Report on doppelganger effects

Haogang Huang

Introduction

Over the years, as the field of machine learning has boomed, it has been applied to a growing number of industries to provide efficient solutions. In the biomedical field, machine learning is widely used in direct-to-consumer (DTC) medical AI/ML apps[1], genetics and transcriptomics, diagnostic imaging, and prediction the risk and progression of many diseases through electronic health records (EHRs) [2]. However, according to a potential problem that affects the accuracy and validity of machine learning models is known as the doppelganger effect, a problem that is widespread but has received little attention. In this report we will discuss the existence, quantification, and uniqueness of the doppelganger effect, and propose possible methods of avoiding the doppelganger effect based on my current knowledge.

The existence of doppelganger effects

In machine learning, it is widely accepted that the training and test data sets should be generated independently for evaluating the performance of a ML model. However, results from validation might still be inaccurate when training and test sets were independently generated because of the similarity between the test and training datasets. Doppelgänger effects happen when samples have unintentional similarities that boost the performance of a trained machine learning model when divided across the training and validation sets. This inflationary impact leads to false trust in the model's deployability. [3] This means that the doppelganger effect gives us a machine learning model that overestimates validity and does not produce data that is as accurate or valid as expected in the real world. They could skew analytical procedures that favor choosing feature sets with high validation accuracy.

Doppelganger effects may be found in a lot of different biological data. Doppelganger effects also exist in areas such as imaging, gene sequencing, and metabonomics. For example, in a study by Cao and Fullwood on chromosome interaction prediction systems, they found that the performance of machine learning models was inflated because they were evaluated on a test set with a high degree of similarity to the training set, i.e. doppelganger data was present in the test set. Goh and Wong also identified doppelganger effects in their genetic data study, certain validation data were guaranteed a good performance given a particular training data, even if the selected features were random. Besides, the doppelganger effect is also present in the QSAR dataset and in the renal cell carcinoma proteome data obtained in NetProt.

Doppelganger effects are not unique

In my own opinion, based on my existing knowledge of machine learning, doppelganger effects will

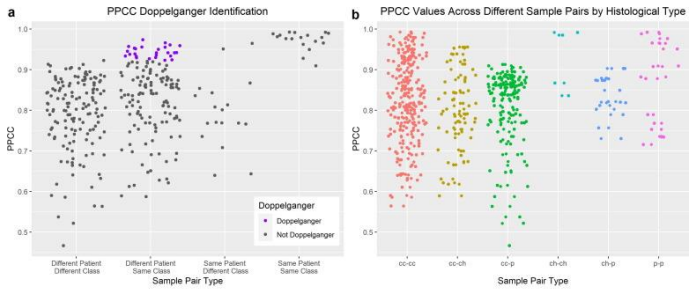
occur not only in the field of biomedical data, but also in other areas of industry. In practice, there will always be cases where the test set is similar to the training set, thus exaggerating the utility of the model in the real world. I think there is a link between problems like doppelganger effects and data leakage and overfitting. These are very common problems in ML, and when these problems occur, doppelganger effects are likely to occur.
In today's popular social platforms, recommendation systems are crucial. However, in practice, we find that although most of the time the platform is able to push us the exact content we are interested in, sometimes the system will suggest content that is not relevant to our interests. I think the doppelganger effect is one of the reasons for this phenomenon, as social platforms have a huge number of users and many people have similar interests, so the platform will push you other content from people with the same interests that you may not be interested in.

The quantification of doppelganger effects

There are a number of ways to analyse the generation of the doppelganger effect from a quantitative perspective. The first is the dupChecker method, but it is considered not to picked up actual data doppelgängers that are separately obtained samples that are similar by coincidence.

What we would most like to discuss is the pairwise Pearson's correlation coefficient (PPCC), captures relationships between sample pairs of various data sets. A pair of samples that together have an abnormally high PPCC value are known as PPCC data doppelgängers.Sample pairs with extremely high mutual correlations or resemblance are known as data doppelgängers (DDs). To find DDs, for instance, we may use the pairwise Pearson's correlation coefficient (PPCC), where sample pairings with high PPCCs are also known as PPCC DDs. However, functional doppelgängers (FDs) are sample pairs that do not actually "learn" when divided between training and validation data.

To validate the PPCC in a realistic scenario, Wang et al. used the renal cell carcinoma (RCC) proteomics data from Guo et al. for sub-scenario validation, ultimately calculating the PPCC with the largest negative sample pair. They identified PPCC data doppelgängers based on the PPCC distribution of the valid scenario against the negative and positive scenarios. And they observed a high proportion of PPCC data doppelgängers.



