Doppelgänger Effects in Machine Learning

Zheng Wenqing, C2338500

February 2023

1. Introduction

Gene sequencing and protein sequencing have facilitated the development of genomics and proteomics, providing a large amount of biomedical data for humans. With the development of machine learning in recent years, some researchers have used machine learning models for drug discovery or protein structure and function prediction. Data doppelgängers are widely present in biomedical data, which can confound the machine learning models, producing inflationary effects similar to data leakage. Thus, It is crucial to identify and avoid data doppelgängers. In this report, I explain the phenomenon of data doppelgängers, show their prevalence in some fields and propose methods to avoid them.

2. Data Doppelgänger and Doppelgänger effects

Data doppelgangers are independently-derived samples that possess similar characteristics to each other[1]. In machine learning, the training and test sets used for evaluation should be independently derived, but the data doppelgangers cause the ML models to perform surprisingly well regardless of how they are trained. However, when the models are deployed and used in real-world data, they do not perform well as we expected. We say this phenomenon is an observed doppelganger effect. In fact, Wang et al. found that data doppelgangers may not guarantee a doppelganger effect. Therefore, data doppelgangers that can cause data doppelganger effect are teamed functional doppelgangers.[2]

In order to know the doppelganger effect on model performance, Wang et al. generated six datasets, each containing 28 test samples and 8 validation samples as follows: (1) 0 PPCC data doppelgängers in validation; (2) 2 PPCC data doppelgängers in validation; (3)4 PPCC data doppelgängers in validation; (4)6 PPCC data doppelgängers in validation; (5)8 PPCC data doppelgängers in validation;

(6) Positive Control: Perfect leakage in validation. Using the KNN model and the Naive Bayes model as examples, it can be seen from **Figure.1** that the accuracy of models gradually increases (from around 40% to over 80%) as the number of data doppelgängers in the validation set increase. With all data doppelgängers in validation (8 PPCC data doppelgängers in validation), the accuracy of models is similar to that of the positive control (Perfect leakage in validation) dataset (all of them over 85%). This evidence suggests that data doppelgängers have an inflationary effect in ML models and that the inflationary effect becomes more obvious with more data doppelgängers available. However, Wang et al. also found that not all models were affected by data doppelgangers, such as the decision tree and logistic regression models. I guess the decision tree and logistic regression models focus more on common features of all samples, whereas fewer features are used in the Native Bayes and KNN models. Therefore, these models may be confounded by data doppelgangers easily. [2]

Figure.1 The prediction performance of different machine learning (ML) models on pairs of training-validation sets [2]

3. The prevalence of the Doppelgänger effect

Data doppelgangers are very pervasive in biomedical data (e.g. genes, proteins) because the same tissues of different bodies have similar functions. In recent years, some researchers have observed doppelganger effects in bioinformatics. Cao and Fullwood performed a detailed evaluation of existing chromatin prediction systems. They found that these systems were evaluated on test sets that had a high degree of similarity to the training set (data doppelgangers), and therefore the performance has been overstated. And in protein function prediction, the large amount of data doppelgangers lead the models to use the wrong features. As a result, the models could identify proteins with similar sequences and the same function, but could not correctly identify proteins with different sequences and different function. In the field of face recognition[3, 4], lookalikes such as twins can significantly affect the accuracy of the model, requiring other biometric characteristics such as fingerprints and irises for recognition.

However, doppelganger effects are not only present in biomedical data, but also in the field of Text similarity measurement[5, 6]. In most cases, similar texts have the same meaning. However, in some specific situations (e.g. anger), the same text may contain different meanings, or completely different texts may have the same meaning. Thus, there is a large amount of data doppelgangers in the text dataset, which may have a negative effect on the model's recognition ability of text similarity.

4. How to avoid doppelgänger effects

The most crucial step in eliminating doppelganger effects is to identify data doppelgangers in datasets and remove these data. Therefore, some researchers have proposed some methods to identify them.

1) Use correlation metrics.

The pairwise Pearson's correlation coefficient (PPCC) captures the relationship between sample pairs in different datasets. [2] A higher value indicates that the sample pair may constitute data doppelgangers. However, the method was initially used to detect data leakage rather than true data doppelgangers. However, the basic design of PPCC as a quantitative measure is reasonable. So It can be used to identify potential functional doppelgangers.

The Spearman Rank correlation coefficient firstly ranks the values in each sample and then ranks the variables using Pearson's correlation coefficient[7].he Spearman Rank correlation coefficient measures the monotonic relationship between the samples. Like the Spearman Rank correlation coefficient, the Kendall Rank correlation coefficient also measures the monotonic relationship between samples and these are more general than the PPCC[7].

2)use specific software

To deal with data doppelgangers in complex biomedical data, researchers have presented software such as doppelgangerIdentifier, which not only identifies data doppelgangers but can also be used for assaying data outliers and anomalies. The doppelggangerIdentifier can be used as a tool for constructing training and validation sets. Researchers can identify data doppelgangers with getPPCCDoppelgangers before the training-validation split. And verifyDoppelgagers performs functionality tests to check whether the identified data doppelgangers are functional doppelgangers and whether all functional doppelgangers have been identified.

3) Use meta-data for cross-checking

Metadata contains most of the information about a sample. Therefore, using thes information from the meta-data, we can identify potential data doppelgangers and classify them all into the training sets or validation sets, effectively preventing the doppelganger effects. [2]

At present, it is difficult to identify data doppelgangers in datasets, doppelganger effects can be eliminated by processing datasets. For example, we can perform data stratification and choose an appropriate dataset for prediction, or we can add the amount of as much data as possible.

I think it is feasible to increase the number of data. After the model has been trained, it should be tested using multiple datasets. We can add the data that do not perform well to the training sets or test sets, and remove or put back the data that perform better into the training set. By adding new datasets continually to the model, I guess the data doppelgangers can be removed a little and the effects can be eliminated.

5. Conclusion

In conclusion, doppelganger effects are not only limited to biomedical data, but also exist in other areas. Unfortunately, doppelganger effects are not easily resolved by analysis, and there is currently no algorithm that can identify all data doppelgangers in a dataset correctly. Therefore, the training and validation dataset should be carefully checked before we train the model. For example, we can increase the number of samples or perform careful cross-checks using meta-data to eliminate doppelganger effects in models.

Reference

1. Wang, L.R., X. Fan, and W.W.B. Goh, *Protocol to identify functional doppelgangers and verify biomedical gene expression data using doppelgangerIdentifier.* STAR Protoc, 2022. **3**(4): p. 101783.
2. Wang, L.R., L. Wong, and W.W.B. Goh, *How doppelganger effects in biomedical data confound machine learning*. Drug Discov Today, 2022. **27**(3): p. 678-685.
3. Rathgeb, C., et al., *Reliable detection of doppelgängers based on deep face representations.* IET Biometrics, 2022. **11**(3): p. 215-224.
4. Rathgeb, C., et al., *Impact of Doppelgängers on Face Recognition: Database and Evaluation*, in *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*. 2021. p. 1-4.
5. Li, C., F. Liu, and P. Li, *Text Similarity Computation Model for Identifying Rumor Based on Bayesian Network in Microblog.* The International Arab Journal of Information Technology, 2020. **17**(5): p. 731-741.
6. Wang, J. and Y. Dong, *Measurement of Text Similarity: A Survey.* Information, 2020. **11**(9).
7. Wang, L.R., X.Y. Choy, and W.W.B. Goh, *Doppelganger spotting in biomedical gene expression data.* iScience, 2022. **25**(8): p. 104788.