Doppelganger Effects

Introduction

In contemporary society, machine learning (ML) models have been increasingly used in various kinds of fields, such as biomedical data, finance, speech recognition, computer vision and autonomous vehicles. They are efficient and can help people deal with complex data that cannot be handled by other traditional methods. These complex data, such as high-dimensional data, appear more frequently in the biological field, which has contributed to an increased interest in machine learning models in this area. Several prominent instances include drug discovery, RNA secondary structure, and protein structure (Wang, Choy and Goh, 2022). In machine learning, it is widely accepted that the training and test data sets should be generated independently for evaluating the performance of a classifier.

Nevertheless, this validation method is sometimes unreliable due to the existence of data doppelgangers. Data doppelgangers happen when two sets of data that were separately derived are strikingly similar to one another, leading to models functioning well regardless of how they are trained, and this phenomenon is called doppelganger effects (DEs) (Wang, Wong and Goh, 2021). Doppelganger effects (DEs) happen when samples have unintentional similarities that, when divided between the training and validation sets, boost the performance of the trained machine learning (ML) model. This inflationary effect leads to misleading confidence in the deployable nature of the model and finally a lack of reliability in the predictions it makes (Wang, Choy and Goh, 2022).

Due to the nature of data in the biomedical field, doppelganger effects seem to occur more frequently in this domain. Whether this means that doppelganger effects are unique to the biomedical field will be clarified in the Section 2 of this report. Meanwhile, since it has some negative effects on machine learning models, some approaches to avoiding doppelganger effects will be suggested in Section 3.

Doppelganger Effects' Uniqueness

From the review of the literature, I do not think doppelganger effects are unique to biomedical data; instead, it is unique to biological data. In terms of the definition of doppelganger effects, doppelganger effects occur when the training and validation sets are highly similar by chance or otherwise. This feature is not limited to biomedical data. The next three examples will illustrate that the effects caused by similarity occur elsewhere in the biological domain.

First, Cao and Fullwood (2019) conducted a thorough analysis of the chromatin interaction prediction tools already in use. Their research showed that the reported performance of these systems had been overestimated due to issues with the assessment methodology used. These systems are specifically assessed on test sets that are highly similar to training sets. Second, in a protein function prediction model, proteins with similar sequences are considered to be derived from the same ancestral protein, and therefore both are considered to inherit the function of that ancestor. The model generated in this case has the sequence of the proteins as the independent variable and the function of the proteins as the dependent variable in the training sets, so that the independent variables in both training and validation sets are similar and the function of the proteins in both training and validation sets are similar, leading to a fairly high prediction accuracy. In fact, there are proteins that do not have similar sequences but have similar functions, in which case predictions using the previously obtained model would give incorrect results (Friedberg, 2006). Third, in a RNA structure prediction model (Szikszai et al., 2022), the sequence of the RNA is the independent variable and the structure of the RNA is the dependent variable in the training sets. The rest of this example and the problems that arise are similar to the previous protein example.

Additionally, in areas other than biology, the data can to some extent be artificially filtered and adjusted to avoid a high degree of similarity between the training and validation sets, thus avoiding doppelganger effects. Therefore, doppelganger effects are not unique to biomedical data, but unique to the larger category of biological data that encompasses biomedical data.

Ways to avoid Doppelganger Effects

Since doppelganger effects are a concern because it might overstate how well the machine learning model performs on actual data and could potentially muddle model selection procedures that are exclusively based on validation accuracy, it is necessary to propose some methods to avoid doppelganger effects. There are several measures for identifying data doppelgangers, such as the pairwise Pearson's correlation coefficient (PPCC) (Waldron et al., 2016) and batch correction (Johnson, Li and Rabinovic, 2007). Simultaneously, there is another way to solve doppelganger effects, which is to adapt and improve machine learning models in its practice and development. By referring to the relevant literature, I have come up with one approach. The test data are first divided into two categories, data doppelgangers and non-data doppelgangers, using the previously mentioned method of identifying data doppelgangers. Since data doppelgangers are potential doppelganger effects generators, we only evaluate the performance of the model in the non-data doppelgangers category. Next, we place the doppelgangers identified in the training and validation sets into the training set, build the model and measure the model accuracy first, and then into the validation set, build the model and evaluate it. Finally, we compare the two models and choose the better one.

Conclusion

Machine learning models' performance is evaluated according to the accuracy of the models by using validation data sets. This method is usually correct if the data of the training sets and the data of the validation sets are independent of each other; moreover, this prerequisite always seems to be true or met by default. However, when doppelganger effects are present, this widely accepted assumption may be no longer true. In addition to this, it is clear from the literature that doppelganger effects appear to be common in validation data and can have an exaggerated impact on the accuracy of machine learning models, leading to negative effects on the models and results. Therefore, it is significant to find reasonable ways to identify and avoid doppelganger effects.

References

Cao, F. and Fullwood, M. J. (2019). Inflated performance measures in enhancer–promoter interaction-prediction methods. *Nature genetics*, 51(8), 1196-1198.

Friedberg, I. (2006). Automated protein function prediction—the genomic challenge. *Briefings in bioinformatics*, 7(3), 225-242.

Johnson, W. E., Li, C. and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1), 118-127.

Szikszai, M., Wise, M. J., Datta, A., Ward, M. and Mathews, D. (2022). Deep learning models for RNA secondary structure prediction (probably) do not generalise across families. *bioRxiv*.

Waldron, L., Riester, M., Ramos, M., Parmigiani, G. and Birrer, M. (2016). The Doppelgänger effect: Hidden duplicates in databases of transcriptome profiles. *JNCI: Journal of the National Cancer Institute*, 108(11).

Wang, L. R., Choy, X. Y. and Goh, W. W. B. (2022). Doppelganger spotting in biomedical gene expression data. *Iscience*, 25(8), 104788.

Wang, L. R., Wong, L. and Goh, W. W. B. (2021). How doppelganger effects in biomedical data confound machine learning. *Drug Discovery Today*.