Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2017-468-RC1, 2018

Atmospheric Measurement Techniques Discussions

Interactive comment on "A Machine Learning Approach to Aerosol Classification for Single Particle Mass Spectrometry" *by* Costa D. Christopoulos et al.

Anonymous Referee #1

Received and published: 15 March 2018

The authors apply a machine learning technique (random forests) which is known for its good predicting capabilities to a very interesting and unique SPMS data set. The goal is to predict the composition of an unknown artificial aerosol mixture. Whose constituents are known to belong to a group of aerosol classes that had been analyzed beforehand. The technique is a very promising approach and especially the information extracted from the training-algorithm seems to be very valuable.

But in my opinion several major issues need to be discussed:

-The main scientific work and ideas that were put into the paper are the growing of the forests and their validation including variable reduction and creation of the confusion

matrix. Accordingly, four out of five figures are about these topics. And the results are interesting and offer new ways of looking at this kind of datasets. In contrast the abstract, mainly the introduction and the conclusion strongly focus on the prediction of aerosol classes .

-The paper is in its present form hard to follow. Often the nomenclature is not consistent throughout the paper or doesn't fit to the cited literature.

For example, they do not use the proper term "random forest" but call it machine learning classifier, predictive model, classification model, rule-based probabilistic classification of a decision tree ensemble or supervised classification

and even more important

-While the basic algorithm to grow a random forest is presented. The underlying concepts (randomness, law of big numbers, assumptions, input parameter) and details of the validation process remain unclear.

For example, k=1000 trees have been used for each forest but no further explanation is given why exactly this number of trees is the right one. Or a plot of the test set error against number of trees presented which would make this decision obvious.

The number of random variables used to select the best split from is not specified nor its implications discussed.

The treatment of the "out-of-bag" observations, which is the central means of validation is not comprehensible.

The resultant classification accuracies are not put into perspective; thus the reader can't judge if the algorithm performs is a major improvement to other methods, of which the simplest would be to just use mean values of each aerosol class and use the most similar one as a prediction. It is not given which implementation of the algorithm is useed. Nor how long a typical random forest generation lasts and how this runtime scales with respect to number of particles, number of trees, number of split variables,

etc. . Along with the memory requirements which are missing too, these are basic and easy to provide information that help to compare this method to other methods.

-The random forests have been grown on chemical information and the size of individual aerosol particles, but some of the aerosol classes are not chemically defined. (e.g. multiple fertile soil classes, ATD) This basic contradiction is not clearly addressed.

-To me the section dealing with the blind test data does not fit to the abstract and introduction which present the random forest as a tool specifically suited for this use-case. After showing 80+

So my suggestion would be to split the paper and resubmit both parts in a thoroughly revised version one part with a clear focus on the algorithm and its general applicability to SPMS data including a real comparison to methods currently used (fuzzy-cmeans, manual decision tree, k-means).

And the other with a thorough analysis of the blind data set explaining in a comprehensible way the measured spectra based on all available information, statistics and assumptions. If it is not possible to explain the measured spectra in a controlled laboratory experiment like the one described, the use of the instrument to characterize atmospheric aerosol populations would be quite limited.

Interactive comment on Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2017-468, 2018.