Understanding and Avoiding the Doppelganger Effect in Machine Learning Models

Introduction

Machine learning (ML) has become increasingly important in various fields, including health and medical science, as it can help identify potential drug candidates and predict disease outcomes (Li Rong Wang et al., 2022). In particular, ML models have been widely adopted in drug development for faster identification of potential targets. However, the reliability of such models can be affected by the presence of data doppelgängers, which occur when independently derived data are very similar to each other, causing models to perform well regardless of how they are trained (Li Rong Wang et al., 2022). This issue is not unique to biomedical data, but rather a general issue that can arise in any type of data, including imaging (Rathgeb et al., 2021), and gene sequencing (Waldron et al., 2016), among others. Despite the abundance of data doppelgängers and their confounding effects, they remain poorly understood and are not always identified before model training and evaluation.

This report aims to examine the confounding effects of data doppelgängers on machine learning models in biomedical data. It will explore the prevalence of doppelgänger effects and provide an understanding of how they arise, as well as recommendations for avoiding or checking for them in the development and practice of machine learning models in health and medical science. Additionally, the report will examine examples from other fields to demonstrate how doppelgänger effects can arise in different types of data, providing a comprehensive understanding of this phenomenon.

Prevalence of Data Doppelgängers

Data doppelgängers can cause significant problems in machine learning models trained on biomedical data. These models can perform well, regardless of how they are trained, due to the presence of samples that are very similar to each other, known as data doppelgängers. These functional doppelgängers generate a doppelgänger effect, which confounds ML outcomes. Data doppelgängers have been observed in many areas of modern bioinformatics, including chromatin interaction prediction systems, protein function prediction, and quantitative structure-activity relationship (QSAR) models in drug discovery.

Data doppelgängers are not limited to biomedical data, as they can be observed in other data types as well. In face recognition, data doppelgängers can occur when faces are highly similar, making it difficult for machine learning models to differentiate between individuals. For example, identical twins may have nearly indistinguishable facial features, creating doppelgänger effects that can impact the accuracy of face recognition systems (Rathgeb et al., 2021).

Similarly, in gene sequencing, data doppelgängers can arise when gene or genome sequences are highly comparable, leading to poor model generalization. For instance, cancer genomics research often involves the sequencing of tumor samples from different patients with the same cancer type. In such cases, the mutations in the genes or genomes of these patients can be nearly identical, making it difficult for machine learning models to distinguish between them (Waldron et al., 2016).

In fraud detection, data doppelgängers can arise when legitimate and fraudulent transactions have similar patterns or characteristics, making it hard for machine learning models to accurately distinguish between them. For instance, fraudulent transactions may mimic legitimate ones by using similar transaction amounts, locations, or purchase types. This can lead to a high rate of false positives, where legitimate transactions are flagged as fraudulent, or false negatives, where fraudulent transactions go undetected (Marks, 2020).

Therefore, understanding and avoiding data doppelgängers is a critical issue in the development and application of machine learning models, particularly in the field of biomedical data. In the next section, we will delve deeper into how data doppelgängers arise and the potential effects they can have on machine learning models.

Emergence of Doppelgänger Effects from a Quantitative Angle

Further understanding the emergence of doppelgänger effects from a quantitative angle, it is vital to consider the underlying factors that contribute to their occurrence. One such factor is the similarity of the data samples used in model training, which can lead to overfitting and poor generalization performance (Kong et al., 2022). Additionally, the size of the dataset can impact the emergence of doppelgänger effects, as smaller datasets are more prone to the presence of highly similar samples (L. R. Wang et al., 2022).

To address these issues, the given paper suggests identifying data doppelgängers before the training-validation split to mitigate the doppelgänger effect. However, detecting data doppelgängers can be challenging, and methods such as ordination and embedding methods coupled with scatterplots may not always be effective as data doppelgängers are not always distinguishable in reduced-dimensional space. Another proposed method is to use the pairwise Pearson's correlation coefficient (PPCC) to identify data doppelgängers. This metric calculates the linear relationship between pairs of samples, with a high PPCC value indicating a high degree of similarity between the two samples. While the PPCC can be a useful tool for identifying data doppelgängers, it is not always sufficient to determine their functional impact on the model's performance. Further research is needed to establish a more direct link between the presence of data doppelgängers and their functional impact on the model's performance (Li Rong Wang et al., 2022).

