Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2016-214-RC2, 2016

Atmospheric Measurement Techniques Discussions

Interactive comment on "Evaluation of Machine Learning Algorithms for Classification of Primary Biological Aerosol using a new UV-LIF spectrometer" *by* Simon Ruske et al.

Anonymous Referee #2

Received and published: 7 September 2016

This paper presents a number of different automatic clustering and machine learning techniques which are potentially useful for classification of bioaerosol observed by UV-LIF techniques. The authors compare the performance of these algorithms using populations of both mixed PSL spheres and mixtures of laboratory-generated biological particles. They also briefly discuss the effect of omitting some of the data (i.e. shape information) from the decision and find, seemingly, that the shape information provides a marginal improvement at best and serves as a hindrance in some instances.

Overall this paper can provide a useful contribution to the literature in that it applies a number of automated identification techniques to laboratory generated particles of known biological origin. The paper assumes quite a bit of prior knowledge of clustering

techniques and statistical treatment and I believe the utility of the work for non-experts could be improved if the authors put their results in context a bit more and expanded explanation in a few areas.

Specific comments include (in the order in which they appear in the manuscript):

1. The abstract refers multiple times to the "two datasets" but it is never stated, in the abstract, what those two datasets are. I would recommend adding that information (i.e. mixed PSLs and a mixture of laboratory generated bioaerosol) somewhere in the abstract. Also, it seems like it would be worthwhile to include something in the abstract about the findings regarding the utility of additional data like the shape information.

2. One p4, line 31, you say you have "used the z-score to standardize the data" but I don't really understand what that means. Does that mean that each channel in each dataset has been individually standardized using the z-score? Why does that make sense in populations of laboratory particles for which they need not be (and likely aren't) normally distributed? Also, was this only done for the HCA case or was this done for all of the data?

3. On pg 6, lines 28-30, this is not the most informative example as it is a value that has a much higher probability of being assigned to a particular category. What if the value was 0 instead? Would it then classify the particle as a blue particle because that is the most likely answer even though the particle is nearly as likely to be a red or a green?

4. Section 3, is there no reference specific to the MBS?

5. For Figure 5, please add error bars in the y-direction to show the variability of how a particular kind of well-known particle appears to the instrument. Also, it would be interesting to see size distributions, just as a metric for how the instrument does there.

6. For figure 6, so are you lumping all bacteria, fungi and pollens together for each of these traces? I would think there could be some important differences. For example,

between washed and unwashed bacterial populations, bacterial vs fungal spores as well as different pollen. Are the bacterial spores lumped in with the bacteria or the spores? This graph needs much more explanation and the laboratory methods used to generate it probably also deserve a bit more discussion. Also, again, vertical error bars and an accompanying size distribution would be informative.

7. Pg 13, line 12: The authors state that the division of data into training and testing sets has been described in section 2 but I can't find that text.

8. For table 3, what are the units of "Performance"? Is this the percentage of particles that are correctly identified? For the lab data is that identification as a broad class (bacteria, fungi and pollen) or is it on a sample-by-sample basis? It seems like that number is pretty dependent on the kinds of particles you used for your laboratory component. For example, you only have one sample of spores and they are a small fraction of the total data (by number) such that if they were getting lumped in with bacteria or pollen it wouldn't make the overall fraction of correctly identified particles that different. Similarly, do you classify all the non-fluorescent particles correctly (by number nearly half of your laboratory particles are non-fluorescent if I read your methods correctly) but then get the various classes of bioparticles confused with each other? It would be interesting if you could break down the correct identifications as a function of particle type. This would be complicated to show for all of the methods tested but, perhaps just for the top performers or something. This speaks to the potential of translating these methods to more diverse data sets.

9. Also for table 3, there are several things that don't make sense to me and which seem to be inconsistent with the text. First, for all of the HCA cases, it looks like the algorithm does worse for PSLs than for the lab data. Is this a type-o or are the columns mislabeled? Also, the text says that omitting the shape information leads to improved identification for PSLs but degrades identification for the laboratory data whereas that is not how I interpret the table unless the authors are only referring to the HCA cases. For 6 of the 10 supervised methods the reduced data leads to better identification of

laboratory data unless, again, columns are mislabeled.

10. The other thing that would be helpful in interpreting Table 3 would be if you could label the different methods according to the section headings from section 2.

Interactive comment on Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2016-214, 2016.