Report

Introduction

The paper "How doppelgänger effects in biomedical data confound machine learning" by Li Rong Wang et al. (2021) discusses the issue of doppelganger effects in biomedical data and how they can cause problems for machine learning models used in healthcare. In this report, I will provide an overview of the paper, and then discuss whether doppelganger effects are unique to biomedical data, how they can be avoided in the practice and development of machine learning models, and provide some examples of doppelganger effects in other data types. I will also discuss the quantitative basis of doppelganger effects and propose some ways to avoid or check for these effects.

Overview of the Paper

The paper begins by defining doppelganger effects as the phenomenon where distinct patients may have similar or identical electronic health records (EHRs), and this can cause confusion for machine learning algorithms. The authors argue that doppelganger effects are unique to biomedical data because of the complexity and heterogeneity of such data, as well as the fact that EHRs are often

incomplete, noisy, and contain missing values. The paper discusses some methods to mitigate doppelganger effects, including feature selection, model-based approaches, and adversarial learning.

Are doppelganger effects unique to biomedical data?

While the authors argue that doppelganger effects are unique to biomedical data, there are some examples of doppelganger effects in other data types. For example, in imaging data, it is possible for two different individuals to have very similar or identical images due to factors such as imaging equipment, posture, or body composition. In gene sequencing data, it is possible for two different individuals to have very similar or identical genetic sequences due to factors such as ancestry, shared mutations, or genomic rearrangements. In metabonomics data, it is possible for two different individuals to have very similar or identical metabolic profiles due to factors such as diet, medication, or environmental exposures. Therefore, while doppelganger effects are more common in biomedical data, they are not unique to this domain.

Avoiding doppelganger effects in machine learning models

To avoid doppelganger effects in machine learning models for healthcare, several strategies can be employed. First, feature selection can be used to identify the most informative features in the data and eliminate redundant or irrelevant features that may contribute to doppelganger effects. Second, model-based approaches such as deep learning and Bayesian methods can be used to build models that can learn complex relationships between features and avoid overfitting. Third, adversarial learning can be used to generate synthetic data that can expose and mitigate doppelganger effects.

Interesting examples from other data types

One example of the doppelgänger effect in imaging data is the classification of malignant versus benign lesions in breast cancer. Several studies have shown that machine learning models trained on imaging data can be confounded by the presence of benign lesions that have similar imaging features to malignant ones. As a result, the models may misclassify malignant lesions as benign, leading to delayed diagnosis and treatment.

Another example of the doppelgänger effect in gene sequencing data

is the identification of driver mutations in cancer. Driver mutations are mutations that promote the development of cancer, and their identification is critical for developing effective targeted therapies. However, many driver mutations are rare and occur in patients with different genetic backgrounds, making it challenging to identify them accurately.

Conclusion

In conclusion, the doppelgänger effect can have significant implications for the development and application of machine learning models in health and medical science. While there is no one-size-fits-all solution for avoiding or checking for the doppelgänger effect, researchers can take steps to ensure that their datasets are representative, use sophisticated machine learning algorithms, perform sensitivity analyses, and consider the potential ethical implications of their models. By doing so, they can help ensure that their models are accurate, reliable, and effective for all subgroups of the population.