Question 3:

Doppelganger Effects in Biomedical Data that Confounds Machine Learning

Huang Jiayao

1. Introduction

The use of machine learning models in the biomedical industry has grown in popularity, with applications including disease screening and diagnostics, drug development and discovery, and patient risk assessment. However, the performance and reliability of these models are heavily dependent on the quality and consistency of the data used to train and test them. One potential issue that could compromise the reliability of machine learning models is the "doppelganger effect" [1], which occurs when there is a high similarity between independently derived training and test data, leading to models performing well regardless of how the model is trained [1].

The phenomenon of the doppelganger effect can negatively impact the performance of machine learning models in biomedical applications in various ways. For example, it could lead to overfitting, which occurs when the model is overly fitted to similar data points from both the training and test sets, leading to a poor generalization of new and unseen data [2]. Furthermore, when the data contains similarities between different groups of samples, the model may be biased towards specific groups, resulting in inaccurate or unreliable predictions [3]. This could lead to reduced robustness of the model, making it less able to perform well in certain scenarios. Therefore, it is crucial to take appropriate measures to mitigate the doppelganger effect in biomedical data to improve machine learning models' performance. Failure to take doppelganger effects into consideration could lead to serious consequences, especially in the biomedical field, where the models are used to make important decisions about diagnosis and patient care.

This report aims to delve into the doppelganger effect, specifically in biomedical data, and will also discuss how these doppelganger effects emerge from a quantitative angle. The report also proposes effective ways to check and avoid this effect in the practice and development of machine learning models for health and medical science.

2. Doppelganger Effects in Biomedical Data

2.1. Whether Doppelganger Effects are Unique in Biomedical Data

Doppelganger effects are not unique to biomedical data and can occur in any dataset that contains duplicate or highly similar data points between the training and test sets. An example would be in image recognition, in which the doppelganger effect can occur when the training and test data contain duplicate or highly similar images, leading to a high probability of false matches in the recognition system [4]. This has been studied

in the field of face recognition, where it has been found that a high similarity in the images collected leads to a significant increase in the probability of false matches [4]. This doppelganger effect also applies to recognizing other types of images, when the training and test data contain similar data, thus inflating the machine learning model's performance.

Even though doppelganger effect may also occur in other types of data, biomedical data may be particularly susceptible to doppelganger effects due to the abundance of confounding similarities in the data. These similarities can arise from a variety of sources, such as problems in assessment methodologies, proteins with similar sequences, similar chromosomes, RNA families, or shared ancestry [5]. This high degree of similarity in the data could cause biasness, which leads to inaccurate or unreliable predictions. This is particularly problematic in the field of medical research, where the goal is to develop models that can accurately diagnose diseases or aid in drug development.

2.2. Example of Doppelganger Effects in Different Biomedical Data Types

Doppelganger effects could also occur in different types of biomedical data, such as imaging and gene sequencing data.

In medical imaging data, doppelganger effects can occur when the images have similar characteristics, such as the subject matter being imaged, pose, background, noise, lighting, and colour, as medical images are usually taken based on standard imaging protocols [6]. For example, in cardiology, images of the heart may have similar views and lighting conditions, while in ophthalmology, images of the retina may have similar background and noise levels [6]. These similarities may potentially lead to data doppelgängers in the imaging data, thus causing inaccuracies in machine learning models trained to identify diseases. Furthermore, in the case of standard imaging modalities like computed tomography (CT), magnetic resonance imaging (MRI) and ultrasound (US), consecutive images frames may be captured. If similar images exist in both the training and test datasets, doppelganger effects may occur, which could compromise the model performance. To reduce the duplication of imaging data, Chinn et al. has proposed a method to eliminate redundancy by utilizing a similarity metric [6].

Doppelganger Effects in gene sequencing can occur when the same population is used for both training and testing, or when the data collection processes for the two datasets are not sufficiently independent. A study on the deep learning model for RNA secondary structure prediction, by Szikszai et al., has demonstrated the potential of the model to predict the RNA secondary structure within the same RNA family, but a poor performance when generalizing across families [7]. This is a good example of the doppelganger effects in gene sequencing data, as the training and test datasets were from the same RNA family, resulting in a very high similarity between the two datasets. Due to this high similarity, prediction performance is only good for data within the same

family, with poor inter-family performance [7]. When new data from other RNA families are introduced into the model, prediction performance may decrease due to overfitting.

3. Identification of Data Doppelgangers and Their Confounding Effects

In the study by Wang et al., several ways of identifying data doppelgangers were discussed. One of the most effective methods is to measure the Pairwise Pearson's Correlation Coefficient (PPCC), which captures the relationships between sample pairs of different data sets. A high PPCC would indicate the presence of PPCC data doppelgängers. To identify potential functional doppelgangers from the PPCC data doppelgängers identified, constructed benchmark scenarios were used [1].

