The effect of doppelgänger

——The report of reading 'How doppelgänger effects in biomedical data confound machine learning'

When the cross-validation technique is applied to the evaluation model, potential targets will be quickly identified, followed by the emergence of data doppelgängers . This means that if there are very similar points in the independently derived data, data doppelgängers will appear. So that the correctness and authenticity of the final model will be greatly reduced, even though the final model will perform well. In the process of analyzing biomedical data, data doppelgängers often appears with confounding effects . To mitigate the effects of doppelgänger, identification data doppelgängers will be researched and adopted prior to train-validation.

In the original text, the author analyzed data doppelgängers through the following points:

1. Introduction
2. Motivation
3. Design & Solution
4. Result of above Solution
5. Conclusion

Introduction

ML models are increasingly being used in many ways, not only to speed up drug discovery by faster selection of better drugs, but also to greatly reduce the cost of drug development for the treatment of other diseases. A new anticancer drug candidate, EXS21546, was also discovered through Exscientia's 'Centaur Chemist' artificial intelligence (AI) platform. Therefore, classifiers (training models) have been applied to predict new drug–disease interactions and possible adverse drug reactions.

However, in ML, independent training and test sets can still produce unreliable validation results. For example, due to chance or other reasons, the training and validation sets will have highly similar and so a model trained and validated on the data doppelgängers may perform well . Regardless of the quality of training, we say that there is an observed doppelgängers effect when the classifier mistakenly performs well due to the presence of the data doppelgängers . Therefore, data doppelgängers that also produce the doppelgängers effect (confounding ML

outcomes) are called functional doppelgangers. Here we hope to understand better level of similarity between suspected functional doppelgänger and the acceptable proportion of functional doppelgänger in the validation set.

Motivation

Although several methods for identifying data doppelgänger exist, most are not generic or robust enough. Therefore, the similarity between the data generated by data doppelgänger has not been eliminated so far. It is necessary to study the properties of doppelgänger data and propose an improved doppelgänger recognition method.

Design & Solution

- Firstly, a logical way for doppelgänger to identify data is to use either a sorting method (e.g., principal component analysis) or an embedding method (e.g., T-SNE) with a scatter plot. And see how the sample is distributed in a dimensionality reduction space.
- Second, earlier studies on similar problems also proposed one of the methods for identifying data doppelgänger - dupChecker. Identify duplicate samples by comparing the MD5 fingerprints of their CEL files.
- Another measure, the pairwise Pearson correlation coefficient (PPCC), captures the relationship between sample pairs of different data sets. An unusually high PPCC value indicates that a pair of samples constitutes the PPCC data doppelgänger (note that it is impossible to determine which of the pair of samples is original). Although the PPCC method has some limitations, the basic design of PPCC as a quantitation measure is reasonable methodologically. In the valid scenario, the distribution of PPCC is a broad continuum with no obvious fracture. This suggests that the use of outlier detection methods (as recommended in the original PPCC paper) would not be sensitive enough. It also shows that the data doppelgänger exist naturally (and are not surprising anomalies) as part of a similar spectrum between samples. As to why this occurs, we cannot say for sure whether it is the PPCC itself or because the transcription profile of genes is positively correlated in most cases we examined the distribution of PPCC between the same and different tissue pairs. PPCC values for the same tissue pairs remained high overall, indicating a high correlation between samples, even if they came from different patients. This is not surprising, since many genes share common regulators. However, if we compare the different class effect pairs, the PPCC distribution is definitely lower. In contrast, the PPCC is also very high when we consider duplications from the same samples or tissues. These evaluations indicate that PPCC has significant identifying value.

The Result of above Solution

- Initially, we find that this approach is not feasible because the data doppelgänger are not necessarily distinguishable in dimensionality reduction space.
- For method of dupChecker, identical MD5 fingerprint indicates that the samples are duplicate, which means that the essential replicates and so there are indicative of leakage issues. As a conclusion, it does not detect the true data doppelgänger, which are independently exported samples that are accidentally similar.
- After identifying PPCC data doppelgangers in RCC, we investigate their impact on the validation accuracy of different random training classifiers (a training classifier is an ML model that "learns" from training data).This will determine whether PPCC data doppelgangers act as functional doppelgangers with a significant inflationary effect on ML performance. We note that the presence of PPCC data doppelgangers in the training and validation data improves ML performance even if the features are randomly selected. Thus, it is meaningless, which means that the performance of the model should be poor during validation. Also the paper found that showed consistent reproducibility across different training and validation datasets and across different ML models. In addition, the more doppelganger pairs represented in the training and validation set, the better ML performance. This indicates a dose relationship between the number of PPCC data doppelgangers and the size of the doppelganger effect. The result confirms that PPCC data doppelgangers (based on pairwise correlation), acting as functional doppelgangers (confounds ML results), producing an inflating effect similar to data leakage. The similarities between the doppelganger effect and leakage are evident in our experiments using the K-nearest Neighbor (kNN) model, where the training validation set with 8 doppelgangers showed the same accuracy distribution in the validation as the perfect leak training validation set. However, not all models were equally affected: by comparing with decision trees and logistic regression models, kNN and Naive Bayes models had a clearer linear relationship between performance inflation and doppelganger dosage. Apparently, when all PPCC data doppelgangers were placed in the training set, the doppelganger effect was eliminated. This provides a possible way to avoid the Doppelganger effect. However, containing PPCC data doppelgangers to either training set or validation set are suboptimal solutions. When the size of the training set is fixed, therefore, each contained data results in a less similar sample being excluded from the training set, which means that the model does not generalize well due to lack of knowledge in a model.

After reading this report, there are my own opinion and suggestions. In order to show clearly, this part will be divided by 3 sections.

1. My Understanding
2. Suggestions
3. Extra information

My Understanding

The appearance of data doppelgänger effect is when there is a similarity between the training and test sets through using ML models. This means that there will not be a limitation of the effect of the data doppelgängers on biomedical data uniquely and the doppelganger effects still exist in other area. For example, detect malfunctioning machine parts in industrial and manufacturing applications and predict the data about objects and preferences in the service industry and so on.

Suggestions

If we cannot remove the doppelgängers data, then we should use other methods to prevent it. This article proposes three approaches: The first approach is to carefully perform cross-checking using metadata as a guide. The second approach is to implement data layering. The third approach is to perform very robust independent validation checks involving as many data sets as possible (divergent validation).

Extra information

Whole-genome analysis of cancer specimens is commonplace, and investigators frequently share or re-use specimens in later studies. Duplicate expression profiles in public databases will impact re-analysis if left undetected, a so-called “doppelgänger” effect.[1]

Reference:
[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5241903/