Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2016-7-RC2, 2016

Atmospheric Measurement Techniques Discussions

Interactive comment on "Retrieval of Intensive Aerosol Microphysical Parameters from Multiwavelength Raman/HSRL Lidar: Feasibility Study with Artificial Neural Networks" *by* M. Mustafa Mamun and Detlef Müller

Anonymous Referee #2

Received and published: 30 March 2016

A one-layer neural network (NN) is generated to estimate effective radius from multi-channel Raman lidar observations. The training and validation data are generated from a Mie code for various mono-modal log-normal size distributions. The best choice of input data is evaluated by considering the correlation of the network's outputs with the validation data, finding that the use of three backscatter and two extinction observations is usually optimal.

I reject this paper as I find the description of the methods excessively vague, consider the experiments poorly designed to answer the questions posed, and do not believe the

data justify the conclusions. The paper was particularly disappointing as Prof. Müller has previously produced exemplary research in this extremely interesting and relevant area.

The omission of uncertainty in the inputs, in my opinion, makes the results of Figs. 6–13 extremely unconvincing. This algorithm is presented with ideal inputs — noise-free simulations of perfectly spherical, mono-modially distributed particles — and yet barely produces results of equivalent quality to the studies cited in the conclusion. I struggle to see this algorithm producing useful results when exposed to the real world of noisy data and uneven particles. To be frank, if this data had been used to conclude that the NN technique is unsuitable for estimating effective radius from lidar data, I would have accepted it with minimal comment.

I must ask what is this paper trying to achieve? By evaluating only simulated data, it fails to show the algorithm is practically useful. By not analysing the validation data with any alternative algorithms, it fails to show the NN technique is any improvement on the status quo. The differences in performance shown are insufficient to establish that any one NN is superior. The omission of uncertainties precludes a sensitivity study. The title proclaims this is a feasibility study. The paper certainly shows that you *can* determine effective radius using a NN but it provides no convincing reason *why* or if this is done sufficiently well to be worthwhile. If the authors simply wanted to show it could be done, they could have written a one-page letter.

I liked that the introduction outlined the broad concept of the technique as this paper is the first application of NN to lidar and readers are unlikely to be familiar with it. Figure 2 is particularly well designed. However, the authors go too far in omitting any details of the algorithm used; simply listing MATLAB functions is insufficient. In particular,

- How, exactly, are the components of the refractive index determined? The paper often implies they are retrieved (such as at L307), but Fig. 3 indicates that they are inputs to the neural net. Fig. 4 implies there is some manner of iteration, but

the text constrains that iteration to the creation of the net.

- Further to that point, I am deeply concerned that the effective radius and refractive index are determined separately from the same input data. Why are they not both outputs of the algorithm? Even in the worst case considered, there are three inputs from which three outputs would be theoretically possible. NNs already present a "black box" and the vague manner in which the refractive index is determined presents the possibility that the same data is being used to determine multiple parameters in separate steps, which is a misuse of the available information.

- How many neural networks were used to produce the data presented — one or five? The text does not clarify if a single network was produced (for case A, taking 3β and 2α) and then operated using only some of the inputs or if five separate networks were trained for the cases A–E. The text implies the former, such that I wonder why the authors believe the network will work when presented with fewer inputs? Was it trained with various numbers of outputs? It seems entirely unsurprising that case A has superior performance if that's what the network was built to do.

  The extensive comparisons make the most sense if multiple networks were built, but then I would like to know how continuous the networks are? How do the result of the case A network compare to the case E network on the same simulation? Why would we expect different networks using different inputs to produce equivalent data?

- By your earlier description, the NN contains two components — the weights/biases applied to the inputs and the activation functions of the nodes. It seems as if the weights were determined during network training, but on L398 you say that the weights/biases are determined by trial-and-error. How important is the selection technique? How many such techniques exist?

As to the biases, L157 states they are fixed and random but L198 indicates they are determined within a numerical optimisation. Which is it?

- If each node has a weight and bias, why do you need to map the input data onto a normalised distribution (L219–224)? What are the standard deviations before mapping? Scaling to $[-1, 1]$ I can understand, but standard deviation scaling can be a non-linear transform of the data that I find concerning. Why was it necessary? "To achieve the best results" doesn't tell me what goes wrong if I don't.

- I agree with the referee #3 that there should be a discussion of why the NN technique is being used when others (by one of the authors and others) are available and could have been optimised for computational time.

  One suspects from L93 that it was used simply to see what would happen. While I've no particular problem with that, it would be good to hear a little more justification of why this particular NN was used over all those available. Why use only one layer? The original submission of this paper indicated that, in addition to evaluating the optimal number of measurements, there was a study of the optimal number of hidden neurons. Though there was too much detail then, it's complete removal provides too little.

- The mean squared error (MSE) is normalised by the degrees of freedom (DOF), being dependent on the number of measurements. However, the backscatter and extinction data have different units, different magnitudes, and very different uncertainties. The definition in Eq. (3) attempts to avoid some of these by normalising by the true value, but this still treats uncertainties in backscatter and extinction as equivalent, which is wrong. The denominator of Eq. (3) should be the uncertainty of the input.

