Doppelganger effects in biomedical data

WU SIYI

Abstract Machine learning has many applications in the field of biomedical engineering, which requires a large amount of biomedical data to support it. But data doppelgangers occur when independently derived data are very similar to each other, causing ML models to perform falsely well. This report presents an overview of doppelgangers effects in biomedical data and how to avoid them in the practice and development of machine learning models for health and medical science. The report also illustrates that the effects are not unique to biomedical data，but also exist in other different fields. The pairwise Pearson's correlation coefficient (PPCC) as a worthy indicator is also highlighted, which facilitates the quantitative analysis of data doppelgangers such as the identification of data doppelgangers in samples and the impact of doppelgangers effects on ML models. It provides the reader with an understanding of the topic and perspective on further potential developments.

1 Introduction

In recent years, the rapid development of advanced computing and imaging systems in biomedical engineering has given rise to new dimensions of research, and the increasing scale of biomedical data requires accurate data mining algorithms based on machine learning[1]. These algorithms learn key attributes of the training dataset and correctly predict disease diagnosis in biomedical applications[2]. Machine learning models are also providing researchers with better ideas in the field of medicine, for example, machine learning-based classification models can be used to predict new drug-disease interactions[3], which can help screen drugs and reduce the cost of drug development.

However, using efficient machine learning algorithms to analyze biomedical data for classification and prediction is challenging because of its noisy and irrelevant characteristics[4]. Traditional validation methods such as Cross-validation are unable to guarantee the quality of ML models due to potentially biased training data and the complexity of the validation procedure itself. Suppose for some reason the training and validation data are extremely similar, this would

guarantee a good validation outcome, yet it informs little on whether a model has been learned intelligently (i.e., the doppelgangers effect)[5]. Therefore, how to effectively identify data doppelgangers in biomedical data and how to avoid doppelgänger effects in biomedical machine learning models is an area of current research.

2 Plenitude of doppelgangers effects in data

2.1 doppelgangers in biomedical data

The biomedical data have been categorized into a few broad data types including sequences (data generated by Omics technologies), signals, images and so on. High-throughput omics data (e.g., genomics, proteomics, and transcriptomics) is one of the most common types of biomedical data, which is often used to help understand the mechanisms that support disease onset and progression[6]. And data doppelgangers are not uncommon in high-throughput omics data. For instance, the reuse of tissue samples is common in clinical genomic studies, and this hidden duplication term may overstate the statistical significance or apparent accuracy of genomic models when combining data from different studies[7]. In addition, Cao and Fullwood provided a detailed analysis of existing systems for predicting chromatin interactions，the results showed that the performance of these systems can inflate statistical significance or apparent accuracy of genomic models when combining data from different studies[8]. We can also find the effects in the proteins structural data: in protein function prediction, proteins with similar sequences are inferred to be descended from the same ancestor protein and thereby inherit the function of that ancestor, this approach is unable to correctly predict functions for proteins with less similar sequences but similar functions[9]. In terms of images, face recognition is a biometric technology based on biological data about a person's facial features for identification. However, impact of doppelgangers on face recognition is always highlighted. Scientists presented a preliminary study on the ability of humans and automated face recognition to distinguish lookalikes, which means that the data of faces is similar. Their analysis showed most of automatic face recognition models were not able to correctly recognize lookalikes[10].

2.2 doppelgangers effects in other fields

Even though doppelgangers effects in other fields are poorly recorded, I do not think

doppelganger effects are unique to biomedical data. The paper of Wang et al suggested that data doppelgangers exist naturally as part of the similarity spectrum between renal cell carcinoma (RCC) samples[11]. This suggests that the effects are likely to exist in nature. Besides, in any database, it is possible that the training and validation data are extremely similar. This also may lead to doppelgangers effects. It is also possible to think of it in another way, where doppelgangers effects specific to the ML models and therefore the concept of data is less important. ML models are used not only in biomedical data, but also in data from other fields such as environmental science, business, etc. In short, further research is required to confirm these ideas.

3 Avoidance of doppelgangers effects in biomedical machine learning

3.1 Identification of data doppelgangers

Before using ML models, identifying the data d doppelgangers that exist in the training and validation sets is necessary. There were some studies working on this problem to identify data doppelgangers. When identifying duplicates in a pair of datasets, scientists calculated Pearson's Correlation Coefficient (PCC) between every sample in one dataset against every sample in the other dataset, i.e. Pairwise Pearson's Correlation Coefficient (PPCC). They found very similar expression profiles in cancer types with high PPCC, such as thyroid carcinoma, were hard to distinguish based on expression data only[7]. It means that the exceptional high PPCC value is likely to indicate that the samples contain data doppelgangers. But it does not build a conclusive connection between data doppelgangers and their ability to obfuscate ML tasks. And further reanalysis showed that their reported doppelgangers were the result of leakage instead of true data doppelgangers[11]. Sheng et al. developed a Bioconductor package "Dupchecker" to efficiently identify duplicate samples by generating MD5 fingerprints for the raw data[12]. However, dupChecker does not detect true data doppelgangers when samples are similar by chance. In fact, most biomedical datasets contain a large number of features, noisy information and complex patterns. It is not known which factors are responsible for the doppelganger effects.

