could be further improved as to align more with the true classifications.

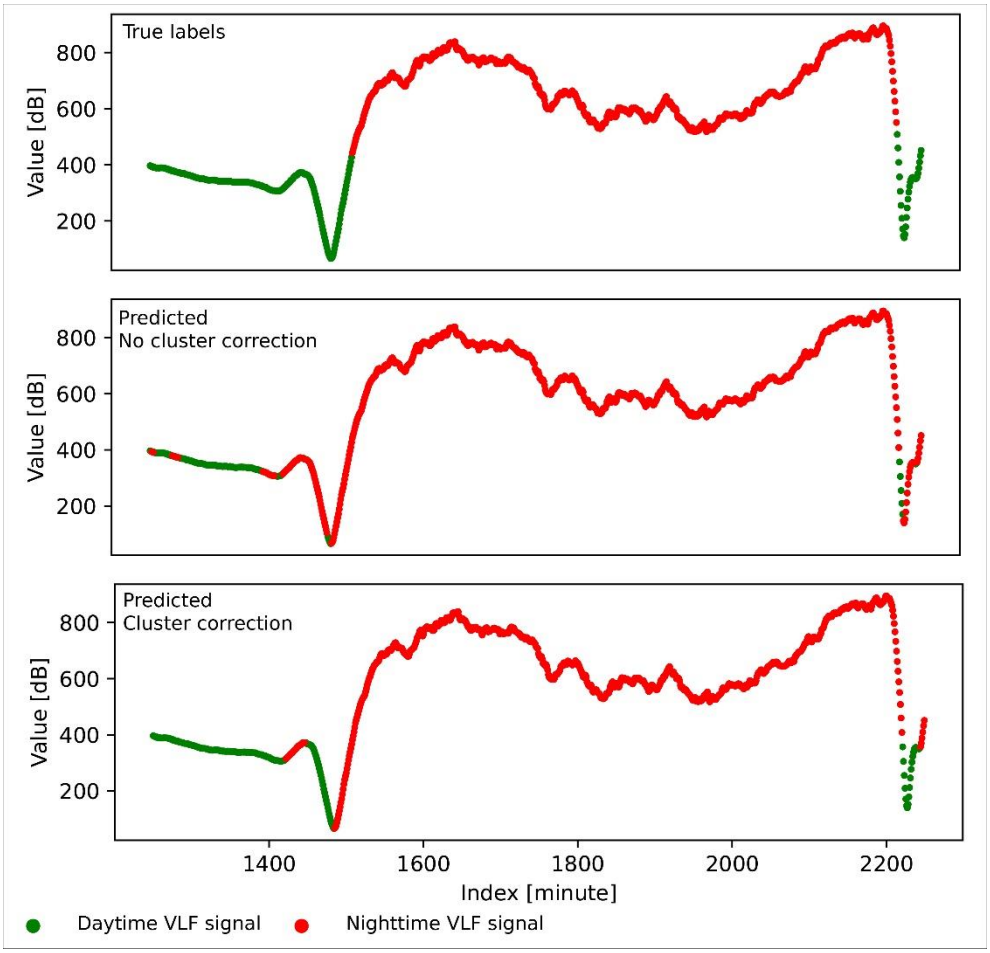

Figure 1. True daytime and nighttime labels (upper panel); Predicted labels without the cluster correction (middle panel); Predicted labels with the cluster correction (lower panel)

Cluster analysis was utilized to correct the minor groups of nighttime labels in the predicted signal. The cluster analysis statistics revealed a significant positive skew in the distribution of cluster lengths, suggesting that the model generates a considerable number of brief nighttime predictions. Conversely, the nighttime signal is expected to be present for a longer duration; consequently, all clusters with a consecutive prediction length of fewer than 25 were

reclassified as daytime signals. Figure 1 (bottom panel) illustrates the reclassification of small clusters previously identified as nighttime signals to daytime signals. The cluster analysis functioned as a corrective measure in this instance and exhibited satisfactory characteristics.

The comparison of evaluation metrics between cluster-corrected and non-cluster-corrected values reveals relatively similar results, occasionally favoring the cluster-corrected values (Table 1). The precision and F1-score for the nighttime class exhibit increased values, whereas the recall parameter shows a lower value for the nighttime class in the cluster-corrected predictions.

Table 1. Comparison between evaluation metrics for the non-cluster corrected and the cluster corrected predictions

	Prec. NC	**Prec. C**	**Rec. NC**	**Rec. C**	**F1 NC**	**F1 C**
Daytime class	0.88	0.86	0.74	0.82	0.8	0.84
Nighttime class	0.71	0.77	0.86	0.82	0.78	0.8
Macro averaged	0.8	0.82	0.8	0.82	0.79	0.82
Weighted averaged	0.81	0.82	0.79	0.82	0.79	0.82

NC- No cluster C- Cluster

The integration of the cluster correction yielded satisfactory results; however, it requires additional refinement to fully automate the process and produce improved, more precise predictions.

Conclusions

The complete automation of ionospheric VLF signal classification will require considerable time and extensive research effort. This communication presents the enhancement of ML classification for daytime and nighttime VLF signal conditions. The improvement was achieved through case-by-case manual labeling, replacing random undersampling with SMOTE for training dataset

balancing, substituting the random forest model with the XGB model, expanding the feature list and employing a cluster correction after the classification process. The results are promising; however, additional refinement and improvement is necessary, which will be the focus of subsequent research.

Acknowledgement

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., 2023. Machine learning approach for distinguishing daytime and nighttime ionospheric conditions on VLF signals related to solar flares during 2011. In XX Serbian Astronomical Conference, 16-20 October 2023, Belgrade, Serbia. Book of Abstracts, Astronomical Observatory of Belgrade and Faculty of Mathematics, pp. 79

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.

Arnaut, F., Kolarski, A. and Srećković, V.A., 2024. Machine Learning Classification Workflow and Datasets for Ionospheric VLF Data Exclusion. Data, 9(1), p.17.