The 99% accuracy club

Kajsa Møllersen, UiT The Arctic University of Norway

Melanoma Classification - a 10,000$ competition

For the 2020 Melanoma Classification competition[1] hosted by kaggle[2], 33,126 images were made available for training (of which 2% were melanomas), and an additional 10,982 were used for final ranking of the 3,308 teams who entered the competition with eyes on the 10,000$ prize. The task was simple: provide a probability for melanoma (deadly skin cancer) for each image. The ranking was based on the area under the ROC curve (AUC). Up until the deadline, contestants could submit their training results for an intermediate ranking.

And the prize goes to ...

A team of three kaggle grandmasters ran away with the first prize with an AUC of 0.9490. Their intermediate ranking was 881st - not even in the top 25%. The dynamics between intermediate and final ranking is easily explained by overfitting - the real enigma is how come computer scientists seemingly never learn.

How kaggle solves some problems - but also creates new ones

In a toy example[3], with an AUC of 0.9490 (same as the best team) produced by a linear + flat curve for each class, the DeLong[4] 95% confidence interval was 0.9368-0.9611.

A huge problem in the scientific community is the abuse of test sets, by modifying a method after test results, and then retesting and reporting the better result. 762 teams achieved a higher AUC than the upper limit of the confidence interval in the intermediate ranking. Assuming that the confidence interval includes the state-of-the-art AUC, 25% of the teams reported better performance when a data set was available for repeated testing. This enlightens a serious problem: when state-of-the-art is achieved by re-using the test set, it is impossible to beat it. Competitions like kaggle make sure that the test set is truly independent, and good performance is not a result of fitting the method to the test.

At the lower end, the confidence interval includes the AUC of 336th ranked team. In other words, if we had a number of test sets drawn from the same distribution, the AUC would vary each time, and the variation is so large that it could completely re-rank the top 10% of the teams.

Is it possible that also 0.9490 is an overestimation of the state-of-the-art? If we assume that the 10% best performing teams in reality each have reached the upper threshold of what a method can perform, and all variation is due to some random nature of the methods, then half of them will overestimate the AUC. By this reasoning, the true state-of-the-art AUC is 0.9401, the AUC of the 166th ranked team.

Competitions like kaggle solves the problem of test-set abuse, but creates a new one similar to multiple testing, and can be seen as an example of regression to the mean.

---

[1] International Skin Imaging Collaboration. SIIM-ISIC 2020 Challenge Dataset. International Skin Imaging Collaboration https://doi.org/10.34970/2020-ds01 (2020)

[2] https://www.kaggle.com/c/siim-isic-melanoma-classification/overview

[3] https://github.com/kajsam/NOBIM2021

[4] Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Non-parametric Approach, Elizabeth R. DeLong, David M. DeLong et al., Biometrics, 44, 3, 9 1988