Report of doppelgängers Effects on Biomedical Data

Abstract

The advancements of information technology, computational science and related processing techniques have created a fertile foundation for human progress. Artificial intelligence (AI) is widely used, and Machine learning (ML), an essential component in AI, has been integrated into many fields. However, the reliability of such ML methods can be affected by the presence of data doppelgängers. Data doppelgängers occur when independently derived data are highly similar to each other, causing models to perform well regardless of how they are trained (doppelgängers effect). In this report, I try to seek, sort out and summarize the methods of identifying data doppelgängers and the methods to reduce or eliminate this effect in relevant papers. In addition, I also expressed my own views on the identification and effect of data doppelgängers and put forward my own opinions. (In this report, as for the answers to the questions, I will put my thoughts and solutions in italics for easy identification.)

Introduction

In the field of biology, ML techniques have been used for the discovery and development of novel drug candidates. ML models have been increasingly used in drug discovery to accelerate drug development. They can shortlist better drug candidates faster, thereby reducing the time spent on discovery and testing. Classifiers in classification models have been used for the prediction of new drug-disease interactions and possible adverse drug reactions. Given the expensive drug-testing process, it is important that these classifiers are properly trained and tested to identify suitable drug candidates. [1]

When evaluating the classifier's performance, the classifier could still yield unreliable results after deriving training and test datasets independently. For example, models trained and validated on data doppelgängers (training and validation sets are similar) might perform well regardless of the quality of the training. When a classifier incorrectly performs well because of data doppelgängers, there is an observed doppelgänger effect. Data doppelgängers are called functional doppelgängers when they generate the doppelgänger effect. Although the biomedical data science community appears to be aware of data doppelgänger problems, it is surprising that procedures to eliminate or minimize similarities between test and training data still do not constitute standard practice. In fact, the re-use of tissue specimens is more widespread in clinical genomic studies, creating a doppelgängers effect in publicly available datasets, if left undetected, can inflate the apparent accuracy of genomic models. [2]

Data doppelgängers have been observed in modern bioinformatics as well. It is found that the performance of many models has been overstated recently due to issues in assessment methodologies, which may affect the reliability of subsequent studies. Therefore, in this paper, researchers and scholars attempt to investigate the nature of data doppelgängers, propose improved methods for identifying doppelgängers and find ways to mitigate the doppelgänger effect. It is important to reveal the true and objective performance of the model by eliminating or weakening the doppelgänger effect.

After searching a lot of relevant paper online, I found that regarding the doppelgänger effect (DE), the feature paper *only demonstrates the existence of DE within a single proteomics dataset. Hence, it does not show the pervasiveness of the DE across other data modalities such as high-throughput gene expression (genomics).* [3] *But researchers indicate that they are still trying to explore DE in two types of gene expression data sets, namely a well-studied microarray gene expression data from the study of Belorka and Wong* [4] *and widely available RNA-Seq gene expression data from the Cancer Cell Line Encyclopedia (CCLE) project* [5]*. It has probability that in the future they will find the existence of DE within other types of dataset.*

Possible Solutions

To better address the questions just mentioned in that paper, there are several solutions. Given the potential confusions of DE, it is crucial to be able to identify the presence of data doppelgängers between the training and validation sets before validation.

Firstly, this paper indicate that the data can be identified in a logical approach by using ordination methods (e.g., principal component analysis) or embedding methods (e.g., t-SNE), to see how samples are distributed in reduced-dimensional space. In this method, if the distribution of data doppelgängers can be showed in reduced-dimensional space and displayed by scatter plots, then this method could be very useful. However, such a method is unfeasible because the data doppelgängers are not necessarily distinguishable in reduced-dimensional space. So it is necessary to find some other methods that possibly help.

Then, some earlier studies working on similar problems have proposed measures for identifying data doppelgängers. One of these methods, dupChecker, identifies duplicate samples by comparing the MD5 fingerprints of their CEL files. But the same fingerprints indicate that samples are duplicates, which means that there are leakage issues. Thus, the results show the dupChecker does not actually detect true data doppelgängers that are independently derived samples. Another method, the pairwise Pearson's correlation coefficient (PPCC), captures the relationship between sample pairs of different data sets. An anomalously high PPCC value indicates that a pair of samples constitutes PPCC data doppelgängers (but it's impossible to determine which one is original). The prime limitation of the original PPCC paper was that it never conclusively made a link between PPCC data doppelgängers and their ability to confound ML tasks. However, the basic design of PPCC as a quantitation measure is methodologically reasonable.

Therefore, in this paper, researchers use the design idea of PPCC for identifying potential doppelgängers from constructed benchmark scenarios. After constructing benchmark scenarios with renal cell carcinoma (RCC) proteomics data taken from the NetProt software library, scientists simulate these scenarios across the two batches of the RCC dataset. They identified PPCC data doppelgängers based on the PPCC distribution and observed a high proportion of PPCC data doppelgängers surprisingly. This result suggests that data doppelgängers exist naturally as part of the similarity spectrum between samples. But as for why this happens, scientists cannot say whether this is a problem in PPCC or because the transcriptional profile of genes is, for the most part, positively correlated.