In summary, the similarity of DDs and their proportion in the validation set are two key quantitative criteria. the higher the PPCC coefficient, and the higher the proportion of PPCC DDs in the validation set, the more pronounced the DE and the more pronounced the exaggeration of ML model performance.

Besides, other correlation metrics such as the Spearman Rank correlation coefficient and Kendall Rank correlation coefficient can also be used to identify DDs.

How to avoid doppelganger effects[4]

Cross-checks

The first approach is to conduct thorough cross-checks using the meta-data as a reference. Here, we created negative and positive examples using the RCC meta-data. This allowed us to predict the PPCC score ranges for instances when leakage is present and doppelgängers are not possible. Samples from the same class but distinct patients are the conceivable data twins that call for concern. In order to effectively prevent doppelgänger effects and enable a more objective assessment of ML performance, we are able to use the meta-data to identify probable doppelgängers and group them all into either training or validation sets.

Data stratification

Data stratification is the second technique. We can divide test data into strata of varying similarity (for example, PPCC data doppelgängers and non-PPCC data doppelgängers), and then evaluate model performance on each stratum separately, rather of evaluating model performance on the entire test data. Assuming that each stratum corresponds with a known percentage of the real-world population, we may still evaluate the classifier's performance in the real world by taking into account a stratum's real-world prevalence when evaluating the performance at that stratum. However, strata with poor model performance also reveal weaknesses in the classifier.

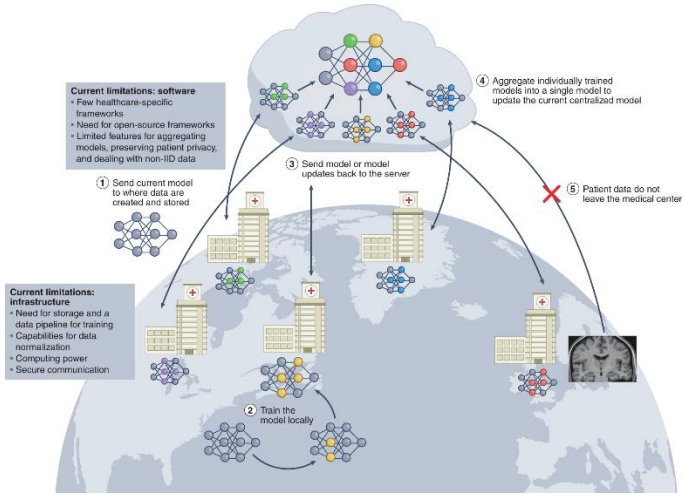
Augmenting datasets

Large, diversified, balanced, and well-labeled datasets have been a prerequisite for many successful models during the last ten years. The difficulty of obtaining large quantities of standardized clinical data is made worse by the fact that these data frequently represent the patient populations of a single or small number of institutions.

Strong answers to these issues are provided by GANs. To enhance model performance, training data may be supplemented using GANs. A convolutional neural network (CNN) for the categorization of liver lesions, for instance, improved the performance of the model by 10% compared to a CNN trained purely on classically supplemented datasets. [5]

Federated learning

In order to maximise the amount and diversity of data, we can use federated learning. Federated learning is a paradigm for training ML models when decentralized data are used collaboratively under the orchestration of a central server. In contrast to centralized training, where data from various locations are moved to a single server to train the model, federated learning allows for the data to remain in place. This mitigates concerns about privacy breaches, minimizes costs associated with data aggregation, and allows training datasets to quickly scale in size and diversity. The successful implementation of federated learning could transform how deep-learning models for healthcare are trained. [5]



Reference

[1] Babic, B., Gerke, S., Evgeniou, T. *et al.* Direct-to-consumer medical machine learning and artificial intelligence applications. *Nat Mach Intell* **3**, 283–287 (2021). https://doi.org/10.1038/s42256-021-00331-0

[2] Rajkomar, A., Oren, E., Chen, K. et al. Scalable and accurate deep learning with electronic health records. npj Digital Med 1, 18 (2018). https://doi.org/10.1038/s41746-018-0029-1

[3] L.R. Wang, X.Y. Choy, W.W.B. Goh, Doppelgänger spotting in biomedical gene expression data iScience, 25 (2022), p. 104788https://doi.org/10.1016/j.isci.2022.104788

[4] Zhang, A., Xing, L., Zou, J. et al. Shifting machine learning for healthcare from development to deployment and from models to data. *Nat. Biomed. Eng* 6, 1330–1345 (2022). https://doi.org/10.1038/s41551-022-00898-y

[5] Frid-Adar, M. et al. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing* 321, 321–331 (2018). https://doi.org/10.1016/j.neucom.2018.09.013