REPORT on
the article entitled How doppelgänger effects in biomedical data confound machine learning

ZHOU Yihang
Macau University of Science and Technology

An applicant to the program of Msc of Biomedical Data Science at Nanyang Technological University
Application number: C2213722

1. INTRODUCTION

In this article, Wang *et al* [1] proposed that doppelgänger has an inflationary effect on the research results of biomedical data in machine learning. Doppelgänger's definition is that in machine learning, a pair of data that should be obtained independently are very similar, even if they come from different individuals. If researchers randomly divide a large amount of data collected into the training set and validation set, this doppelgänger data may be allocated to the training set and validation set respectively. This will make the trained models perform well, no matter how they are trained. For an extreme example, if the training set and validation sets are full of a large number of doppelgängers, and the data that are not similar to these doppelgänger data account for a relatively small number, when the existence scale reaches a certain degree, the validation set is almost equal to the training set, and then using the validation set to evaluate the performance of the model is equivalent to using the data of the training set to verify the model. This evaluation of the model will undoubtedly lead to the conclusion that the model performs very well. However, when these doppelgänger data are reduced or removed, or when the model is applied to practical applications, the model may perform otherwise very poorly.

Unfortunately, this phenomenon exists not only in biomedical data science but also in machine learning in other fields. This will greatly affect people's cognition of the objective accuracy of the model. The authors believed that this phenomenon should be eliminated, and the industry should also have objective standards to eliminate the impact of doppelgänger. However, at present, few people have noticed or taken practical actions to curb this phenomenon, and the methods existing are not perfect. Aiming at solving this problem, the authors refer to some studies done by others [2,3] and point out the advantages and limitations of their methods. At the same time, the authors also used the algorithm in Waldron's article [2] to carry out further research, which proves that the existence of doppelgänger does have an impact on the evaluation of the model, and finally puts forward three solutions.

2. EXPERIMENTS

In the research of Waldron [2], pairwise Pearson's correlation coefficient (PPCC) was proposed to find the doppelgängers data in the training set and validation set. However, Wang *et al* believed that it never conclusively made a link between PPCC data doppelgängers and their ability to confound ML tasks, and there is also a leakage phenomenon in the data used in this article, so the authors believe that the conclusion of this article cannot prove that doppelgängers will expand the evaluation model. However, Wang *et al* believed that the PPCC proposed in this article is very worthy of reference, so the method of finding doppelgängers data was applied to renal cell carcinoma (RCC) proteins data. RCC was chosen for its utility in constructing clear-cut scenarios: (i) negative cases, in which doppelgängers are nonpermissible by constructing samples pairs of different class labels; (ii) valid cases, in which doppelgängers are permissible by constructing sample pairs assigned to the same class label but from different samples. These effects can then be compared against positive cases (pairs constructed by taking technical replicates arising from the same sample; these constitute obvious leakage issues and, therefore, are not considered doppelgängers). Wang *et al* conducted experiments when doppelgänger data exists in data set, randomly select features and train the data with KNN, Naive Bayes, Decision Tree, and Logistic Regression models. The final experimental results showed that in the presence of doppelgängers, even the model of a meaningless random selection of features has a good training effect, and the more doppelgängers exist, the better the model effect. In the control experiment, when doppelgängers were all put into the training set for training, the performance of the model was close to 0.5, which is in line with the model results of randomly selected features. KNN, Naive Bayes, Decision Tree, and Logistic Regression are affected differently by doppelgängers. By comparing with the results of the control experiment, it can be obtained that doppelgängers will expand the model, and the more doppelgängers account for, the closer the training result is to leakage. Through this experiment, Wang *et al* successfully proved that doppelgängers will inflate the model.

3. METHOD

About how to solve doppelgängers, Wang *et al* summarized three points according to the experience of others and their experiments.

First, all doppelgängers data can be put into the training set or validation set as done in the experiment to reduce the impact of doppelgängers and thus ensure that the number of data is not affected. However, this method also has its limitations. When all doppelgängers are put into the training set, if the size of the training set is fixed, each newly included data doppelganger will cause a less similar sample to be excluded from the training set, this may lead to the model not being well generalized because of the lack of knowledge. When all doppelgängers are put into the validation set, it may end with a winner taking all scenario (doppelgängers' prediction is either correct or wrong).

Second, the data can be filtered in a more rigorous way like Cao and Fullwood [4]. This can be achieved by segmenting training and validation data according to a single chromosome (rather than all chromosomes considered together), and generating training evaluation pairs using different cell types, and then establish a good practice/standard in this field. However, this is difficult to do in practice because it depends on a priori knowledge and the existence of high-quality context/benchmark data.

Third, doppelgängers data can be removed overall, so that doppelgängers will not exist in the training set and validation set. Finally, the trained model will not be affected by doppelgängers. However, the limitation is that if the original data set is not large and there are a large number of doppelgängers, deleting all doppelgängers will sharply reduce the size of the data set, and may even reach the point that the data set is too small to use.

Although these three methods can solve doppelgängers, they all have certain limitations to use and cannot solve doppelgängers without a large number of changes in the data set and less human intervention.

4. RECOMMENDATIONS

Finally, given the limitations of the above methods, Wang *et al* gave three recommendations to solve the doppelgängers.

