Doppelganger effects in machine learning model for biomedical science

Xiao Xia (C2308025)

Introduction

Machine learning models are increasingly being used to discover and accelerate drug development. Classification models based on machine learning and artificial intelligence can predict the interactions between new drugs and diseases and identify drug candidates after training is completed (Shi et al., 2018). This greatly reduces the time and money required for drug testing during drug development, making machine learning models a promising area of research in biomedical science.

In a machine learning model, datasets for training and validation should be derived independently. However, due to chance or many other reasons, the datasets used for training and verification may have high similarity, which leads to the inflated accuracy of the evaluated data sets used for verification regardless of the training effect, resulting in poor model construction effect, which is called doppelganger effects (Ho et al., 2020).

In this report, the following issues are discussed to gain a deeper understanding of what the doppelganger effect is, how it emerges from a quantitative angle, and how to avoid it in research in machine learning in the biomedical field.

What is the doppelganger effect and how does it manifest outside the biomedical field?

Doppelganger effects, also known as "data clones," refer to the phenomenon of multiple observations in a dataset having identical or highly similar characteristics. This can make it difficult for machine learning models to accurately classify, predict or generalize new data. It is not unique to biomedical data. In image data, doppelganger effects may occur when multiple images are nearly identical, such as in the case of stock photos or images of popular landmarks. This can make it difficult for image recognition algorithms to accurately identify and classify images. In genomics data, doppelganger effects may occur when multiple individuals have similar genetic profiles. This can make it difficult to accurately identify genetic variations that are associated with specific diseases or traits.

How do doppelganger effects emerge in a quantitative way?

Doppelganger effects can occur in any type of data and can be caused by data duplication, data imbalance, poor data quality, insufficient data, or high correlation between observations (Abdur & Alspector, 2003). Data duplication means that the same data is recorded many times in the dataset because of the overlap of data sources or collection errors. Data imbalance means excessive recording of data in the same category or group so that they are over-represented in datasets. Poor data quality and insufficient data mean that there is insufficient or high-quality data, which leads to the over-fitting or difficult classification of the model in the training process. High correlation means that multiple data in a dataset are highly similar due to how they are fetched or for other reasons. For example, datasets composed of data produced in the same laboratory are likely to be highly correlated.

How to avoid doppelganger effects in research in machine learning in the biomedical field?

In my opinion, doppelganger effects can be avoided or mitigated in three aspects: data acquisition, data preprocessing and machine learning model construction.

When obtaining data, you should avoid obtaining data from a single data source, but from multiple sources, such as multiple hospitals or multiple laboratories. The diversity of data needs to be emphasized. Biomedical data of different genders, countries and ethnicities can be selected to avoid doppelganger effects. Data augmentation can be used when the amount of data is insufficient or when there is a single source of access. New and unique data can be created by adding noise to the data or rotating the image data to increase the diversity of the dataset.

Datasets should be preprocessed before being used for the training and validation of machine learning models. First, duplicate or highly similar data should be removed. After that, you can use the metadata as a guide for cross-checking. Metadata can be used to construct negative and positive cases to confirm the scoring range of the model. Using this information from the metadata, we can identify potential data doppelgangers and place them all in a training or validation set, effectively preventing doppelganger effects (Wang et al., 2021).

During model training, the data can be stratified. The data is divided into different layers with different similarities, and the model trained by each layer of data is verified during verification. If the population proportion in a certain data layer is consistent with the overall data, the performance of the model trained by the overall data can still be evaluated by the performance of the model trained by the data of this layer. If the verified performance of a certain layer of data is poor, this layer can be used to point out gaps in the model. (Venet et al., 2011).

Conclusions

Doppelganger effects are a common phenomenon in datasets, which may affect the accuracy of machine learning models. In the biomedical field, machine learning has a wide range of applications. However, Doppelganger effects cause problems in biomedical data machine learning due to data redundancy, imbalance, and high correlation. There is currently no effective way to completely avoid Doppelganger effects. Doppelganger effects can be avoided or mitigated by some strategies in data acquisition, preprocessing and model construction steps.

Reference

J.-Y. Shi, X.-Q. Shang, K. Gao, S.-W. Zhang, S.-M. Yiu, An integrated local classification model of predicting drug-drug interactions via Dempster-Shafer theory of evidence, Sci Rep 8 (2018) 1–11.

S.Y. Ho, K. Phua, L. Wong, W.W.B. Goh, Extensions of the external validation for checking learned model interpretability and generalizability, Patterns 1 (2020) 100129.

Chowdhury, Abdur, and Joshua Alspector. "Data duplication: an imbalance problem?." ICML’2003 workshop on learning from imbalanced data sets (II), Washington, DC. 2003.

Wang, Li Rong, Limsoon Wong, and Wilson Wen Bin Goh. "How doppelgänger effects in biomedical data confound machine learning." Drug Discovery Today (2021).

D. Venet, J.E. Dumont, V. Detours, Most random gene expression signatures are significantly associated with breast cancer outcome, PLoS Comput Biol 7 (2011) e1002240.