Analysis of doppelgänger effects for machine learning

Jingyuan Chen

Introduction

Drug development is increasingly using machine learning (ML) models to speed up the identification of potential targets and improve the efficacy of drug testing. When dealing with biomedical data, ML always has the effect of data doppelgängers, which is hardly uncharacteristic. However, there have been several methods identified to mitigate the doppelgänger effect**.**

Doppelgänger effects

In machine learning, we usually use classification models which contain a classifier to predict data. To obtain a precise result, the classifiers need to be trained properly and to be more precise, the training and test data set need to be independent. Otherwise, it will contain data doppelgängers create the doppelgänger effect.

Data doppelgängers occur when independently derived data are like each other. Data doppelgängers will then cause models to perform well regardless of how they are trained, which is known as the doppelgänger effect. However, this may not guarantee a doppelgänger effect, data doppelgängers that generate a doppelgängers effect are termed functional doppelgängers.

Data doppelgängers occur in bioinformatics. It has been observed in many cases, for example, Cao and Fullwood evaluated existing chromatin interaction prediction systems. These systems are overstated because they used high-similarity training sets. In the established field of bioinformatics such as protein function prediction, a protein with similar sequences is inferred to be descended from the same ancestor protein which gives us a false impression of prediction. What's more, in drug discovery, the QSAR model. The Doppelgänger effect is not only existed in biomedical data. For instance, it also occurs in psychology and refers to the experience of a direct encounter with oneself and his/her double which shares the same personality and identity. The perceptual element is usually a hallucination. This effect has been described in individuals suffering from overwhelming fear, severe anxiety or intoxication, epilepsy, as well as in the sleep-wakefulness transition. It has also been reported in major psychoses [1]. Another example could be the doppelgänger effect in gene sequences. Analysing the DNA sequences of unrelated look-alikes, revealed that doppelgängers are likely to share genetic similarities which could influence their

facial appearance [2]. Whole-genome analysis of cancer specimens is also an example. Investigators share or re-use specimens in later studies. Duplicate expression profiles in public databases will impact re-analysis if left undetected. [3] Given the potential for unrecognized duplication to falsely inflate prediction accuracy.
From the quantitative perspective, to express the doppelgängers effect. it should be run by when we are going to use machine learning to analyse the data set, we will have the procedure training, validation, and test. We use one labelled dataset and chop it into three and use the rule of thumb of 2:1:1. The process is mainly about we put one of the labelled data set into a training model to train the data for a different model, then wen compares trained models on validation data pick the best one and compete against someone else on test data. This process is mainly to look for the lowest validation error.

Identification of data doppelgängers

Although scientists are aware of data doppelgänger problems and trying to identify the data doppelgangers using the following three methods. The first one is through reducing dimensionality using ordination such as PCA or t-SNE. But they still can't feasibly see the problem because data doppelgängers are not distinguishable in reduced dimensional space. Secondly, working on similar probability by using the dupChecker method or PPCC. Although they still have problems, we still pick PPCC as it is a quantitative measure, which makes it reasonable to use.

Effect of PPCC on machine learning

PPCC data doppelgänger has an inflation effect on machine learning which is like data leakage. We constructed an experiment on PPCC data doppelgänger and non-PPCC doppelgänger and find out that with PPCC doppelgänger, the data performed better with the large data set. However, this effect is unbalanced, it occurs in some models but not all models.

In this example, RCC proteomics data is used to create benchmark scenarios that could identify the data doppelgänger. There are two scenarios: negative cases and valid cases. For negative cases, the sample pairs are constructed from different class labels; while for valid cases, samples are assigned to the same class label but from different samples.

With the comparison of negative cases and valid cases, we plot a graph and see a high PPCC value. This indicates that similarities exist naturally as part of the similarity spectrum between samples. The high PPCC value might occur because of the transcriptional profile of genes that shares common regulators. This example supports that the PPCC has meaningful discrimination value although with less sensitivity.

Methods avoiding doppelgängers effect

Firstly, we could perform careful cross-checks using meta-data as a guide to identify potential doppelgängers and assorts them into training or validation sets. Secondly, stratify data into strata of different similarities to evaluate model performance on each stratum separately. The last one is to use divergent validation techniques to inform on the objectivity and generalizability of the model. What's more, we could reduce the doppelgängers effect by other methods such as support vector machines, that is, by theory, using separating hyperplanes to cut up the D-dimensional inputs space and dividing classifiers into K classes. The solution optimises the margin between data points and the hyperplane. For non-linear boundaries we could use the kernel trick this method is like Knn but with a better result. Otherwise, we might use artificial neural works that work well in high dimensions with supervised classification.

Conclusion

Doppelgänger effects inflated machine learning which can provide a negative impact on the usefulness of machine learning for phenotype analysis and drug identification. We could check the potential doppelgängers in data before training and validation to avoid inflation.

Reference

[1] Cristiano Barbieri, Gabriele Rocca, Caterina Bosco, Lucia Tattoli, Ignazio Grattagliano & Giancarlo Di Vella (2022) The Doppelgänger phenomenon and death: a peculiar case of homicide by a subject with first-episode psychosis, ForensicSciencesResearch, 7:4, 798-802, DOI: 10.1080/20961790.2022.2055827

[2] New York Times (2022), Your doppelgänger is out there and you probably share DNA with them.

[3] Levi Waldron, Markus Riester, Marcel Ramos, Giovanni Parmigiani, Michael Birrer (2016), The Doppelgänger Effect: Hidden Duplicates in databases of Transcriptome Profiles, JNCI J Natl Cancer Inst 108(11): djw146