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

CHENGLE OUYANG

1 Abstract

Doppelganger effects have increased in frequency in biological data over the past several years as machine learning algorithms have been used more and more in medication discovery. The doppelganger effect happens when samples show unintentional resemblance. No matter how well-trained the model is, this will lead to good performance. The link between doppelganger effects and biological data is discussed in this study, along with tips on how to prevent and recognise them while using machine learning to construct health and medical models. Additionally, the paper provides a fascinating example of cancer genomes.

2 Relationship between the doppelganger effect and biomedical data

Medication research is increasingly using ML (Machine Learning) models to speed up drug development. In a variety of ways, ML increases the effectiveness of drug development. For example, the ML model enables quicker screening of superior drug candidates (targets), which shortens the time required for discovery and testing. Additionally, they can locate US Food and Drug Administration (FDA)-approved medications already in use for treating other illnesses (drug reuse), significantly lowering the price of developing new medications. Classifiers (training models) have been used to forecast the potential side effects of innovative medications.

These methods have demonstrated promise. Exscientia's "Centaur Chemist" artificial intelligence (AI) platform developed a novel anticancer medication candidate, EXS21546, which is presently undergoing clinical trials (NCT04727138) [1]. For the treatment of COVID-19, several ML-identified medications and medication combinations have also entered clinical trials. These medications include melatonin and torimiphan, NCT04531748 and baritinib using a web-based approach [2], Recognition by Benevolent's knowledge Maps (NCT04373044 and NCT04401579) [3,] and melatonin and torimiphan alone.

When assessing the effectiveness of classifiers, it has been established that training and test data sets should be generated independently. However,

findings from validation might still be erroneous when training and test sets are independently created. A model trained and validated on a data set, for instance, is likely to perform well regardless of the quality of the training if data duplication between the training set and the validation set is substantial. They are still uncharacterized despite the prevalence of data dopes and the inflationary effects they have on biological data. This might not, however, ensure the doppelganger effect. Therefore, a functional doppelganger is a data doppelganger that causes a doppelganger effect (confuses ML findings).

Modern bioinformatics has seen a lot of data doppelganger. A thorough analysis of current chromatin interaction prediction methods was carried out by Cao and Fullwood [4]. According to their research, there are issues with the assessment techniques utilised to report these systems' success, which causes it to be overstated. These systems are specifically assessed on a test set that closely resembles the training set. Goh and Wong also noted the existence of data dopes, which means that even though the characteristics are chosen at random, specific validation data can ensure successful performance in the presence of certain 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 doppelganger effect. Especially in the object detection of small objects, it is easier to produce the doppelganger effect.

3 Methods to avoid the doppelganger effect

We still need to minimise doppelganger impacts even if it has proven challenging to explicitly remove doppelganger from data.

First, you may carefully cross-check your work using metadata as a reference. We were able to identify probable doppelgams using the metadata and categorise them all into training or validation sets, essentially eliminating doppelgams and enabling a more objective evaluation of ML performance. The same should be done each technical repeat of the same sample.

Layering data comes next. We can layer the test data into several levels of similarity rather than assessing the model performance on the complete set of test data (for example, PPCC data clones and non-PPCC data clones, and evaluate the model performance of each layer separately).

The next piece of advice is to do as many diverse data sets of really thorough independent validation tests as you can (divergent validation). While not a direct defence against the data doppelganger, many validation strategies can help the

classifier be more objective. Despite the potential availability of data doppelgads in the training set, it also influences the model's extensibility (in terms of practical application) [5].

4 Methods to identify the doppelganger effect

A excellent place to start is being able to spot data doppelgangers. Earlier studies on related topics also recommended steps to spot data doppelgangers. By comparing the MD5 fingerprint of a sample's CEL file, one approach, dupChecker, finds duplicate samples [6]. The sample is a duplicate, as evidenced by the same MD5 fingerprint (basically copied and therefore indicates a leak problem). Because it is a random sampling of a comparable sample, dupChecker misses actual data doppelganger. The pairwise Pearson correlation coefficient (PPCC) is another metric that reflects the connection between sample pairs from various data sets [7]. Methodologically sound, the PPCC is a fundamental design for quantitative measurement. In order to find possible functional doppelgangers in the developed base (derived from PPCC data doppelgangers), we employ it. As a result, we employ it to find prospective functional doppelgangers (from the PPCC data doppelganger) in the created base scenario.

Even if the attributes are randomly chosen, the existence of PPCC data dopes in training and validation data might increase ML performance (and therefore meaningless; In other words, the model should perform poorly during validation). This result was consistently reproducible using several training and validation datasets as well as various ML models. Additionally, the exaggeration of ML performance increases with the proportion of doppelgams in the training and validation sets. This shows that the strength of the doppelgangers effect and the number of doppelgangers in the PPCC data are correlated dose-dependently. The doppelganger effect is abolished when all PPCC data doppelgangers are combined in the training set. This suggests a potential method for preventing the doppelganger effect. However, it is not the best method to restrict the PPCC data clone to the training or verification set. The former creates the potential that the model may not generalise effectively since the model lacks information when the size of the training set is fixed (i.e., each data clone added results in the exclusion of less comparable samples from the training set). The doppelganger was either properly or wrongly predicted in the latter instance.

Other correlation indicators, such as the Spearman rank correlation coefficient and the Kendall rank correlation coefficient, can also be used to detect DD in addition to the Pearson correlation coefficient[8]. In order to get the Spearman rank correlation coefficients, each sample's values are ranked before the ranking variables are given Pearson correlation coefficients. Unlike Pearson's

correlation coefficient, which only assesses linear correlations, Spearman's rank correlation coefficient measures monotone relationships between samples. A recognition technique similar to the PPCC was used to determine the pairwise Spearman rank correlation coefficient (PSRCC) between the sample pairs of the two data sets. The ranking order of the variables in each sample provides the basis for the Kendall rank correlation coefficient, on the other hand. Kendall rank correlation coefficients, which are more versatile than PPCC and assess monotone associations between samples, are similar to Spelman rank correlation coefficients in this regard. Using a recognition technique similar to PPCC, a pairwise Kendall rank correlation coefficient (PKRCC) was computed between the sample pairs of the two data sets.

Using the software package doppelganger Identifier[9] is another useful and intriguing technique to identify and confirm the doppelganger phenomenon. an application that uses doppelganger identities to different illnesses and data kinds as part of a software package for doppelganger identification and verification. Users may quickly find PPCC DDS between and within data sets using the doppelganger Identifier R programme, and they can confirm how these found PPCC DDS affect the validity of ML models.

5 Interesting examples in other data types

Cancer specimens' whole genomes are frequently analysed, however the cancer transcriptome experiences alterations that are quite distinctive but more challenging to condense into a form that is uniquely recognisable. For future investigations, researchers frequently exchange or reuse specimens. The duplicate expression profile in the public database will have an impact on the re-analysis, which also results in the doppelganger effect, if it is not recognised. When merging data from several research, hidden duplicates may, if not found, inflate the statistical significance or apparent accuracy of genetic models. When sequence information at the nucleotide level is not available, Levi Waldron et al. provided a widely used approach [8] to reliably match duplicate cancer transcripts even samples obtained by various microarray methods or microarray plus RNA sequencing. Utilising the transcript IDs present in both datasets, batch correction, Pearson correlation coefficients (PCCS) between every sample in one dataset and every sample in the other dataset, and duplication-oriented outlier identification

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).