Avoiding or Checking

As mentioned earlier, one of the main challenges with identifying doppelgänger effects is that they can arise due to chance or other factors that make independently derived data very similar to each other. This can occur in various types of data, not just biomedical data.

To avoid doppelgänger effects in machine learning models for health and medical science, there are several approaches that can be taken. One approach is to carefully evaluate the data to identify potential doppelgängers before training and validation, which can be done using various methods such as ordination, embedding methods coupled with scatterplots, or by comparing MD5 fingerprints of samples. Data stratification is also important to ensure that the model is performing well across different subgroups of the data. Additionally, performing robust independent validation checks with multiple datasets can help ensure the objectivity and generalizability of the model despite the possible presence of doppelgängers in the training set (Li Rong Wang et al., 2022).

Another approach that could be explored is identifying functional doppelgängers directly (Li Rong Wang et al., 2022). This can be done by looking for subsets of a validation set that are predicted correctly regardless of the ML method used and then

avoiding these subsets during model evaluation. In addition to these approaches, researchers can consider using alternative techniques, such as ensemble methods, to reduce the impact of doppelgänger effects by combining multiple models that are trained on different subsets of the data (Kong et al., 2022). The domain-specific knowledge can be used to guide the development of machine learning models, which can help identify potential sources of data doppelgängers and provide insights into the characteristics of the data that should be considered when building the model.

Furthermore, data augmentation techniques can be used to generate additional training samples that are diverse and representative of the target population, which can help reduce the impact of data doppelgängers by increasing the variability of the data (Li et al., 2020). Data augmentation techniques can also help reduce the risk of overfitting, which is another issue that can arise in the development of machine learning models. While data doppelgängers can be problematic for model development, they can also be used to improve the efficiency and effectiveness of model development by generating synthetic data that is representative of the target population (Al Amin et al., 2023).

Conclusion

In conclusion, doppelgänger effects are a significant challenge in the development and application of machine learning models for health and medical science. While they can arise in various types of data, they are particularly prevalent in biomedical data due to the nature of the data and the high degree of similarity that can occur between samples. To avoid doppelgänger effects, it is essential to carefully evaluate the data and perform data stratification, as well as explore new methods for identifying functional doppelgängers directly. Taking these steps can help ensure that machine learning models are properly trained and validated and can be used to effectively identify potential drug candidates and other important discoveries in health and medical science.

References

Al Amin, M. A., Shetty, S., Formicola, V., & Otto, M. (2023). Assessing the quality of differentially private synthetic data for intrusion detection. *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, 473–490. https://doi.org/10.1007/978-3-031-25538-0_25

Kong, W., Hui, H. W. H., Peng, H., & Goh, W. W. B. (2022). Dealing with missing values in proteomics data. *PROTEOMICS*, *22*(23-24), 2200092. https://doi.org/https://doi.org/10.1002/pmic.202200092

Li, Y., Hu, G., Wang, Y., Hospedales, T., Robertson, N. M., & Yang, Y. (2020). Differentiable Automatic Data Augmentation. *Computer Vision – ECCV 2020*, 580–595. https://doi.org/10.1007/978-3-030-58542-6_35

Marks, P. (2020). Dark web's doppelgängers aim to dupe antifraud systems. *Commun. ACM*, *63*(2), 16–18. https://doi.org/10.1145/3374878

Rathgeb, C., Drozdowski, P., Obel, M., Dorsch, A., Stockhardt, F., Haryanto, N. E., Bernardo, K., & Busch, C. (2021). Impact of doppelgängers on face recognition: Database and evaluation. *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*. https://doi.org/10.1109/biosig52210.2021.9548306

Waldron, L., Riester, M., Ramos, M., Parmigiani, G., & Birrer, M. (2016). The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles. *J Natl Cancer Inst*, *108*(11). https://doi.org/10.1093/jnci/djw146

Wang, L. R., Choy, X. Y., & Goh, W. W. B. (2022). Doppelgänger spotting in biomedical gene expression data. *iScience*, *25*(8), 104788. https://doi.org/10.1016/j.isci.2022.104788

Wang, L. R., Wong, L., & Goh, W. W. (2022). How doppelgänger effects in biomedical data confound machine learning. *Drug Discovery Today*, *27*(3), 678–685. https://doi.org/10.1016/j.drudis.2021.10.017