Confounding Effects of PPCC Data Doppelgängers

After identifying PPCC data doppelgängers in RCC, researchers explored their effects on validation accuracy across different randomly trained classifiers. In result, the more doppelgänger pairs represented in both training and validation sets, the more inflated the ML performance. This conclusion showed a dosage-based relationship between the number of PPCC data doppelgängers and the magnitude of the doppelgängers effect. In my opinion, this result also confirms that the PPCC data doppelgängers act as functional doppelgängers, producing inflationary effects. *And as the quantity of PPCC data doppelgängers increases in training set and validation set, the doppelgängers effect is more obvious and the performance of ML models are more inflated, which exactly shows how effect of PPCC data doppelgängers emerges from a quantitative angle.*

Ameliorating Data Doppelgängers

Comprehensive assessment strategies could be achieved by splitting training and test data based on individual chromosomes, as well as using different cell types to generate the training-evaluation pair. But this is difficult to do practically because it predicates the existence of prior knowledge and good quality contextual data.

In this research, authors also tried to alleviate the doppelgänger effect with methods that would not lead to a significant reduction in sample size or require a high amount of contextual data, although their attempts have met with failure thus far. However, they observed no change in the inflationary effects of the PPCC data doppelgängers after the removal of correlated variables. This observation hints at the extreme complexity of the doppelgänger effect, given that the reason for high correlations between sample pairs cannot simply be explained by a subset of highly correlated variables. In the future, scientists can look toward novel feature engineering and normalization approaches to address this more successful.

In conclusion, doppelgänger effects are not easy to resolve unfortunately and data doppelgängers are elusive to remove directly after trying different methods. Therefore, I suggest that to avoid performance inflation, it is important to check for potential doppelgängers in data before assortment in training and validation data.

For potential research directions in the future, there are three recommendations mentioned at the end of the paper. Recommendation one, perform careful cross-checks using meta-data as a guide. Recommendation two, perform data stratification instead of evaluating model performance on whole test data, stratifying data into strata of different similarities. Recommendation three, perform extremely robust independent validation checks involving as many data sets as possible. In the future, research could explore other methods of functional doppelgänger identification that do not rely that much on metadata. In such approaches, hopefully we could identify functional doppelgängers.

Supplement to Extra Questions

In the study of the Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles[2], they demonstrate the effectiveness of the method in databases containing dozens of datasets and thousands of breast, bladder, and colorectal cancer microarray profiles and of matching microarray and RNA sequencing expression profiles from The Cancer Genome Atlas (TCGA). They also identified probable duplicates among more than 50% of studies, originating in different continents, using different technologies, published years apart, and even within the TCGA itself. This study shows that there are data doppelgängers maybe exists in other data type such as RNA and gene sequencing. Moreover, scientists from the study provide the doppelgangR Bioconductor package for screening transcriptome databases for duplicates. Given the potential for unrecognized duplication to falsely inflate prediction accuracy and confidence in *differential expression, doppelganger-checking should be a part of standard procedure for combining multiple genomic datasets.* [2]

In the article named "Protocol to identify functional doppelgängers and verify biomedical gene expression data using doppelgangerIdentifier" [6]*, the author details the use of doppelgänger Identifier (DI), providing software and packages installation, data preparation, doppelgänger identification, and functional testing steps. They demonstrate examples with biomedical gene expression data and provide guidelines for the selection of userdefined function arguments.*

References

[1] Patel Lauv,Shukla Tripti,Huang Xiuzhen,Ussery David W & Wang Shanzhi.(2020).Machine Learning Methods in Drug Discovery.. Molecules (Basel, Switzerland)(22). doi:10.3390/molecules25225277.

[2] Waldron Levi,Riester Markus,Ramos Marcel... & Birrer Michael.(2016).The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles.. Journal of the National Cancer Institute(11). doi:10.1093/jnci/djw146.

[3] Wang Li Rong,Choy Xin Yun & Goh Wilson Wen Bin.(2022).Doppelgänger spotting in biomedical gene expression data. iScience(8). doi:10.1016/J.ISCI.2022.104788.

[4] Belorkar Abha & Wong Limsoon.(2016).GFS: fuzzy preprocessing for effective gene expression analysis.. BMC bioinformatics(Suppl 17). doi:10.1186/s12859-016-1327-8.

[5] Broad Institute. (2018) Cancer Cell Line Encyclopedia. Available at: https://sites.broadinstitute.org/ccle/ (Accessed: 12 March 2022).

[6] Wang Li Rong,Fan Xiuyi & Goh Wilson Wen Bin.(2022).Protocol to identify functional doppelgängers and verify biomedical gene expression data using doppelgangerIdentifier.. STAR protocols(4). doi:10.1016/J.XPRO.2022.101783.