Ann. Geophys. Discuss.,
https://doi.org/10.5194/angeo-2019-149-RC2, 2019

Annales Geophysicae Discussions
Open Access
EGU

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

Anonymous Referee #1

Received and published: 28 November 2019

I thank the authors for answering my questions. Unfortunately, it is now clear that what I suspected were a major flaw of this paper is indeed an obvious mistake in the methodology, which explains the almost perfect prediction accuracy reported. In my opinion, this paper has no value and should be rejected.

The problem is with the automatic labeling of data, starting with a minuscule portion of labeled data. This is in essence what you are doing (of course, correct me if I am wrong). You manually label a very small portion of the data («1%). The decision tree partitions your 8-dimensional input space in a given way. Each partition (8-D hypercube) is then assigned to the majority class (the class that has the majority of

Printer-friendly version

Discussion paper

training points in that portion, gets it all. Note that by starting with 6 hours = 360 points in 8 dimensions, you have about 2 points per dimension!). You will agree that by no means this is expected to be a good classifier for the rest of the data. Then you take some unlabeled data, classify them with this decision tree and put them in the training set. If you now train another decision tree with the new enlarged training set, what do you expect? The new classifier will be pretty much identical to the old one, but now each partition will have a larger majority class (because the new training points you have added all belong to the majority class). So the classifier is more confident. You repeat this procedure a number of times, sure enough each leave of the decision tree ends up having a ratio of positive/negative close to 100% or 0%. You have cooked up a perfect classifier (ie perfect with respect to the labels that it has itself produced!)

This is nonsensical.

If you don't have labeled data you are better off with an unsupervised technique.

In any case, I urge you to collaborate with a machine learning expert, to avoid wasting your time with other silly mistakes. Good luck!

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

Printer-friendly version

Discussion paper