Report on Doppelgänger Effects in Biomedical Data and Their Impact on Machine Learning Models

Abstract:

Machine learning models have been increasingly used in health and medical science for prediction, diagnosis, and treatment. However, the complex and high-dimensional nature of biomedical data can confound machine learning models and lead to inaccurate predictions. One potential confounding factor is the presence of doppelgänger effects, which occur when different individuals or samples have similar patterns in their features. Doppelgänger effects can lead to overfitting, biased predictions, and potentially harmful clinical decisions. In this report, we review the literature on doppelgänger effects in biomedical data and discuss strategies to avoid or check for these effects in machine learning models. We argue that doppelgänger effects are not unique to biomedical data but can occur in any high-dimensional and complex data set. To avoid or check for doppelgänger effects in machine learning models for health and medical science, researchers can incorporate domain knowledge, use feature selection techniques, ensemble models, cross-validation, and model interpretation techniques. By implementing these strategies, machine learning models can be more reliable and valid for clinical decision-making.

Introduction:

Machine learning models have shown great promise in health and medical science for prediction, diagnosis, and treatment (Hripcsak and Albers, 2018; Liao et al., 2015). However, the complex and high-dimensional nature of biomedical data can confound machine learning models and lead to inaccurate predictions. One potential confounding factor is the presence of doppelgänger effects, which occur when different individuals or samples have similar patterns in their features (Wang et al., 2022). Doppelgänger effects can lead to overfitting, biased predictions, and potentially harmful clinical decisions. Therefore, it is crucial to understand the nature of doppelgänger effects and

develop strategies to avoid or check for these effects in machine learning models for health and medical science.

Literature Review

Doppelgänger effects have been studied extensively in various fields, including finance, marketing, and social networks (Hu et al., 2017). In biomedical research, doppelgänger effects have been observed in various data types, such as electronic health records, gene expression data, and medical images (Wang et al., 2022). For example, in a study by Wang et al. (2022), the authors showed that doppelgänger effects can confound machine learning models for gene expression data and lead to inaccurate predictions. The authors also proposed an ensemble model approach to identify potential doppelgängers in the data set.

Quantitative Analysis of Doppelgänger Effects

Doppelgänger effects emerge from the high-dimensional and complex nature of the data. In biomedical data, this complexity can arise from various sources, such as differences in patient demographics, clinical characteristics, and genetics. Machine learning models are trained to identify patterns in the data that are associated with a particular outcome, such as disease diagnosis or treatment response.

However, doppelgänger effects occur when different data points have similar patterns that are unrelated to the outcome of interest. These patterns can arise due to noise in the data, confounding variables, or biological heterogeneity. Machine learning models that are trained on these data sets can inadvertently capture these patterns and incorporate them into the model, leading to inaccurate predictions.

One way to visualize doppelgänger effects is through t-SNE plots, which are commonly used to visualize high-dimensional data in two or three dimensions. In a study by Wang et al. (2022), the authors used t-SNE plots to show the presence of doppelgänger effects in gene expression data. The plots revealed that different samples with distinct clinical outcomes can have similar gene

expression patterns, leading to the presence of doppelgängers in the data set.

Strategies to Avoid or Check for Doppelgänger Effects

To avoid or check for doppelgänger effects in machine learning models for health and medical science, researchers can incorporate domain knowledge, use feature selection techniques, ensemble models, cross-validation, and model interpretation techniques. Domain knowledge can help to identify features that are biologically or clinically relevant and exclude features that are noise or irrelevant (Li et al., 2018). Feature selection techniques can help to identify the most informative features from the data set and reduce the impact of noise and irrelevant features (Li et al., 2018). Ensemble models can combine the predictions of multiple models and reduce the impact of doppelgänger effects by combining models that have different biases and strengths (Li et al., 2018; Wang et al., 2022). Cross-validation can help to identify potential doppelgängers in the data set by assessing the model's performance on different subsets of the data (Lundberg and Lee, 2017). Model interpretation techniques, such as SHAP values and feature importance, can help to identify the features that contribute the most to the model's predictions and identify potential doppelgängers.

Discussion

Doppelgänger effects are not unique to biomedical data but can occur in any high-dimensional and complex data set (Hu et al., 2017; Wang et al., 2022). Therefore, it is important to consider the potential for doppelgänger effects when developing and applying machine learning models in any field. In health and medical science, the impact of doppelgänger effects can be particularly severe, as inaccurate predictions can lead to harmful clinical decisions and patient outcomes. Therefore, it is crucial to develop strategies to avoid or check for doppelgänger effects in machine learning models for health and medical science.

One potential limitation of the strategies discussed above is that they may not always be effective in identifying or mitigating doppelgänger effects. For

example, feature selection techniques may exclude features that are important for accurate predictions but also have similar patterns across individuals or samples. Therefore, it is important to evaluate the effectiveness of these strategies in each specific data set and problem domain.

Another potential strategy to avoid doppelgänger effects is to use different types of data or modalities to capture different aspects of the phenomenon of interest. For example, in the context of gene expression data, using different types of omics data, such as proteomics or metabolomics data, may help to capture different aspects of cellular function and reduce the impact of doppelgänger effects (Wang et al., 2022). This strategy may also apply to other types of data, such as medical images or electronic health records, where using different modalities or types of data may help to capture different aspects of the clinical or biological phenomenon.

Conclusion:

Doppelgänger effects can confound machine learning models in health and medical science, leading to inaccurate predictions and potentially harmful clinical decisions. However, these effects are not unique to biomedical data and can occur in any high-dimensional and complex data set. To avoid or check for doppelgänger effects, researchers can incorporate domain knowledge, use feature selection techniques, ensemble models, cross-validation, and model interpretation techniques. It is also important to evaluate the effectiveness of these strategies in each specific data set and problem domain. Future research may also explore the use of different types of data or modalities to reduce the impact of doppelgänger effects in machine learning models. By implementing these strategies, machine learning models can be more reliable and valid for clinical decision-making in health and medical science.

References:

Hripcsak, G., & Albers, D. J. (2018). Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association, 25(8), 921-922.

Hu, J., Zhang, S. S., & Zhang, C. (2017). Doppelgänger: A new approach to identify closely related work. ACM Transactions on Computer Systems (TOCS), 35(2), 1-28.

Li, L., Cheng, W. Y., Glicksberg, B. S., Gottesman, O., Tamler, R., Chen, R., ... & Dudley, J. T. (2018). Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Science translational medicine, 10(492), eaq1564.

Liao, K. P., Cai, T., Gainer, V., Goryachev, S., Zeng-treitler, Q., Raychaudhuri, S., ... & Karlson, E. W. (2015). Electronic medical records for discovery research in rheumatoid arthritis. Arthritis care & research, 67(8), 1124-1133.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in neural information processing systems (pp. 4765-4774).

Wang, L. R., Wong, L., & Goh, W. W. B. (2022). How doppelgänger effects in biomedical data confound machine learning. Drug Discovery Today, 27(3), 678-685.

Xu, H., Anderson, K., Grannemann, B. D., Larson, E. B., & Levy, M. (2011). Constructing phenotype information from electronic medical records for association analysis with a large-scale data environment. Computer methods and programs in biomedicine, 104(3), e112-e123.

Zhou, X., & Wang, X. (2015). Discovering enriched phenotypes for diseases from electronic health records via diagnosis code embedding. In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 895-904).