Data Doppelganger Effect

Abstract

Date doppelganger occurs when independently derived data are similar to each other, resulting in models that perform well no matter how they are trained, this is a phenomenon known as the data doppelgänger effect. Data doppelganger effect can affect the reliability of machine learning models. With the large amount of data doppelganger in biological data and the increasing use of machine learning models for drug development, the impact of data doppelganger effects on biomedicine is evident. In earlier studies, some solutions to avoid data doppelgangers have been proposed, which have some drawbacks but provide directions for future research.

Introduction

Machine learning models are used extensively in drug discovery to speed up drug development as a way to raise the speed of drug discovery. In machine learning, when evaluating the performance of a classifier, the training and testing datasets should be derived independently. Yet, separately derived training and testing sets could still produce unreliable validation results. When the model is trained in the presence of data doppelganger, then it can incorrectly perform well. (Wang et al., 2022) However, there are times when data doppelganger does not cause the data doppelganger effect, and Wang et al. referred to data doppelganger that does induce data doppelganger effect as functional doppelganger. (Wang et al., 2022) Functional doppelganger is present in many studies, and it is then particularly important to attempt to avoid data doppelganger effect.

Examples of Data Doppelganger Prevalent in Biomedical Data

Data doppelganger is very common in biomedical data. In the study by Waldron et al. (2016), they examined the databases of ovarian, breast, bladder and colorectal cancers of cell lines and assesses their accuracy against a ‘gold standard’ of duplicate samples generated by means of further manual

Table 1.

Overview of confirmed doppelgängers in all studies*

Dataset identifier by type of cancer	Total No. samples	No. of doppelgängers	Institutional source of doppelgängers
Bladder			
GSE1827, GSE13507, GSE31189, GSE31684, GSE37317, PMID: 17099711	570	0	Various, no doppelgängers identified
GSE19915, GSE32894	490	84	University Hospital of Lund, Sweden
GSE89, GSE5287	70	2	Aarhus University Hospital, Denmark
Breast			
MAINZ, NKI, VDX	881	0	Various, no doppelgängers identified
TRANSBIG, UNT, UPP	586	78	Uppsala County, Sweden

Table 1 Hidden duplicates in databases of transcriptome profiles

examination of expression data, clinical annotations, and sample identifiers. In more than half of all their studies, doppelgangers were confirmed (part of the results are shown in Table1). (Waldron et al., 2016) However, Wang et al. (2022) realized on re-analysis of the Waldron et al. (2016) data that the doppelgangers they reported were the result of leakage, meaning samples are duplicated. Therefore these data might not constitute true doppelganger effect. But the fundamental technique used by Waldron et al. (2016) to identify data doppelganger is considered reasonable by Wang et al. (2022) and I would focus on this model (PPCC) in the next section. In addition, Cao and Fullwood (2019) conducted a detailed evaluation of existing chromatin interaction prediction systems. Significantly, these systems were evaluated on a test set with a high degree of similarity to the training set. Their work shows that the performance of these systems is exaggerated because of the problematic evaluation methods used in reporting them. (Cao & Fullwood, 2019) Furthermore, Goh and Wong also observed the existence of data doppelganger in the way that

certain validation data were guaranteed to perform well under a given training data, despite the chosen features being random. (Goh & Wong, 2009)

With these studies, it is readily apparent that the data doppelganger effect is ubiquitous and influential in biomedical data. If the data doppelganger effect is not avoided properly, then the training of learning machine models would tend to be ineffective, making it impossible to identify true biological activity.

