Effects of the doppelganger effect on biomedical data on machine learning models

CHENGLE OUYANG

1 Abstract

In recent years, as machine learning models have been increasingly applied to drug development, doppelganger effects have become more common in biomedical data. The doppelganger effect occurs when samples exhibit accidental similarity. This will result in the model performing well no matter how trained. This report explains the relationship between doppelganger effects and biomedical data and suggests how to avoid and identify doppelganger effects in the practice and development of machine learning models in health and medicine. The report also gives an interesting example of cancer genomes.

2 Relationship between the doppelganger effect and biomedical data

ML(Machine Learning) models are increasingly being used in drug discovery to accelerate drug development. ML improves the efficiency of drug discovery in a number of ways: the ML model allows faster screening of better drug candidates (targets), thereby reducing the time spent on discovery and testing. They can also identify existing US Food and Drug Administration (FDA) approved drugs for other diseases (drug reuse), thus greatly reducing the cost of drug development. Classifiers (training models) have been used to predict novel drugs and possible adverse drug reactions.

These approaches have shown promise. A new anticancer drug candidate, EXS21546, was discovered by Exscientia's "Centaur Chemist" artificial intelligence (AI) platform and is currently in clinical trials (NCT04727138) [1]. Several ML identified drugs and drug combinations for the treatment of COVID-19 have also entered clinical trials, such as melatonin and torimiphan, NCT04531748 and Baritinib through a web-based approach [2], Recognition by Benevolent's knowledge Maps (NCT04373044 and NCT04401579) [3].

ML, it has been established that training and test data sets should be derived independently when evaluating the performance of classifiers. However, independently derived training and test sets can still produce unreliable validation results. For example, when the training set and the validation set are high, data duplication occurs, resulting in a model trained and validated on the data set that is likely to perform well regardless of the quality of the training.

Despite the large number of data dopes and their inflationary effects in biomedical data, they remain uncharacterized. However, this may not guarantee the doppelganger effect. Therefore, a data doppelganger that produces a doppelganger effect (confuses ML results) is called a functional doppelganger.

Abundant data doppelganger has been observed in modern bioinformatics. Cao and Fullwood conducted a detailed evaluation of existing chromatin interaction prediction systems [4]. Their work suggests that the performance of these systems is exaggerated because of problems with the evaluation methods used to report them. In particular, these systems are evaluated on a test set that is highly similar to the training set. Goh and Wong also observed the existence of data dopes, so that even if the selected characteristics are random, certain validation data can guarantee good performance given specific training data.

But I don't think the doppelganger effect is unique to biomedical data. For example, SSD neural network model is often used in deep learning of target detection. It is also possible for its test sets and training sets to be highly similar, resulting in the doppelgge effect. Especially in the object detection of small objects, it is easier to produce the doppelganger effect.

3 Methods to avoid the doppelganger effect

While directly removing doppelganger from data has proven difficult to achieve, we still need to prevent doppelganger effects as much as possible.

First, you can use metadata as a guide for careful cross-checking. Using this information from the metadata, we were able to identify potential doppelgams and classify them all into training or validation sets, effectively preventing doppelgams and allowing a relatively more objective assessment of ML performance. Similarly, technical repetition of the same sample should be treated similarly.

Second, perform data layering. Instead of evaluating the model performance of the entire test data, we can layer the data into different layers of similarity (for example, PPCC data clones and non-PPCC data clones, and evaluate the model performance of each layer separately).

The third recommendation is to perform extremely robust independent validation checks involving as many data sets as possible (divergent validation). Although not a direct hedge against the data doppelganger, different validation techniques can inform the objectivity of the classifier. It also informs the extensibility of the model (in terms of actual use), despite the possible existence of data doppelgads in the training set [5].

4 Methods to identify the doppelganger effect

Being able to identify data doppelgangers is a good start. Earlier research on similar issues also suggested measures to identify data doppelgangers. One method, dupChecker, identifies duplicate samples by comparing the MD5 fingerprint of its CEL file [6]. The same MD5 fingerprint indicates that the sample is duplicate (basically copied and therefore indicates a leak problem). As a result, dupChecker does not detect true data doppelganger, which is a random sample of a similar sample. Another measure is the pairwise Pearson correlation coefficient (PPCC), which captures the relationship between sample pairs of different data sets [7]. PPCC as the basic design of quantitative measurement is methodologically sound. Therefore, we use it to identify potential functional doppelganger (from the PPCC data doppelganger) in the built base scenario.

