The Influence of Doppelganger Effects on Biomedical Data in Machine Learning

Wenjun Gu

Introduction

With the development of technology, data plays an important role in every field. With the growth of data, how to extract and analyze large amounts of data effectively has become one of the popular research problems. Machine learning (ML), as a subset of artificial intelligence, is increasingly applied to data analysis within biomedical field (Uçar et al., 2020). It can analyze and examine large amounts of data automatically and make predictions by building and training data models (Uçar et al., 2020). For example, during the process of drug discovery, trained machine learning models (classifiers) can improve the efficiency of drug development by predicting interactions between new drugs and diseases, and possible side-effects (Wang et al., 2021). However, the presence of confounding effects in data may be affect the reliability of machine learning models, when high similarity is found within independently derived data, the performance of ML models will be inflated, in other words, they will perform better than its actual situation regardless of how they will be trained, this is known as the doppelganger effect (Wang et al., 2021). In this essay, the doppelganger effect in all different kinds of data and specific biomedical data examples will be given, and potential ways of avoiding and examining doppelganger effect on biomedical data will be analyzed.

Different types of data doppelganger

It is clear that doppelganger effects are not unique in biomedical data, in fact, doppelganger effect is common in big data in our daily life. Many internet companies using data doppelganger to improve their offerings, for example, when the online shopping platforms like Amazon or social media like Facebook or Instagram will recommend similar products or contents to users who has similar interest based on results of analyzing users' data like searching history (Sirvaiya, 2021). Also, doppelganger effect is common in other types of data like climate,

economic and financial data, when meteorologists do weather forecast or economist do prediction and estimation of future scenarios of economic and stock market using ML models, data doppelganger may influence the accurate of predicting (Gogas & Papadimitriou, 2021).

Doppelganger effect in biomedical-related data

Doppelganger effect is common in biomedical data, next, three examples of different types of biomedical data will be analyzed to show the effect of data doppelganger to ML models.

The first example is medical diagnosis in precision medicine, in simple terms, the data of patients' medical record of known correct diagnosis are used to run a learning algorithm to build and train a ML model, the model can derive medical diagnostic knowledge automatically from solved cases in the past, and deal with missing and noisy data due to the uncertainty, errors and lack of medical data to improve the diagnostic speed and accuracy (Kononenko, 2001). It is clear that many similar data are existed when build and train diagnosis ML models since different diseases may have similar symptoms and features, and it may cause doppelganger effect. The ML models often have generalizability but not all diseases have. Thus, patients' data doppelganger may influence the accurate and speed of ML diagnosis models.

The second example is about genomics and bioinformatics, machine learning often used in genomics to recognize patterns in DNA sequences, assign functional annotations, and help to understand the mechanisms underlying gene expression (Waldron et al., 2016). For example, when scientists do cancer genome analysis and try to find specific target DNA sequences for diagnosis and treatment, they often use public data from cancer atlas or database like NCBI. Due to privacy problems, publicly available data for human cancer genome are limited, this may cause doppelganger effect since cancer transcriptomes are always highly distinctive and it is difficult to identify uniquely (Waldron et al., 2016). It is unavoidable that limited data re-use in different research. When ML models is built and trained based on these hidden duplicates data, genomic models may be inflated and influence statistical significance or accuracy

(Waldron et al., 2016).

The third example is about medical imaging. In recent years, more and more research show machine learning can be used as a technique to recognize patterns of medical imaging, the algorithm system identifies combinations of medical image features for classify and computing some metric for specific regions to make prediction and diagnosis (Erickson et al., 2017). For example, ML models can be used to analysis functional MR imaging of brain to detect the cognitive state and diagnose potential neurologic disease like Alzheimer disease (Erickson et al., 2017). Due to the complexity of medical image and clinical situation, datasets bias will occur since available dataset can only reflect some medical condition partially, when the similar training data has different distribution than the actual test data, doppelganger effect will be caused and the accuracy of diagnosis will also be affected (Varoquaux & Cheplygina, 2022).