The study also provides evidence from a quantitative angle of the confounding effects of PPCC data doppelgängers on machine learning performance. Different machine learning models were trained with the data that contains doppelgängers, including K-Nearest Neighbours (KNN), Naïve Bayes, Decision Tree, and logistic regression. The study has found that data doppelgängers identified by PPCC were present in both training and validation datasets, even with randomly selected features, and they could inflate machine learning performance. This effect is consistent across different datasets and different machine learning models. Furthermore, the study also found a dosage-based relationship between the number of doppelgängers and the magnitude of the effect, as the more doppelgänger pairs represented in both training and validation sets, the more inflated the ML performance is. The study confirms that PPCC data doppelgängers act as functional doppelgängers, as all machine learning models showed higher performance on PPCC data doppelgängers than on non-PPCC data doppelgängers. Moreover, it is also found that different machine learning models are affected differently by data doppelgängers. For instance, k-Nearest Neighbour (kNN) and Naive Bayes models show a linear relationship between performance inflation and doppelganger dosage compared to the decision tree and logistic regression models [1].

4. Methods of Checking & Avoiding Doppelganger Effects in Machine Learning

The Doppelganger effects in machine learning could be checked by independent validation. The model could be evaluated on a diverse range of data from different sources and data types to assess its generalizability and objectivity. A more comprehensive assessment of the model's performance could be achieved using a diverse set of data and features. This would also provide insight into how well the model will perform on unseen data and suggest whether there might be potential data doppelganger confounding the model [1].

To avoid the doppelganger effects, several methods have been discussed by Wang et al. [1]. One way is to place all PPCC doppelgängers in the training set. However, this method may not be practical as the models may not generalize well due to a lack of knowledge.

Another method is to remove PPCC data doppelgängers identified by the package doppelgangR [8]. By directly removing the data doppelgängers, their confounding effects on the machine learning models would be significantly reduced. However, this approach is not practical if the datasets are too small with a high proportion of PPCC data doppelgängers, as directly removing them could significantly reduce the sample size. The study also attempted to alleviate the doppelganger effect with methods that would not significantly reduce sample size or require a large amount of contextual data, but with no success [1].

Based on the recommendations provided by Wang et al., this report proposes two methods to mitigate the impact of the doppelganger effect. The first method proposed is Meta-data-guided sample selection. PPCC can be used to identify potential doppelgangers in the meta-data. During the sample selection process, all the potential doppelgangers could be assorted into either training or validation sets. This would significantly reduce the similarity between training and test data, effectively preventing doppelgänger effects in machine learning models [1].

Another proposed method is data stratification based on similarity. This approach involves first identifying similarities in the data by PPCC. Then, the data is divided into different strata based on their similarities, such as PPCC data doppelgängers and non-PPCC data doppelgängers. This is followed by a separate evaluation of the performance of the machine learning model. Assuming that each stratum corresponds with a known proportion of the real-world population, a more realistic assessment of the model's performance could be achieved. This would also provide an understanding of how the model performs on data with different levels of similarity. More importantly, potential weakness areas of the models that need improvement could also be identified [1].

5. Conclusion

In conclusion, doppelganger effects can confound the performance of machine learning models in different types of biomedical data. However, it is important to note that doppelganger effects are not unique to biomedical data and can also occur in other data types. Many methods could be utilized to identify data doppelgangers and reduce their confounding effect on machine learning. However, the doppelganger effect still cannot be eliminated entirely. Future research could focus on identifying new methods for detecting and mitigating doppelgänger effects that do not rely heavily on meta-data. In addition, novel feature engineering and normalization approaches could also be explored.

6. References

[1] L. R. Wang, X. Y. Choy, and W. W. Goh, “Doppelgänger spotting in biomedical gene expression data,” *iScience*, vol. 25, no. 8, p. 104788, 2022.

[2] “What is overfitting?,” *IBM.* [Online]. Available: https://www.ibm.com/topics/overfitting. [Accessed: 16-Jan-2023].

[3] E. Shimron, J. I. Tamir, K. Wang, and M. Lustig, “Implicit data crimes: Machine learning bias arising from misuse of public data,” *Proceedings of the National Academy of Sciences*, vol. 119, no. 13, 2022.

[4] C. Rathgeb et al., "Impact of Doppelgängers on Face Recognition: Database and Evaluation," *2021 International Conference of the Biometrics Special Interest Group (BIOSIG*), Darmstadt, Germany, 2021, pp. 1-4.

[5] L. R. Wang, X. Y. Choy, and W. W. Goh, “Doppelgänger spotting in biomedical gene expression data,” *iScience*, vol. 25, no. 8, p. 104788, 2022.

[6] E. Chinn, R. Arora, R. Arnaout, and R. Arnaout, “Enrich: Exploiting image similarity to maximize efficient machine learning in medical imaging,” 2021.

[7] M. Szikszai, M. Wise, A. Datta, M. Ward, and D. H. Mathews, “Deep learning models for RNA secondary structure prediction (probably) do not generalize across families,” *Bioinformatics*, vol. 38, no. 16, pp. 3892–3899, 2022.

[8] L. Waldron, M. Riester, M. Ramos, G. Parmigiani, and M. Birrer, “The doppelgänger effect: Hidden duplicates in databases of transcriptome profiles,” *Journal of the National Cancer Institute*, vol. 108, no. 11, 2016.