Doppelgänger effects and their solutions

Abstract

Doppelgänger effects refer to the phenomenon of two or more individuals having very similar characteristics or features. These effects can occur in any type of data, including biomedical data, financial data, customer data, and social media data. In the context of machine learning, doppelgänger effects can pose a challenge because they can lead to confusion and bias in the model. This paper proposes several feasible solutions to Doppelgänger effects.

Background

In biomedical data, Doppelgänger effects can negatively impact the performance of machine learning models because there can be a lot of noise and uncertainty in the biomedical data, which can lead to unstable output of the model. For example, in medical image analysis, the same medical image may produce different outputs under different scanning conditions, as the quality of the medical image can be affected by many factors (e.g., the quality of the scanning device, the patient's position, the scanning time, etc.). In addition, there may be other noises in the biomedical data, such as measurement errors, human factors, etc., which can also cause the output of the model to be unstable.For example, if we are training a machine learning model to predict the likelihood of a patient developing a certain disease, and we have two patients with very similar characteristics, it may be difficult for the model to accurately predict which patient is more likely to develop the disease. This is because the model may be unable to distinguish between the two patients and their respective outcomes, leading to inaccurate predictions.

In the given paper, pairwise Pearson’s correlation coeffificient (PPCC) is used to identify data doppelgängers. The Pearson's correlation coefficient is a measure of the linear correlation between two variables. It is a statistical measure that ranges from -1 to 1, where a value of -1 indicates a strong negative correlation, a value of 0 indicates no correlation, and a value of 1 indicates a strong positive correlation.The Pearson's correlation coefficient is often calculated on a pairwise basis, meaning that it is calculated between two specific variables.The Pearson's correlation coefficient is a widely used measure of correlation in statistical analysis, and is particularly useful for understanding the linear relationship between two variables. In the field of statistics and machine learning, Pearson coefficient is often used to analyze the correlation between samples or features for statistical analysis or feature dimensionality reduction. Therefore, it is reasonable to use PPCC to measure the doppelgängers.

Doppelgänger effects can occur in any type of data where there is the potential for

duplicate or similar data to be present. In the field of imaging, for example, doppelgänger effects can occur if there are multiple copies of the same image in a dataset, or if there are similar images that are difficult to distinguish from one another. In gene sequencing, doppelgänger effects can occur if there are multiple copies of the same gene sequence present in a dataset, or if there are similar gene sequences that are difficult to distinguish from one another 1. In metabonomics, doppelgänger effects can occur if there are multiple copies of the same metabolic profile present in a dataset, or if there are similar metabolic profiles that are difficult to distinguish from one another.

For example, doppelgänger effects can potentially occur in MRI data of gliomas, as well as in any other type of medical imaging data. In the context of MRI data of gliomas, doppelgänger effects could potentially arise if there are multiple copies of the same MRI scan present in a dataset, or if there are similar MRI scans that are difficult to distinguish from one another. Another example is the data of genome when researchers use same batch of samples in their different studies. As a result, duplicate expression profifiles in public databases will impact re-analysis if left undetected.

Obviously, doppelgänger effects are not unique to biomedical data.Here are a few examples of how doppelganger effects can occur in different types of data:

Financial data: In financial data, doppelganger effects can occur when a transaction is recorded multiple times, such as when a payment is entered into a system twice or a transaction is recorded in two different accounts. This can result in inaccurate financial reports and make it difficult to track the flow of money within an organization.

Customer data: In customer data, doppelganger effects can occur when a customer is represented by multiple records due to changes in their contact information or the use of different systems to store customer data. This can make it difficult to track customer interactions and history, and can lead to duplicate marketing efforts.

Marketing data: In marketing data, doppelganger effects can occur when a customer is represented by multiple records due to the use of different systems to store customer data or the collection of customer data from multiple sources. This can make it difficult to accurately segment and target marketing efforts and can lead to the waste of marketing resources.

Solutions

To avoid doppelganger effects when developing machine learning models for health and medical science, we can solve the problem from two aspects: data and model.
In terms of data, it is important to ensure that the data used to train and test the model is accurate and up-to-date. This may involve the use of unique identifiers for each

patient, regular data cleansing and de-duplication, and the integration of data from multiple sources.

Here are a few specific steps that can be taken to avoid doppelganger effects in the practice and development of machine learning models for health and medical science:

1.Use unique identifiers for each patient: By using unique identifiers for each patient, such as a patient ID number or a combination of personal information, it is possible to accurately identify and track individual patients and prevent the creation of duplicate records.

2.Regularly clean and de-duplicate the data: By regularly cleaning and de-duplicating the data, it is possible to identify and merge duplicate records, ensuring that the data used to train and test the model is accurate and up-to-date.

3.Integrate data from multiple sources: By integrating data from multiple sources, it is possible to create a more complete and accurate representation of each patient, which can improve the performance of the machine learning model.

Ensemble learning is also a effective technique to reduce the impact of doppelganger effects on machine learning models,, which combines the predictions of multiple models to improve the overall performance of the model. Ensemble learning can be used to avoid doppelganger effects in a number of ways:

Data de-duplication: By de-duplicating data before training an ensemble model, it is possible to minimize the impact of doppelganger effects on the model. This can be done through the use of software tools or by manually reviewing and merging duplicate records.

Model diversity: By training multiple models with different algorithms or using different sets of features, it is possible to create an ensemble model that is less susceptible to the effects of doppelganger records. This is because the models in the ensemble will have different strengths and weaknesses, and will be less likely to be affected by any errors or inconsistencies in the data.

Data quality checks: By including data quality checks as part of the model training process, it is possible to identify and correct any errors or inconsistencies in the data before the model is trained. This can help reduce the impact of doppelganger effects on the model.

Conclusion

Doppelgänger effects generally exist in various data, especially biomedical data. Its existence will lead to inaccurate data, waste of resources, poor model performance

and other problems. The impact of Doppelgänger effects can be reduced by integrating and sorting data, using ensemble learning and other methods, but the method to completely avoid this effect is still under research.

Reference

1. Waldron L, Riester M, Ramos M, Parmigiani G, Birrer M. The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles. J Natl Cancer Inst. 2016 Jul 5;108(11):djw146. doi: 10.1093/jnci/djw146. PMID: 27381624; PMCID: PMC5241903.