Methods of avoiding and checking potential doppelganger effect

Some methods will be given to check or avoid potential doppelganger effects when build and train ML models on biomedical data.

First, meta-data can be used as a guide to perform cross-checking since we can use meta-data to identify potential data doppelganger and differentiate them to training or validation sets, it can help us prevent doppelganger effects of ML models and get more objective evaluation of its performance (Wang et al., 2021). For example, from aspects of bioinformatics, when we do DNA or RNA sequence analysis, meta-data like which organism or cell line this sequence comes from, the sequencing platform and data processing information could give us more background information about genomic data and help us identify potential data doppelganger and performance of ML models.

In addition, data stratification is another way to check potential doppelganger effect. Evaluating stratified data based on different similarities is more efficiency than evaluating the whole datasets in ML models (Wang et al., 2021). Also, strata with poor model performance

can be found and improved quickly to reduce the impact of doppelganger effects on ML models (Wang et al., 2021). Also, under condition of permissible, multiple data sets from different sources can be used to increase statistical power and reduce uncertainty of biomedical data (Wang et al., 2022). Moreover, software tools can be used to identify data doppelganger when previous methods are not working due to some limitations. However, no effective software tools can be used to identify doppelganger effect for a long time (Wang et al., 2022). Recently, the new R package *doppelgangerIdentifier* were developed to identify pairwise Pearson's correlation coefficient (PPCC) and data doppelgangers (DDs) between microarray and RNA-Seq data set (Wang et al., 2022).

Conclusion

As one of the common issues when building and training machine learning models on biomedical data, doppelganger effects may inflate the performance of ML models, and show low accuracy on medical diagnosis, genomic analysis, and medical imaging analysis. Although methods like performing cross-checking by meta-data and data stratification can check and avoid doppelganger effects on biomedical data, it is necessary for humans to develop more effective ways like software tools to reduce the influence brought by data doppelganger.

Reference

Erickson, B. J., Korfiatis, P., Akkus, Z., & Kline, T. L. (2017). Machine Learning for Medical Imaging. *RadioGraphics*, *37*(2), 505–515. https://doi.org/10.1148/rg.2017160130

Gogas, P., & Papadimitriou, T. (2021). Machine Learning in Economics and Finance. *Computational Economics*, *57*(1), 1–4. https://doi.org/10.1007/s10614-021-10094-w

Kononenko, I. (2001). Machine learning for medical diagnosis: History, state of the art and perspective. *Artificial Intelligence in Medicine*, *23*(1), 89–109. https://doi.org/10.1016/s0933-3657(01)00077-x

Sirvaiya, S. (2021, August 23). *Doppelganger: Your data has a twin?* Medium. Retrieved December 18, 2022, from https://medium.com/analytics-vidhya/doppelganger-your-data-has-a-twin-f52fe53c4ce

Uçar, M. K., Nour, M., Sindi, H., & Polat, K. (2020). The effect of training and testing process on machine learning in biomedical datasets. *Mathematical Problems in Engineering*, *2020*, 1–17. https://doi.org/10.1155/2020/2836236

Varoquaux, G., & Cheplygina, V. (2022). Machine Learning for Medical Imaging: Methodological Failures and recommendations for the future. *Npj Digital Medicine*, *5*(1). https://doi.org/10.1038/s41746-022-00592-y

Waldron, L., Riester, M., Ramos, M., Parmigiani, G., & Birrer, M. (2016). The doppelgänger effect: Hidden duplicates in databases of transcriptome profiles. *Journal of the National Cancer Institute*, *108*(11). https://doi.org/10.1093/jnci/djw146

Wang, L. R., Choy, X. Y., & Goh, W. W. (2022). Doppelgänger spotting in biomedical gene expression data. *IScience*, *25*(8), 104788. https://doi.org/10.1016/j.isci.2022.104788

Wang, L. R., Wong, L., & Goh, W. W. (2021). How doppelgänger effects in biomedical data confound machine learning. *Drug Discovery Today*, *27*(3), 678–685. https://doi.org/10.1016/j.drudis.2021.10.017