Doppelgänger Effects

1 Introduction

Doppelgänger effects, also known as doppelgänger biases, refer to the phenomenon whereby the performance of a machine learning (ML) model is overoptimistic when tested on data that is similar to the training data. These effects can occur when the training and test data sets are not independently derived, leading to biased and unreliable results. It is well established in ML that, when assessing the performance of a classifier, the training and test data sets should be independently derived. However, independently derived training and test sets could still yield unreliable validation results. For example, models trained and validated on data doppelgängers (where training and validation sets are highly similar because of chance or otherwise) may yield unrealistic performance results. When this happens, we say that there is an observed doppelgänger effect.

In the context of biomedical data, doppelgänger effects can be particularly problematic in drug discovery, where the cost and time required for testing is high. It is therefore important to ensure that the training and test data sets are independently derived and representative of the population of interest to avoid doppelgänger effects and ensure the reliability and generalizability of ML models.

2 Non-uniqueness

Doppelgänger effects are not unique to biomedical data, and have been observed in other types of data such as imaging, gene sequencing, and metabonomics. For example, in a study of chromatin interaction prediction systems, Cao and Fullwood found that the performance of these systems was overstated due to the presence of data doppelgängers in the test sets. Goh and Wong also observed doppelgänger effects in their study of gene expression data, where certain validation data were guaranteed good performance given a particular training data set, even if the selected features were random.

3 Other Examples

3.1 Imaging

In the field of imaging, doppelgänger effects can occur when the training and test data sets are not representative of the overall population of images. For example, doppelgänger effects can be present in the performance of a convolutional neural network (CNN) model trained to classify magnetic resonance imaging (MRI) scans of the brain into normal and abnormal categories.

3.2 Gene Sequencing

Doppelgänger effects have also been observed in the analysis of gene sequencing data. In a study by Wang et al., doppelgänger effects were present in the performance of a support vector machine (SVM) model trained to classify gene expression data into different cancer types.

3.3 Metabonomics

In the field of metabonomics, doppelgänger effects can occur when the training and test data sets are not independent and representative of the overall population of metabolic profiles. For example, doppelgänger effects can be present in the performance of a partial least squares discriminant analysis (PLS-DA) model trained to classify metabolic profiles of human urine samples into different disease states.

4 How Doppelgänger Effects Emerge

Doppelgänger effects can emerge from a variety of sources, including sampling biases, selection biases, and inherent correlations in the data.

4.1 Sampling biases

Sampling biases can occur when the training and test data sets are not representative of the overall population, leading to doppelgänger effects due to the similarity between the two sets. Sampling biases can be caused by a variety of factors, such as non-random sampling methods or the inclusion of a disproportionate number of certain types of data points in the training or test sets. For example, if the training set is skewed towards a certain class of data points, the model may perform well on the test set due to the presence of similar data points, rather than due to its actual predictive ability.

4.2 Selection biases

Selection biases can also contribute to doppelgänger effects, as the choice of features used for training and testing can affect the similarity between the training and test sets. If the same features are used for both training and testing, the model may perform well on the test set due to the similarity between the two sets, rather than due to its ability to generalize to new data.

4.3 Inherent correlations

Inherent correlations in the data can also lead to doppelgänger effects, as highly correlated features can result in similar data points even when the data sets are independently derived. For example, if two features are highly correlated, the presence of one feature in a data point may strongly predict the presence of the other feature, leading to doppelgänger effects if the two features are included in both the training and test sets.

These are just a few examples of how doppelgänger effects can emerge from a quantitative angle. It is important to carefully consider these factors when designing training and test data sets in order to avoid doppelgänger effects and obtain reliable performance estimates for ML models.

5 Methods to Avoid

5.1 Data augmentation

There are several methods that can be used to identify and mitigate doppelgänger effects in ML models. One approach is the use of data augmentation techniques, which involve artificially generating new data points from existing data to increase the size and diversity of the dataset. This can help to reduce the likelihood of doppelgängers appearing in the training and test sets. Data augmentation techniques can include techniques such as adding noise to the data, rotating or shifting images, and synthesizing new data points from existing ones.

5.2 Selective sampling

Selective sampling is another method that can be used to avoid doppelgänger effects. This involves carefully selecting a subset of the data that is representative of the overall population, rather than using the entire dataset for training and testing. This can help to ensure that the training and test sets are independent and dissimilar, reducing the risk of doppelgänger effects. Selective sampling can be done using techniques such as stratified sampling, which ensures that the proportion of different classes in the sample is representative of the overall.

6 Reference

Wang L R, Wong L, Goh W W B. How doppelgänger effects in biomedical data confound machine learning[J]. Drug Discovery Today, 2021.

Cao F, Fullwood M J. Inflated performance measures in enhancer–promoter interaction-prediction methods[J]. Nature genetics, 2019, 51(8): 1196-1198.