- You often misuse the terms uncertainty and error. Error is the difference between your measurement and the 'true' value. Uncertainty is your estimate

of the likely magnitude of the error. Robust definitions can be found at http://www.iso.org/sites/JCGM/GUM-introduction.htm. Thus, Eq. (3) is an error because it is the difference between the estimated and modelled values whilst your study neglected uncertainties in the input data.

- Aspects of the comparisons are troubling. On L365, I have no idea what values of MSE would be high or low. Case A certainly performs better than cases B–E. However, all the comparisons confirm is that your NN did what it was designed to do (estimate effective radius from the outputs of a Mie calculation). It doesn't show that your algorithm is any good.

  On L367, I disagree that case A is superior. It has the highest Pearson coefficients, but it's results also show the most structure, with some manner of saturation apparent as the Angstrom exponent exceeds 0.5. This is not evident in cases B–D.

- Figs. 1 and 5 convey very little and seem drawn from a poster on this topic. Figs. 6–13 are appallingly bad. The font is too small, there are too many virtually identical plots for the length of discussion, and (considering the large number of data points indicated in §2.4) should probably be density histograms rather than scatter plots to illustrate the spread of data. At the moment, this algorithm looks fairly bad as the spread is large compared to my understanding of the typical uncertainties in existing techniques. The R^2 values indicate a dense packing of points near the one-to-one line that none of these plots makes evident.

Various more minor points and comments:

- Please refer to the cases of Tab. 2 uniformly by either letter or α/β sets; I prefer the brevity of the letters.

L204 Fig. 3 implies you used more than five input neurons. Please clarify.

L221 Possibly remind the reader that the min/max values are given in Fig. 3.

L230 to L233 doesn't actually explain why you picked the χ^2 test; you just state it's "good". Either write an actual explanation or omit these sentences (as the χ^2 test is fairly common).

§2.3 A mostly unnecessary summary of basic statistical concepts.

L266 The values in Tab. 2 indicate you consider virtually no coarse mode aerosol, which I'd define as at least a few microns. This seems a rather important omission.

L276 to L294. Why do you spend a whole paragraph telling us you shuffled the data? If shuffling is so important, you're going to need to do more than assure me that doing so avoids biases. If it's not, there's no need to mention it.

L306 This isn't a sentence and I don't know if you mean that the ANN was optimised for retrieval of effective radius or complex refractive index. The remainder of the document implies the former.

L309 Can an ANN account for a priori at all?

Fig. 6 This, and subsequent, are scatter plots not correlation plots.

§3.2 The loss of sensitivity with Angstrom exponent is interesting. Both it and lidar ratio are understood as measures of particle size, but the differing sensitivities implies otherwise.

L351 If you aren't going to show us any results from previous algorithms, you need to summarise their results and explain why yours are equivalent.

L354 Is it fair to compare analyses using three observations to four this way?

L382 You are the first to show such strong correlations because you were the first to look, not because of any relative skill in your technique.

L383 Could you explain the reasoning behind this supposition?

L422 These numbers should be one or two significant figures.

L669 Why are you mentioning work not done in a figure caption? That should be in the main text instead. Also, that flow chart implies you select the lowest MSE not by evaluating a number of circumstances but rather by considering it's absolute value. If this is the case, 'lowest' is an extremely poor choice of word.

The English has a peculiar grammar and is overly verbose but was no worse than I usually encounter from non-native speakers. I felt I understood the authors throughout the paper. The technical corrections I caught while reading follow:

L14 I think the following may be preferable: ANN could be a useful alternative or supplement to the traditional approach . . .

L17 allows for investigation of the information

L37 and at the ground are

L40 uncertainty in climate changes forecasts with regard

L43 They appear at various heights in the atmosphere

L54 data acquisition are far less

L64 amount of data, the addition of vertical

L70 reliable, quality-assured instruments and software for data analysis. Automated algorithms

L82 late 1990s to the mid 2000s

L96 results of a large-scale analysis

L97 during a 5-year effort

L98 used for support of data inversion

L99 of the human brain

L104 is unlike that in statistics

L107 area, in part because of the lack of appropriate

L122 among which the most

L186 I'm not certain what this sentence means. My best guess is, "This number of neurons was selected because it resulted in the minimal Mean Squared Error (MSE) in the training phase."

L200 the more difficult it becomes

L203 I'm not certain what the end of this sentence is getting at. My best guess is, "We find that five hidden neurons proves a reasonable compromise between the complexity of the NN, the effort required to train it, and the ability to enhance the network in future studies."

L216 I think 'potential' would be preferable to 'possible' as the former means "any rows that may exist in our case" while the latter means "any rows that could exist in any study", though neither word is actually necessary here.

L227 Shouldn't Pearson's r be italicised?

L229 at the end of the results

L239 lines and identify the slope

L245 The typesetting of this equation could be improved (for example, make the root symbol larger than the subsequent terms).

L263 Either ‘for which we need at least 3 hours computation time’ or ‘for which we need more than 3 hours computation time’.

L299 Either ‘a number of’ or ‘various’; no need to use both.

L361 data. However, extinction profiles

L370 Doesn’t this sentence repeat the previous?

L390 real part are more accurate than the imaginary part

L396 the accuracy of outputs depends on

L423 This result suggests that

Interactive comment on Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2016-7, 2016.