Coralum/sagitto next steps

From Jeremy Grata <grata@coraluma.com>
To Michael Pitsakis <mike@coraluma.com>
Date 2025-04-03 21:49

users 1, 2 and 3, you pam and Jake, each have at least 5 sessions (you may have 6)

using the 25hz 64pixel normalized PDS, i suggest the following for selecting outliers for removal from presentation to show expected performance with higher SNR.

1. for a given user, select 1 session to be the validation set. use the rest of the users data as the calibration set. produce a calibration, maybe gpr rq, linear least squares, or any other(s) that are interesting for pds. a standard regression learner 5fold approach is fine.

2. generate calibration self-prediction for all calibration data, ie selfpred=TM.predictfcn(cal). identify any predictions with abs error>20%. exclude these from cal set and recompute calibration(s).
3. recheck for outliers as in step 2.
4. make note of all outliers that were excluded from calibration.
5. compute validation predictions with the cal(s) from 2 (or 3 if additional OL were found, excluded and cal recomputed).
6. check validation predictions for an abs error >20% and make note of these.
7. execute 1-6 using the next session as val and the rest as cal.
8. repeat 7 until all sessions have been the validation set.
9. identify any spectra for which it was identified as an OL in step 2(or3) AND in step 6. this assumes step 8 has been completed.
10. Clarke and time plots can be generated for each session when it was the validation set using predictions from 6. this is the real validation performance. the same plots can be generated with OL identified in 9. omitted. this is the expected performance with higher SNR. in both sets of plots, avg by sitting is valid and may help overall.

additional iterations that may be useful:

1. the above defines a full session as a validation set. this is truly an independent set. but, it may be helpful to prevent overfit by defining val as 1st half of ses1 and 2nd half of ses2 (and so on).
2.this process can and should be run with 25hz DC, but the gpr exp as the calibration process.
3. there should be a step between 6and7 that looks at the scale of the time plot preds wrt the refs ... or the slope of the Clarke. if the slope is much less than 1 ... maybe any slope <0.9, the preds should be recomputed with the cal vector scaled by 1/slope. this would be true for self preds in steps 2and3 and in 6. however, aside from linear regression, the "predfcn math" would be needed to do this.
4. all cals from step 3 from the 10 steps, and step3 above should be stored for potential future use.

finally, there are probably other tweaks that could be thrown in along the way (a non single svm novelty cal is at the top of my thought list) such as a different PDS normalization approach, but I know we don't have the expertise and time to implement and test all of these. this is where guiding Sagitto and pushing them to try a little harder would help. I think a talk with them explaining were trying to maximize the presentation of our existing data to help with successful fund raising may help motivate them? But, it's 3pm Friday there now, so not sure we'll get a response until next week.

Jer

Sent from my Verizon, Samsung Galaxy smartphone