Identifying Data Doppelganger and Possible Ways to Avoid It

Given the potential for confusion caused by the doppelganger effect, it is crucial to be able to identify the existence of data doppelganger between the training and validation sets prior to validation in order to avoid the data doppelganger effect. In their study, Wang et al. (2022) point out that the ordination methods and embedding methods, coupled with scatter plots to identify data doppelganger are not feasible because data doppelganger are not always discernable in the reduced dimensional space. (Wang et al., 2022) They also questioned the reasonableness of the method dupChecker because of the leakage issues, which resulted in dupChecker not being able to detect true data doppelganger, that is independently derived samples that are incidentally similar. (Wang et al., 2022) Finally, another measure is the pairwise Pearson's correlation coefficient (PPCC) mentioned in the last section, which captures the relationship between pairs of samples from different datasets. (Waldron et al., 2016) An exceptionally high PPCC value indicates that a pair of samples constitutes data doppelganger. Although PPCC proposed by Waldron et al. has some limitations in that it never ultimately relates the PPCC data doppelganger to its ability to confound the machine learning task, that is it is not possible to determine whether there is a functional data doppelganger. Nevertheless, Wang et al. (2022) argue that PPCC is methodologically sound as a

basic design for quantitative measurement. (Wang et al., 2022) Wang et al. analyzed whether PPCC data doppelgangers act as a functional doppelganger, that is having a significant inflationary effect on machine learning performance, by identifying the effect of PPCC data doppelganger in RCC (Guo et al.'s (2015) renal cell carcinoma proteomics data) on the validation accuracy of different random classifiers. (Wang et al., 2022) The results of the study reflected the prediction of Wang et al. that good accuracy is easily obtained when there are many similar examples, without ensuring generalizability to less similar examples. However, if there are few similar examples, gaps in the model are exposed, and therefore the model tends to perform poorly. PPCC data doppelganger could act as functional doppelganger.

Wang et al. (2022) argued that direct removal of data doppelganger to avoid the data doppelganger effect is difficult to achieve, one reason being that removing data doppelganger might result in too little experimental data to use. (Wang et al., 2022) There possible suggestions were made by Wang et al., the first suggestion is to use metadata as a guide for careful cross-checking. (Wang et al., 2022) The second suggestion is to perform data stratification, that is layering the data into layers of different similarities, rather than assessing the model fitness of the entire test data. (Wang et al., 2022) The third suggestion is to perform extremely robust independent validation checks involving as many datasets as possible. (Wang et al., 2022)

Conclusion

The performance of machine learning models is susceptible to data doppelganger effects that can invalidate validation data. It is particularly important to avoid data doppelganger effects, however, they could not be easily solved, and the optimal approach is to check for functional doppelganger in the data before classifying the training and test data. More experiments need to be investigated in the future to find more effective ways to avoid data doppelganger effect.

Reference

Cao, F., & Fullwood, M. J. (2019, July 22). *Inflated performance measures in enhancer–promoter interaction-prediction methods*. Nature News. Retrieved December 25, 2022, from https://www.nature.com/articles/s41588-019-0434-7#citeas

Goh, W. W. B., & Wong, L. (2009, January). *Turning straw into gold: Building robustness into gene signature inference*. Drug discovery today. Retrieved December 25, 2022, from https://pubmed.ncbi.nlm.nih.gov/30081096/

Guo, T., Kouvonen, P., Koh, C. C., Gillet, L. C., Wolski, W. E., Röst, H. L., Rosenberger, G., Collins, B. C., Blum, L. C., Gillessen, S., Joerger, M., Jochum, W., & Aebersold, R. (2015, March 2). *Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps*. Nature News. Retrieved December 25, 2022, from https://www.nature.com/articles/nm.3807/

Waldron, L., Riester, M., Ramos, M., Parmigiani, G., & Birrer, M. (2016, July 5). *The doppelgänger effect: Hidden duplicates in databases of transcriptome profiles*. Journal of the National Cancer Institute. Retrieved December 25, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5241903/

Wang, L. R., Wong, L., & Goh, W. W. B. (2022, March). *How doppelgänger effects in biomedical data confound machine learning*. Drug discovery today. Retrieved December 24, 2022, from https://pubmed.ncbi.nlm.nih.gov/34743902/