Understanding and Analysis of Doppelgänger Effects

Shuchen Ye

Abstract

Doppelgängers are data when samples appear to be similar in their measurements, resulting in a doppelgänger effect that exaggerates the effectiveness of a neural network training system, so the evaluation results are not credible. The doppelgänger effect is very common in biomedical data, where it is inevitable that similar molecules with similar activity are classified into training and validation sets. But the doppelgänger effect is certainly not unique to biomedical data, and there are many cases in the social sciences and economics where models are inaccurate due to the doppelgänger effect. The disadvantages of the doppelgänger effect are obvious, but it is not necessary to avoid the doppelgänger effect in all cases. In specific cases, we need the high accuracy of the model due to the doppelgänger effect without making the model general. However, in the vast majority of cases, we still need to identify the doppelgänger effect and deal with it. Several identification methods are mentioned in the paper. 1. Sorting method 2. Embedding method. However, these two methods are not feasible because we cannot determine whether the data can still distinguish the doppelgänger effect in the reduced dimensional space. We therefore also propose methods to identify doppelgängers data by comparing MD5 fingerprints of CEL files and using pairwise Pearson correlation coefficients. To address the doppelgängers effect, we try to place the doppelgängers data in the training or validation set, but this is of course not optimal. We also speculate that the doppelgängers data can be deleted to reduce the effect, but if only variables that are significantly related to the doppelgängers effect are deleted, the doppelgängers effect cannot be removed. We need to consider other ways to avoid the doppelgängers effect. Such as careful cross-checking with metadata as a guide, data stratification, or very rigorous independent validation checks.

Content

Understanding of doppelgänger effects

When constructing a neural network, we need to divide the data into training and test sets, and this process is usually random. This stochastic process thus opens up the possibility of doppelgänger effects. Since the data of the training set and the test set generally come from the same database, the data of the training set and the test set are highly similar. Training in Quantitative Structure-Activity Relationship (QSAR) Models to Predict the Biological Activity of Molecules from their Structural Properties QSAR models assume that structurally similar molecules have similar activities. In most cases, this assumption is true. As described in the paper, classifying similar molecules with similar activity into training and validation sets can easily confuse model validation because a poorly trained model may still perform well on these molecules.

We can infer from this concept that such doppelgänger effects are not specific to biomedical data. For example, when we study the effect of household income on children's educational attainment, the data we get may be collected by a city's statistics bureau, so the data may have a high degree of regional consistency. It is not difficult to infer that the data collection should be carried out in different types of schools, so the data may also have class similarities due to limited statistical channels. Because if the family income is too low and children drop out of school early, it is difficult to be counted into the database. Building a machine learning model in such a situation is very lacking in generality. In theory, this model should fit poorly because of the lack of complete types of data. However, since the validation set is also from this database, the

validation results may be very good. (Wang, 2016)

This example supports my point: doppelgänger effects affect any kind of data analysis, not just biomedical data.

Evaluation of doppelgänger effects in machine learning models

The repeated use of tissue samples is common in clinical genomics research, creating a "cloning effect" in publicly available datasets: hidden duplications, if undetected, may exaggerate the power of genomic models when combining data from different studies Statistical significance or apparent accuracy. (Waldron, Riester, Ramos, Parmigiani, & Birrer, 2016)

A well-trained model could theoretically still perform well on particular instances since they are trained on the structural properties of the information and thus be able to detect small changes.

The downside of doppelgänger effects seems well understood, but does it have to be addressed? In some specific cases, when there is little data and a lack of doppelgängers, the model tends to perform poorly. So as mentioned in the paper: All ML models perform better on PPCC data doppelgängers than non-PPCC data doppelgängers. Consistent with the paper's point about protein sequence-function prediction, we do not need to guarantee generality to less similar examples in some practical cases.

Attempts to avoid doppelgängers effect in machine learning model training

Although procedures to eliminate or minimize the similarity between test and training data before classifier evaluation still do not constitute standard practice, in this paper

we propose several logical approaches and workarounds for identifying data avatars.

Recognition method We propose a ranking method and an embedding method, but we also find that the principal component analysis method in the ranking method is not feasible, because the data avatars are not necessarily distinguishable in the dimensionality reduction space. After judging that the two methods are not feasible, we further proposed to compare the MD5 fingerprints and use the pairwise Pearson correlation coefficient to find the specific data that constitutes the doppelgänger of the data. dupChecker identifies duplicate samples by comparing the MD5 fingerprints of the CEL files of the test data, but the samples detected by this method are completely consistent and cannot detect the real clone data. And we can also monitor the Pearson correlation coefficient (PPCC) between samples. If the PPCC value is abnormally high, it proves that the two sample data are separate. The basic design of PPCC as a quantitative measure is methodologically sound.

To remove the doppelgängers effect, we try to include all the doppelgängers data in the training or test set and remove all or only the data that is highly correlated with the doppelgängers effect. Although restricting the data doppelgängers to training and validation sets is suboptimal and reduces accuracy, it does eliminate the doppelgängers effect. However, after removing the doppelgängers effect by only removing data that is highly correlated with the doppelgängers effect, the PPCC data did not change significantly, so we can guess that the doppelgängers effect is extremely complex, so we cannot simply apply the doppelgängers effect to several groups. highly correlated variables to explain.

Therefore, we realize that the doppelgängers effect is a complex process that we cannot easily eliminate, but we still need to work hard to prevent the doppelgängers effect. The paper presents three approaches, using metadata as a guide for careful cross-checking, performing hierarchical data analysis, and performing very robust independent validation checks. Metadata can be used as a guide for careful cross-checking when it is difficult to directly avoid avatar data generation. This allows us to predict the range

of PPCC scores and whether there is leakage in the absence of avatars. With this information in the metadata, we can identify potential avatars and classify them all into training or validation sets, effectively preventing the avatar effect and allowing a relatively more objective assessment of ML performance. We can also stratify the data into layers with different degrees of similarity and evaluate the model performance of each layer separately, rather than the entire experimental data. This method can largely avoid the double effect. In addition to this, we can perform independent validation checks to evaluate the classifier using as many datasets and different validation techniques as possible. This method is very time-consuming and energy-intensive, but it can effectively prove the generality of the model. Even if there is avatar data in the training set, the avatar effect is negligible due to the richness of the overall data.

We can explore other methods of identifying doppelgänger during future data analysis, but whichever method is used, we can find the subset of correct predictions in the validation set after the training of the machine learning model, and use it in future model evaluations Avoid using these subsets. These correctly predicted subsets are potential doppelgänger data, and it makes more sense to analyze the incorrectly predicted subsets.

In summary, the impact of avatars is not easy to resolve analytically. Therefore, to avoid performance bloat, it is important to examine the data for potential avatars before training and validating data classification.

Reference

Waldron, L., Riester, M., Ramos, M., Parmigiani, G., & Birrer, M. (2016, July 05). *The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles* . Retrieved froOXFORD ACADEMIC: https://academic.oup.com/jnci/article/108/11/djw146/2576926?login=false

Wang, X. (2016, 2 16). The effect of family background on academic performance. *Social Sciences II*.