Atmospheric
Measurement
Techniques

Open Access

EGU

Discussions

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

Anonymous Referee #2

Received and published: 9 September 2020

[journal abbreviation, manuscript]article

etoolbox mathptmx [11pt]moresize blindtext, xfrac

General comment
The study "Classification of Lidar Measurements Using Supervised and Unsupervised Machine Learning Methods" by Ghazal Farhani, Robert J. Sica, and Mark Joseph Daley deals with the important topic of automatic selection of "good" and "bad" data. The application is on Raman and Rayleigh lidar signals in cloudy, clear and aerosol-loaded conditions. The presented methodology is based on machine learning (ML) technique (both supervised and unsupervised) and allows to efficiently creating clusters of data

to facilitate the processing of the lidar data and to set up an alert system. The study presents first an overview of the most representative techniques using ML concept and then provides real examples of supervised and unsupervised techniques applied to the PCL signal. The topic is without any doubt of high interest, as it provides a solid solution to improve the representation and interpretation of lidar data.

The ML algorithm overview is comprehensive and well structured, as well as the part dedicated to the real examples from the PCL. However, starting from Section 3 the description and discussion of the tests results become somewhat confused, with an increasing number of typos and grammar mistakes that make the comprehension more difficult. Especially the discussion around the results shown in Figures 7-9 struggles to get to the point and loses in effectiveness to highlight the strengths of the t-SNE and DBSCAN methods. I recommend the authors to pay special attention to improve this part.

Despite these imperfections, I support the publication of this study as scientific article in AMT after revision.

Technical comments
Throughout the text

1. Why the authors use the term *scan* to indicate a lidar profile? I am used to refer to a scan strictly as for a scanning device, e.g., a scanning wind lidar will give a profile of wind strength and direction. The recording in time of the photon-counting signal when transformed into altitude is better called a profile.

2. When the authors use the relative pronoun "which", this shall come after a comma otherwise "that" has to be used instead.

3. Plenty of punctuation, articles and auxiliary verbs are missing through the text. I have tried to highlight a part of them, but the authors should perform a thorough re-reading of the manuscript and correct these typos and errors.

Sect.1, Pg 2, ln 20: the PCL acronym has already been defined.

Sect.2, Pg 3, ln. 8: you can remove "which are as follows".

Sect.2, Pg 3, ln. 23: "minimizes"

Figure 1, caption: I'd say the line color in the left panel is Cyan rather than blue.

Sect. 2.2, Pg 4, ln 5: "we have tested" (remove "been")

Sect. 2.4, pg6, ln 3: As it is mentioned here for the first time, it could be useful to add the extended name of bagging in brackets "bagging (bootstrap aggregating)" .

Sect. 2.5, pg 6, ln 28-29: the definition of overfitting should be provided in Sect. 2.4 when it is first introduced.

Sect. 3.1, pg 9, ln 4: "….different years AND represent different…"

Sect. 3.1, pg 9, ln 21: change "True negatives (FN)" to "False negative (FN)"

Sect. 3.1, pg 9, last paragraph: first you define precision and recall and then you present results of accuracy in Table 1. I would show Table 2 before Table 1, right below equations 9.

Sect. 3.1, pg 10, ln 5: "clear scans"

Sect. 3.3, pg 13, ln 9: replace "smoke does not present" with "smoke is not present".

Sec. 3.3., pg 13, ln 12: replace "during late June and early June 2002" with "during early June and late June 2002".

Sec. 3.3., pg 13, ln 13: "As, the smoke" without a comma

Sec. 3.3., pg 13, ln 16: "1961 lidar scans"

Sect. 3.3, pg 14, ln 2: replace "To investigate if scans with layers are clustered together the particle extinction …" with "To investigate if scans with layers are clustered together, the particle, extinction…".

Sect. 3.3, pg 14, ln 10-11: please consider rephrasing and correcting the English.

Sect. 3.3, pg 14, ln 11: "no anomalies WERE detected"

Sect. 4, pg 16: "We are planning to expand our unsupervised learning method to both Rayleigh and Nitrogen channels to be able to correctly identify and distinguish cirrus clouds from smoke traces in the UTLS.". Does the PCL have a depolarization channel, or is in the forthcoming plans to implement one to discriminate between different

particles based on their asphericity?