The first recommendation is the data processing. The data should be carefully cross-checked under the guidance of metadata. Just like the RCC data set used by the authors in the experiments, because there are clear labels of positive cases, valid cases, and negative cases, it is possible to predict the PPCC score range (different classes and negative cases) and whether there is leakage (the same patient, the same grade according to repeated cases and positive cases) in the experiments in the absence of doppelgängers. With this information from metadata, we can identify potential doppelgängers and thus classify them into training sets or validation sets, effectively preventing the doppelgängers effect and making the evaluation of the model more objective and reliable.

The second recommendation is that data stratification should be implemented. We can layer the data according to different similarities, rather than evaluate the model performance according to the whole test data. Through layering, we can know which data is not so similar to other data and the quantity is small. With this information, we can supplement such data to improve the reliability of the model.

The third recommendation is the validation. Extremely strict independent validation check should be performed, involving as many data sets as possible. Although this cannot directly eliminate the influence of doppelgängers, it can at least point out the reliability of the machine learning model more objectively.

5. MY THINKING

I think that to solve the doppelgängers problem, there is a way to find the best objective performance model in that the current data set can achieve without human intervention and reducing the size of the data set. Firstly, we need to find all doppelgängers in the data set through PPCC, and then put a specific number of doppelgängers data into the validation set to make the proportion of doppelgängers in the validation set reach a specific proportion (e.g. 60% or 70%), generate multiple training sets and data set pairs through different proportions, and then train them separately to obtain multiple models. Finally, these models are verified with the validation set with doppelgängers accounting for 50% (ensuring that there is no training set data in the validation set). In this case, the best model found is the best model that can be found without human intervention in the data set and reducing the size of the data set.

After reading this article, it reminds me of my experience when I wrote about classifying ginseng and American ginseng with yolov5. Now I think of the impact of doppelgängers on my model. Since there is no ready-made ginseng image data set, my training data is the images downloaded by the web crawler. After the initial data processing, I began to train the training set. When the data set is small, the precision I trained is 0.956 (Fig. 1.). When I wanted to further improve the accuracy of my model, I thought of refining the data set. Therefore, after deleting some duplicate, similar and unclear data in the data set, I trained the model again. However, the accuracy of this model is 0.917 (Fig. 2.), which is worse than the performance before refinement. At that time, I still wondered why the data set was better, but I trained a worse model. Now I understand that it may be the existence of doppelgängers, which inflated the model I trained for the first time. After accidentally eliminating doppelgängers in some validation sets, the performance of my model becomes worse. At that time, I was not aware of the existence of doppelgängers and the expansion effect of doppelgängers on the model, so I restored the data set. Now, I think I should find out the objective optimal model of the data set at that time according to the above method and try to reduce the impact of doppelgängers on my model. Although I cannot improve my model and show the improved results in this report due to the limitation of time and computing resources, I look forward to refine my model after studying deep learning for biomedical science and other courses in Nanyang Technological University in the future.

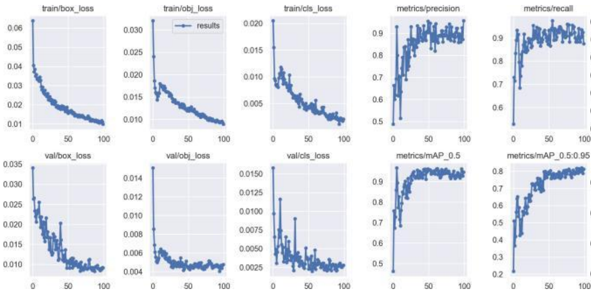

Fig. 1. First Train Result

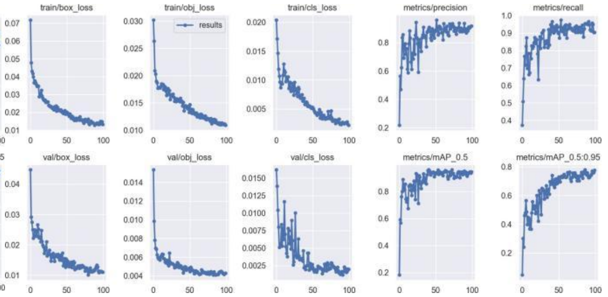

Fig. 2. Second Train Result

6. SUMMARY

After reading this article, I realized that it is because of the existence of doppelgänger, many models that performed well after training did not perform well in practical application. At the same time, I also learned about the existing solutions, and put forward my views on solving doppelgänger. Reading this article is of great benefit to my future learning career and research career, which can make me aware of the shortcomings of trained models and also provide me with many good ways to solve it.

7. REFERENCES

[1] Wang LR, Wong L, Goh WWB. How doppelgänger effects in biomedical data confound machine learning. *Drug Discovery Today*, 2021; https://doi.org/10.1016/j.drudis.2021.10.017

[2] Waldron L, Riester M, Ramos M, Parmigiani G, and Birrer M. The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles. *JNCI: Journal of the National Cancer Institute*, 2016;108(11): djw146.

[3] Sheng Q, Shyr Y, and Chen X. Dupchecker: a bioconductor package for checking high-throughput genomic data redundancy in meta-analysis. *BMC bioinformatics*, 2014; 15(1): 323

[4] Cao F and Fullwood MJ. Inflated performance measures in enhancer-promoter interaction-prediction methods. *Nature genetics*, 2019; 51(8): 1196–1198