Response to reviewers

6/6/2020

We wish to thank three anonymous reviewers for their feedback on the second version of our paper. In this letter we respond to their comments and describe the changes made to the paper in response. Reviewer comments are in black and our responses in blue.

Reviewer 1

No comment to authors.

Thank you for reading the paper again.

Reviewer 3

This paper is very well done and the revisions have improved it even further.

Thank you for your earlier suggestions to improve the paper.

Reviewer 4

Authors mentioned that the WAPE are still less than one percent for every model/ensemble of models in Table 6. This is true. But it does not change the fact that the models are overfitting. If you evaluate the model performances based on other criteria, the in-sample and out-of-sample performances might be quite different. As such, these models might be biased.

> We recognize that the models might be biased, although the amount of bias is unknowable in this case since we do not know the true data generation process. See the following in the concluding remarks:
>
> > "There was some evidence that subsetting the sample can improve the results in some cases (e.g., when modelling the severity of crashes involving active travelers or motorcyclists), but possibly at the risk of overfitting the process. It is well known that overfitting can increase the accuracy of in-sample predictions at the expense of bias in out-of-sample predictions. Alas, since the true data generating process is unknowable in this empirical research, it is not possible to assess the extent of estimator bias."

I'm very confused by the calculation process of "percent correct". Not sure if it is a proper performance measure.

> Thank you for this comment. We suspect that the confusion might be caused by the differences in terminology. The terminology we are using is as follows:

Table 1: Example of a two-by-two confusion matrix

	Observed		
Predicted	Yes	No	Marginal Total
Yes	Hit	False Alarm	Predicted Yes
No	Miss	Correct Non-event	Predicted No
Marginal Total	Observed Yes	Observed No	

> An alternative terminology is used by He and Garcia in their 2009 paper "Learning from Imbalanced Data" (see Fig. 9 in their paper):

Table 2: Example of a two-by-two confusion matrix

	True	
Hypothesis	Yes	No
Yes	True Positive (TP)	False Negative (FN)
No	False Positive (FP)	True Negative (TN)
Marginal Total	P_c	N_c

> Percent correct in our paper is simply the accuracy of the forecast, in the terminology of He and Garcia (see formula 13 in their paper):

$$PC = \frac{\text{Hit} + \text{Correct Non-event}}{\text{Observed Yes} + \text{Observed No}} = Accuracy = \frac{\text{TP} + \text{TN}}{P_c + N_c}$$

As the authors mentioned, that crash data is highly imbalanced. In this case, there are other performance metrics which are more adapted into imbalanced data problem, such as ROC, G-mean, or F-measure.

> He and Garcia also reveal that the verification metrics we use report are all useful for imbalanced data. For instance, Precision is (formula 14 in He and Garcia):

$$Precision = \frac{TP}{TP + TF}$$

This is equivalent to the Post-Agreement, which we don't report, but that is a complement of False Alarm Ratio, which we do report:

$$PA = \frac{\text{Hit}}{\text{Hit} + \text{False Alarm}} = 1 - FAR$$

Recall in He and Garcia is simply the Probability of Detection:

$$Recall = \frac{TP}{TP + FN} = POD = \frac{\text{Hit}}{\text{Hit} + \text{Miss}}$$

As you can see, we do use performance metrics for imbalanced data. Other measures, such as ROC, G-mean, or F-measure, are combinations of the same metrics (Hits, False Alarm, Miss, Correct Non-event), similar to the Skill Scores that we report. Hopefully this will clear the matter.

You can refer to the following paper. Learning from class-imbalanced data: Review of methods and applications

https://www.sciencedirect.com/science/article/pii/S0957417416307175

Thank you for bringing this paper to our attention. We now cite it (as well as He and Garcia) in the concluding remarks.