Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2020-78-RC2, 2020

Atmospheric Measurement Techniques Discussions

Interactive comment on "Mixing height derivation from aerosol lidar using machine learning: KABL and ADABL algorithms" *by* Thomas Rieutord et al.

Anton Sokolov (Referee)

anton.sokolov@univ-littoral.fr

Received and published: 5 May 2020

In the paper "Mixing height derivation from aerosol lidar using machine learning: KABL and ADABL algorithms" two machine learning algorithms presented for the definition of the atmospheric boundary layer height, which is a principal parameter for the atmospheric modelling and air pollution dispersion. The detection of ABL is not a straightforward task, and even though the precision of proposed algorithms is mediocre, I think that the article contains results that could be interesting to the scientific community and correspond well to the scope of the Atmospheric Measurement Techniques journal. An important positive point is that developed programs are available on the internet at the GitHub site under a fully open access option.

Printer-friendly version

Discussion paper

Nevertheless, there are a few issues on training, validation and testing, that I mentioned below in the General comments section. In my opinion, the quality of English should also be improved, and the paper would benefit from proofreading by a native speaker. Some theoretical explanations and expressions are often not accurate. There are also problems with text structure: some variables appear in the text before explanations in later paragraphs.

General (Major) Comments

1. When training the supervised ML algorithm, the estimation of the accuracy of ADABL on the validation ensemble (by the cross-validation technique) is presented in line 171: 99.5%. I think it is important to present also the accuracy on some testing ensemble, at least on the case study of April 19, 2017. It will justify the generalization ability of the applied AdaBoost algorithm showing the algorithm performance in an independent dataset. In my opinion, the training dataset could be insufficient.

2. Another reason why the accuracy of ADABL on validation ensemble is unrealistically high is the application of the random cross-validation split for time-correlated data (line 171). Using random selection from correlated datasets can lead to loss of generalization ability of the algorithm. The proposed ML method should be applied to new (independent) measurements. It means that AdaBoost should be trained-validated on uncorrelated parts of the dataset. If the data points were selected randomly from the whole dataset by the cross-validation procedure, it is highly probable that the similar neighbouring time points would be placed in both training and validation ensembles, which gives unrealistically good accuracy estimate on training-calibration datasets, but the worse result on another independent (test) dataset. I suggest the application of the block cross-validation.

3. If I understand right, the final configuration of the unsupervised ML algorithm (KABL) produces the classification using just one parameter - RCS0. In this case, the phrase in the conclusion (line 435) is misleading – “Both take the same input: one day of data

generated by raw2l1 routine; . . .”

4. Line 272: “number of invalid values (NaN or Inf) are recorded.” - Please explain why algorithms return these kinds of values. Another question is how algorithms deal with undefined values in Lidar measurements. Specific (Minor) comments

5. Line 15: “. . . boundary layer height (BLH). . .” – please give somewhere a definition of the BHL.

6. Line 105: “. . . we chose to derive boundary layer height with parcel method for the 11:15 sounding and bulk Richardson number for the 23:15 one.” Please justify why two different methods were used for morning and evening radiosounding.

7. Line 113: Does false positives on cloud detection perturb a BLH detection? Please explain.

8. Line 113: In the following text, some basic ML concepts are introduced for readers, who are not familiar with the scope of ML. In this case, the “false positives” should also be explained or referenced.

9. Line 127: As the number of seconds, since midnight is a periodical function, the ’classical’ distance could not take it in consideration correctly this variable. It means that the classical distance between one 00:01 and 23:59 will be nearly 24 hours. Please make sure, that ADABL algorithm works as expected in this case.

10. Line 142: I do not see any subsampling in figure 3. Is it a five-forks weak learner creation part? Please specify.

11. Line 142: How these shallow decision trees are fitted? I have never heard about resampling in the classical AdaBoost. Is it Bagging? Please give a reference or explain the algorithm in detail.

12. Line 143: “. . .the error of the classifier is the number of misclassified points.” - I am not sure that error is defined like that. Please explain or give a reference.

13. Line 146: The explanation is not sufficient. I propose to present here a reference to any popular textbook on AdaBoost or carefully introduce the algorithm. For example, in the expression, the performance was not introduced, the upper limit in the sum should be capitalized, etc...

14. Line 170: “trade-off between accuracy and computing time” - I do not think that the limiting factor for this problem is the computing time. Normally this kind of problem could be sufficiently well resolved by parallel computing.