The presence of PPCC data dopes in training and validation data can exaggerate ML performance, even if the characteristics are randomly selected (and therefore meaningless; In other words, the model should perform poorly during validation). This finding was consistently repeatable across different training and validation datasets and across different ML models. In addition, the more doppelgams are represented in the training and validation sets, the more ML performance is exaggerated. This suggests that there is a dose-based relationship between the number of doppelgangers in PPCC data and the size of the doppelgangers effect. When all PPCC data doppelganger is placed together in the training set, the doppelganger effect is eliminated. This offers a possible way to avoid the doppelganger effect. However, limiting the PPCC data clone to the training or verification set is a suboptimal solution. In the former, when the size of the training set is fixed (therefore, each data clone included results in the exclusion of less similar samples from the training set), it leads to the possibility that the model may not generalize well because the model lacks knowledge. In the latter case, the doppelganger was either correctly predicted or incorrectly predicted.

In addition to Pearson correlation coefficient, other correlation indicators (such as Spearman rank correlation coefficient and Kendall rank correlation coefficient) can also be used to identify DD[8]. Spearman rank correlation coefficients are calculated by first ranking the values in each sample and then applying Pearson correlation coefficients to ranking variables. Spearman's rank correlation coefficient measures monotone relationships between samples and is more general than Pearson's correlation coefficient, which only measures linear relationships. The pairwise Spearman rank correlation coefficient (PSRCC) between the sample pairs of the two data sets was calculated using a recognition method similar to that of the PPCC. Kendall rank correlation coefficient, on the other hand, is based on the ranking order of variables in each sample. Like Spelman rank correlation coefficients, Kendall rank correlation coefficients also measure monotone relationships between samples and are more general than PPCC. A pairwise Kendall rank correlation coefficient (PKRCC) was calculated between the sample pairs of the two data sets using a recognition method similar to PPCC.

Another practical and interesting way to identify and verify the doppelgangereffect is to use the software suite doppelgangerIdentifier[9]. A software suite for doppelganger identification and verification that applies doppelganger identifiers to a variety of diseases and data types. The doppelgangerIdentifier R package allows users to easily identify PPCC DDS between and within data sets, and verify the impact of these detected PPCC DDS on the accuracy of ML model validation.

5 Interesting examples in other data types

Whole-genome analysis of cancer specimens is commonplace, whereas the cancer transcriptome undergoes changes that are very unique but more difficult to summarize in form uniquely identifiable. Researchers often share or reuse specimens for later studies. If not detected, the duplicate expression profile in the public database will affect the re-analysis, which also produces the doppelganger effect. Hidden duplicates, if not discovered, may exaggerate the statistical significance or apparent accuracy of genomic models when combining data from different studies. Levi Waldron et al. proposed a commonly practiced method [8] to accurately match duplicate cancer transcriptome even samples analyzed by different microarray techniques or microarray and RNA sequencing when sequence data at the nucleotide level is not available. Using transcripts identifiers available in both datasets, batch correction, Pearson correlation coefficients (PCCS) between each sample in one dataset and each sample in another dataset, and duplication-oriented outlier detection.

Reference

[1] Savage, N., 2020. Tapping into the drug discovery potential of AI. *Nature. com. https://www. nature. com/articles/d43747-021-00045-7*.

[2] Cheng, F., Rao, S. and Mehra, R., 2020. COVID-19 treatment: Combining anti-inflammatory and antiviral therapeutics using a network-based approach. *Cleveland Clinic Journal of Medicine*.

[3] Richardson, P., Griffin, I., Tucker, C., Smith, D., Oechsle, O., Phelan, A., Rawling, M., Savory, E. and Stebbing, J., 2020. Baricitinib as potential treatment for 2019-nCoV acute respiratory disease. *Lancet (London, England)*, *395*(10223), p.e30.

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

[5] Wang, L.R., Wong, L. and Goh, W.W.B., 2021. How doppelgänger effects in biomedical data confound machine learning. *Drug Discovery Today*.

[6] Sheng, Q., Shyr, Y. and Chen, X., 2014. DupChecker: a bioconductor package for checking high-throughput genomic data redundancy in meta-analysis. *BMC bioinformatics*, *15*(1), pp.1-3.

[7] 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).

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

[9] 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).