3.2 Quantitation indicator: PPCC

Although PPCC does not reveal how data doppelgangers confound machine learning models, its design plausibility enables PPCC to be used as a quantitative indicator. Wang et al. used renal

cell carcinoma (RCC) proteomics data to construct negative, valid and positive cases and calculated PPCC. They eventually defined data doppelgängers as valid sample pairs with PPCC values greater than all negative sample pairs[11]. However, automatic setting of PPCC outlier thresholds is very difficult because its distributions are continuous. Apart from that, the impact of doppelgangers effects on ML models can be qualified by PPCC. It is experimentally proven that the more data doppelgangers in the training and validation sets, the more inflated the performance of ML[11]. Meanwhile, the doppelgangers effects are present to varying degrees in different ML models, which means that the compatibility between dataset and model is critical.

3.3 Prevention of doppelgangers effects

It is elusive to avoid doppelgangers effects completely, but there are some approaches to prevent them. In fact, most scientists ignore the data similarity in the training-validation pair. It is possible to use ordination methods, e.g., principal-component analysis, coupled with scatterplots, to see how the instances are scattered in multi-dimensional space[5]. But it is not an intuitive or robust way so another method regarding meta-data should be considered. According to meta-data, we are able to identify potential doppelgängers and assort them all into either training or validation sets[11]. Besides, performing data stratification, combined with a known data proportion, instead of evaluating model performance on whole test data can pinpoint gaps in the classifier. This would indicate which samples have data doppelgangers[11].

4 Conclusion

Doppelgangers effects are not rare in biomedical data and they are also likely to happen in other fields. Due to the fact that data doppelgangers and their doppelganger effects are poorly documented, the understanding of them and the ways to avoid them are still under discussion. PPCC as a good quantitative indicator of data doppelgangers contributes to identifying them in samples as well as making evaluation metrics for doppelgangers effects. Concerning the future of doppelgangers effects in biomedical data still should be studied for knowledge about them.

References

[1] Park, C. , Took, C. C. , & Seong, J. K. . (2018). Machine learning in biomedical engineering. Biomedical Engineering Letters, 1-3.

[2] Tanwani, A. K. , Afridi, M. J. , Shafiq, M. Z. , & Farooq, M. . (2009). Guidelines to Select Machine Learning Scheme for Classification of Biomedical Datasets. European Conference on Evolutionary Computation. Springer, Berlin, Heidelberg.

[3] Min, O. , Ahn, J. , & Yoon, Y. . (2014). A network-based classification model for deriving novel drug-disease associations and assessing their molecular actions. Plos One, 9.

[4] Das, H., Naik, B., & Behera, H. S. (2020). An experimental analysis of machine learning classification algorithms on biomedical data. In Proceedings of the 2nd International Conference on Communication, Devices and Computing (pp. 525-539). Springer, Singapore.

[5] Ho, S. Y., Phua, K., Wong, L., & Goh, W. W. B. (2020). Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability. *Patterns*, *1*(8), 100129.

[6] Goh, W. , & Wong, L. . (2018). Dealing with confounders in omics analysis. Trends in Biotechnology, 36(5).

[7] Waldron, L., Riester, M., Ramos, M., Parmigiani, G., & Birrer, M. (2016). The doppelgänger effect: hidden duplicates in databases of transcriptome profiles. *JNCI: Journal of the National Cancer Institute*, *108*(11).

[8] Cao, F., & Fullwood, M. J. (2019). Inflated performance measures in enhancer–promoter interaction-prediction methods. *Nature genetics*, *51*(8), 1196-1198.

[9] Friedberg, I. (2006). Automated protein function prediction—the genomic challenge. *Briefings in bioinformatics*, *7*(3), 225-242.

[10] Lamba, H., Sarkar, A., Vatsa, M., Singh, R., & Noore, A. (2011, October). Face recognition for look-alikes: A preliminary study. In *2011 International Joint Conference on Biometrics (IJCB)* (pp. 1-6). IEEE.

[11] Wang, L. R., Wong, L., & Goh, W. W. B. (2021). How doppelgänger effects in biomedical data confound machine learning. *Drug discovery today*.

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