15. Line 169: “RCSco, RCScr” – please make sure that these names for copolarized and crosspolarized range-corrected backscatter signals persist in the following text (notably in tables 2 and 3).

16. Line 175: “It is possible to quantify the relative importance of the predictors (Breiman et al., 1984; Hastie et al., 2009). After the training, the time accounts for 30.3%, RCSco for 28.4%, RCScr for 26.5% and the altitude for 14.8%.” – I have not found this information; could you please specify the corresponding page numbers?

17. Line 184: “distances from all points to all centroids” – Are these the Euclidian distances?

18. Line 196: “If we assume all Gaussian have the same fixed variance and that this variance tends to zero, EM and K-means algorithms are the same.” – Could you provide a reference or explain the statement?

19. Line 203: “Then the data they contain are normalized. . .” – if time and height are used in the KABL algorithm, are these variables also normalized?

20. Line 205: What values are included in predictors? If X matrix contains only signals RCSco and RCScr, it should be stated somewhere.

21. Line 209: “Finally, we look for the first change in clusters attribution, starting from the ground. This gives us the BLH for this profile.” I am not sure that this algorithm is optimal, as it could lead to oscillations of BLH. To understand how it could be improved

I suggest presenting and analyzing the altitude-time plot with pixels representing the results of the classification (like Fig 11 but with classes). Probably it is better not to take the first change, but a height above which the class is not changing, e.g. for three levels. Alternatively, a value of height could be selected that persist in time. These kinds of parameters could be optimally selected by the scores optimization. Another option is to modify the 'distance' definition.

22. Line 213: "The parameters of this computer code. . ." these parameters should be introduced at the beginning of the section 3.3, before they are referenced.

23. Line 298: ". . . figure 8 the distribution (violin plots) of the relevant output conditionally to the parameter value." – I suggest adding here a reference on the construction of this kind of plot.

24. Line 304: "Parameters values are chosen to give the most optimal value for the metrics they have influence on." - The selection of locally optimal combination of parameters does not provide the globally optimal solution. How can you be sure that this combination gives the best precision?

25. Line 316: "As the average gap E1 and the RMSE E2 are very similar. . ." – I suggest excluding the average gap E1 from the article for the sake of simplification.

26. Line 326: "Nighttime (launch of 23:15 UTC)" - If nighttime radiosounding was not used, why to present this dataset in "2.2 Radiosonde data"? Probably it was used in supervised ML? Please specify.

27. Figure 9: Adding the confidence intervals for RMSE and Correlation in Fig. 9 could be quite useful.

28. Line 335 and Line 438: I think it would be advantages to understand how works the lidar manufacturer's software and to give some interpretations.

29. Line 399: "A method to filter these oscillations will be needed, but it can also divest the "real-time" property." – Instead of filtering, the criteria of the lowest transition of the

class for KABL could be somehow modified, as I proposed in my comment for line 209. The filtering could be of the "real-time" if it is done relatively the past classifications.

30. Line 422: “5.6 KABL is "trainingless"” – I suggest that KABL could be used also by an expert to simplify the learning stage of supervised ML.

31. Line 417: “. . .strategies . . . for the training of ADABL. . .” – To decrease the sensitivity to “idealized” diurnal cycle of the BLH, I suggest trying to exclude the time predictor in ADABL.

Technical corrections

32. Line 13: “Atmospheric boundary layer concentrates many scientific challenges (small scale flows, turbulence...) and with high impacts due to its position of the interface between ground and atmosphere.” - awkward English, please correct.

33. Line 27: “(clouds, residual layers..)” -> “(clouds, residual layers. . .)”.

34. Line 88: “SIRTA” - Please decrypt the abbreviation.

35. Line 92: “at 11:15 AM and PM” – Please utilize the same notation for the time here and further. I suggest the UTC format.

36. Line 94: Please explain what the theta is. Is it the potential temperature?

37. Line 127: I suggest inserting a comma after “height above ground”.

38. Line 129: “[[1,N]]”- What does double brackets means? Please explain.

39. Line 145: “tree” -> “trees”.

40. Line 146: “m=200” -> “M=200”.

41. Line 154, 157, fig. 4: “top”->”left”, “bottom”->”right”.

42. Line 208: Init parameter is not defined.

43. Line 208: “specified in algo” -> “specified by algo parameter”

44. Line 270: “(0.20)” Is it the software version? Please specify.

45. Figure 10: Please introduce the INDUS abbreviation.

Interactive comment on Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2020-78, 2020.

Printer-friendly version

Discussion paper