Ann. Geophys. Discuss.,
https://doi.org/10.5194/angeo-2019-149-AC1, 2019

Annales Geophysicae Discussions

Interactive comment on "Automatic detection of the Earth Bow Shock and Magnetopause from in-situ data with machine learning" *by* Gautier Nguyen et al.

Gautier Nguyen et al.

gautier.nguyen@lpp.polytechnique.fr

Received and published: 28 November 2019

We thank the referee for the numerous comments regarding our methodology and the various suggestions proposed to better our work. Please find below the detailed answers to the referee's questions and comments.

1) The authors seem to be aware that a random splitting of the data between training and test set is yields erroneous scores (line 111). Yet, they still present results based on random split. I suggest to completely remove the results obtained in this way and to present only the results obtained with a more correct split in time.

As shown in Table 1, we obtain the same score for both random and temporal split. Thus, we do not expect erroneous scores on the prediction even with a random split. We prove it by attaching to our answer the ROC curve we obtain for THEMIS in the case of a temporal split. This attached figure could eventually replace our Figure 3. Additionally, the whole labelled dataset is used to train the final model we use for the massive detection of magnetopause and bow shock crossings. There is then no risk of erroneous predictions due to this random split in the framework of the massive detection. We are aware that this specificity is not specified yet and this shall be the case in a revised version of the paper.

2) The labeling of the data is completely unclear. Reading from line 53 it seems that they are mixing the 'ground truth' with the result of the classification algorithm. The same argument is repeated on line 118. Obviously you cannot use the same algorithm to label and predict.

The explanation provided in the paper is a bit unclear and shall be modified in the revised version of the paper. We started our work with 2 continuous hours of each of the 3 regions and trained a first algorithm with this 6 cumulated hours of data. We then augmented the quantity of labelled data by adding the predictions of this first model on additional data intervals with a possible manual correction of the mis-predicted points. A second model was then trained with this enlarged labelled dataset. The operation was used to speed up the labelling process, and was repeated until we reach the scores that are shown in the paper.

3) Similarly, on line 134 they do not explain how do they label such large amount of Cluster data. Similar for other data.

We applied the same process as described above for every missions. This point shall be specified in a revised version of the paper.

4) In evaluating the model performance, the authors focus exclusively on the AUC. I suggest to compute and show also the True Skill Score and the Heidke

Skill Score, that are standard skill scores.

The TSS and the HSS both apply to a specific threshold/cutoff (and thus to a specific confusion matrix) while the ROC curve and associated AUC apply to the full range of thresholds. The information brought by these two metrics is then already shown with the Figures 3 and 8 for every decision threshold.

5) Figure 3 is not very informative. It looks like the model can achieve perfect predictions?

The model can indeed achieve almost perfect predictions (the AUC would have been 1 in the prefect case) but still make errors especially when a spacecraft comes across the boundary between two distincts regions. This is what is shown in Figure 8 and explained in section 5.1.

6) Line 75: a reference is needed to the original works on boosting algorithms.
7) Line 76: the compute time depends on the CPU (and/or GPU) We agree with the referee that an original reference and the characteristics of the CPU we used will be specified in the paper. For instance, we used an AMD ryzen threadripper 2990wx processor.

8) Line 81: should be 'has been predicted'

This will be modified in a revised version of the paper

9) Line 197: why a decision tree should require less time than an arbitrary decision boundary? We agree with the referee that the prediction time will be unchanged whether it be a boosted ensemble of decision trees or an arbitrary decision boundary. The difference stands in the fitting time of the Gradient Boosting compared to the time required to define the arbitrary boundary that provides the best output. We shall make this point clearer in the revision.

10) Line 199: Decision trees are also threshold-based methods. I do not understand this distinction.

By "threshold-based methods", we mean "manually-set thresholds" that do not have the ability to go as much into precise decision steps as Decision trees and that are not based on sound statistical properties of the dataset but rather on empirical subjective knowledge of the operator. We agree with the referee that the distinction as mentioned in the paper is a bit unclear and shall be emphasized by the addition of the mention "manually set".

11) Line 203: don't you have cross-calibration issues when you employ the algorithm trained on one satellite to make inference on another satellite?

