Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2019-495-AC1, 2020

Atmospheric Measurement Techniques Discussions

Interactive comment on "Classification of Lidar Measurements Using Supervised and Unsupervised Machine Learning Methods" *by* Ghazal Farhani et al.

Ghazal Farhani et al.

sica@uwo.ca

Received and published: 15 June 2020

1) I understand that for the supervised training around a few thousand scans were selected, and photon counts at altitudes were used as features. How many features did you have? (what is the dimension of the training set)

For the Rayleigh channels we have 2300 features as the lidar samples a broad range of altitudes, 25\km to 110 \km. For the nitrogen channel, the number of features decreases to 300 as the Raman scattering is weaker and above 25 \km altitude the signal to noise ratio drops. The number of features is sufficient in both cases, as they provide acceptable results, meaning using these features we can successfully train and

predict classifications for lidar scans.

2) I realize that t-SNE is a strong unsupervised method, but much slower than some other techniques such as KMeans clustering. Is that any reason that you have not implemented KMeans method?

The concept of KMeans clustering is based on finding spherical clusters which have a defined centroid for each cluster. Moreover, the number of clusters is set as an input parameter, and a wrong number of predefined clusters can result in unphysical results. KMean clustering is a reasonable way to proceed when there are principal physical arguments to hypothesize a specific number of categories/clusters. However, as we are performing exploratory data analysis (we might or might not encounter scans with traces of fire smoke), it is preferable to use a technique that "lets the data speak for itself", in which case t-SNE is a better choice of methods.

Interactive comment on Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2019-495, 2020.