Cross-calibration issues can occur when switching from a mission to another or when an instrument switches from a mode to another and this is the reason why we only kept Cluster data when the HIA instrument was under the magnetosphere or magnetosheath mode (l.133).

12) Since decision trees are easily interpretable, it would be interesting to visualize the boundaries and understand their physical implications.

We agree with the referee about this statement. However, the problem is currently 8D and such boundaries would then hardly be interpretable. From then on, we have two options that are both not especially the best to provide physical interpretations: projections in specific (e.g. the main ones) 2D features planes that will struggle to give a global vision of this boundary Using the principal components, which will provide a global vision of the boundary between the different classes in a feature space that have no real physical meaning The three different regions could eventually be studied from a massive statistical point of view and this would be the logic aftermath of our massive detection in an upcoming work.

13) Line 226: it is well-known that the probabilities output of boosted ensemble of decision trees are not well-calibrated.

The probability calibration of Boosted ensemble of decision trees is indeed well-known

(Niculescu-Mizil et al.) and something we haven't mentioned in the paper, we ensured that in our case, our probabilities were decently calibrated and we attach the calibration curves for each of the three classes.

14) Line 275: Do Appendixes B and C really have no text?

Appendixes B and C are just set to represent ROC curves and a 2D histogram of B and Np in the case of Artemis. A small sentence similar to the one in Appendix A will be added to refer to the associated figures.

15) Finally I suggest to be more specific in the title, by replacing 'Machine Learning' with 'Decision Trees'

The title could indeed be modified accordingly. Nevertheless, we want to focus on the benefits of applying machine learning to this specific problem rather than focusing on the possible (and yet to be evaluated) specific benefits of Gradient Boosting classifiers against other algorithms. We would then be more eager to keep the current title.

Interactive comment on Ann. Geophys. Discuss., https://doi.org/10.5194/angeo-2019-149, 2019.

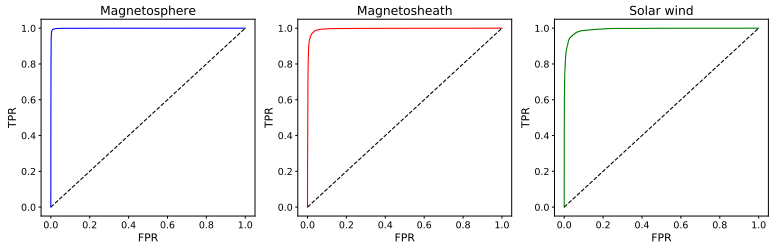

Fig. 1. ROC curve obtained with a temporal split for our model trained on THEMIS data

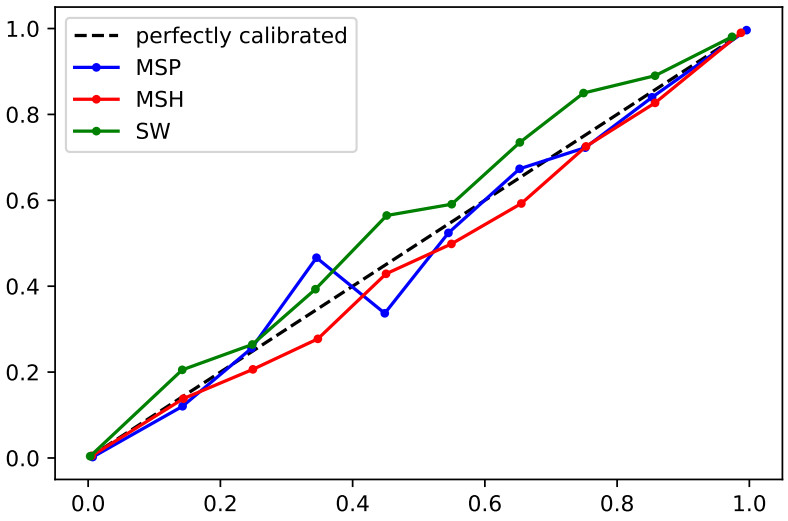

Fig. 2. Probability calibration curve for our model trained on THEMIS data