Jiamin Ma

DOPPELGANGER EFFECTS IN BIOMEDICAL DATA

Introduction

The doppelganger effect has become increasingly prevalent in biomedical data due to the increased adoption of machine learning models (ML) that work to identify potential targets in drug development at rapid speeds. This report serves to examine the doppelganger effect in relation to biomedical data, elucidating whether these links are unique or otherwise. It also serves to explain the avoidance of doppelganger effects in particular relation to machine learning models for health and science, as well as explaining how doppelganger effects emerge from a quantitative angle. Methods of avoiding or checking for doppelganger effects are examined in the report's final section.

Doppelganger effects and biomedical data

Establishing the unique prevalence of the doppelganger effect in biomedical data requires an examination of the Machine Learning Models (ML) that have led to the discovery and duplication of various doppelgangers. ML has become responsible for finding the identified drug combinations for a multitude of biomedical disclosures. One prominent example in recent years is the combination of melatonin and toremifene that was found via ML as a clinical advancement regarding the novel coronavirus 2019 – or COVID 19 – which utilized a network-based approach (Cheng et al., 2020). Artificial Intelligence has also been utilized to succeed with BenevolentAI's knowledge graph which identified baricitinib (Richardson et al., 2020). Doppelgangers emerge to be fairly common in these ML datasets and often influence the subsequent phenotype analysis and drug identification, thus asserting the necessity for systems to check for doppelgangers before analyzing data. The resulting performance inflation greatly affects the potential validation of the data. The uncertainty of ML testing is further amplified by the fact that 'independently derived training and test sets could still yield unreliable validation results' (Wang et al., 2021).

Doppelgangers are particularly abundant within biomedical data, having been observed frequently in modern bioinformatics, taking on forms in drug discovery and the evaluation of existing drug combinations. The extent of current research has covered significant observation of the doppelganger effect, with notable cases including Cao and Fullwood's "detailed evaluation of existing chromatin interaction prediction systems", and the work of Goh and Wong' whereby certain validation data were guaranteed a good performance given a particular training data, even if the selected features were random' (Wang et al., 2021). Doppelganger effects are not unique to biomedical data, though it is here where they show their most profound usefulness. In fact, doppelganger effects are not unique to the medical field at all, but can be found throughout data sets wherever data collection and replication is common. Doppelganger effects are a phenomenon found in artificial intelligence data gathering scenarios, and can function as a confounding middle or end product. The growing use of algorithmic data to process casual relationships into statistical distributions, with these attempts at normalizing curves leads to increased doppelgangers within the data sets. However, the prolific nature of the doppelganger effect in biomedical data forms a certain statistical uniqueness, and it is indeed one of the most thoroughly examined forms of it.

How doppelganger effects can be avoided in the practice and development of machine learning models for health and medical science

In regard to formatively decreasing both the chance of and the avoidance of the doppelganger effect, there is still little quantitative research on either matter. Within the biomedical data sector, both well-trained and poorly trained ML models are prevalent, resulting in an exponential increase in the phenomena of doppelgangers. The identification of doppelgangers within biomedical datasets has been examined, but work towards a significant reduction in their occurrence has yet to be hypothesized to much extent. It should be noted that despite the abundance of doppelgangers in biomedical data, 'it is surprising that procedures for eliminating or minimizing similarity between test and training data still do not constitute standard practice before classifier evaluation' (Wang et al., 2021).

There are several methods of collecting and identifying doppelganger data, for example, dupChecker, which 'identifies duplicate samples by comparing the MD5 fingerprints of their CEL files' (Wang et al., 2021). Also in existence is the pairwise Pearson's correlation coefficient (PPCC), which 'captures relations between sample pairs of different data sets' (Wang et al., 2021). The necessity for developing more rigorous discovery methods that can appropriately integrate within ML models is essential for the elimination of doppelganger effects as a whole. It is clear that the solution lies not from prevention, but the discovery and elimination after the fact, as the empirical data provided on the formation of doppelganger effects, suggest that they will continue to occur despite well-trained ML models.

How doppelganger effects emerge from a quantitative angle

The formation of doppelgangers within biomedical data is relatively complex, but it ultimately involves the re-analysis of undetected duplicate expression profiles. These expression profiles are found on the micro level within datasets, and serve to confound and alter data if they are not detected and removed. The ongoing complexity of doppelgangers relies on their necessity to occasionally be used as required data in smaller datasets, which reveals a scientific conundrum in which doppelgangers must be avoided and discovered, but also analyzed before removal. The classical biomedical study involves frequent re-use of biomedical evidence such as tissue samples or drug profiles, and without these samples being adequately identified prior to the analysis, they will return to the analysis profile pool. These hidden duplicates 'if left undetected, can inflate statistical significance or apparent accuracy of genomic models when combining data from different studies' (Waldron et al., 2016).

The most common circumstantial environment in which doppelgangers arise is public datasets that contain alterations and propositions sourced from various external datasets, but that continue to be replicated and used in ongoing studies. For example, 'the Renal cell carcinoma (RCC) proteomics data of Guo et al. taken from NetProt software library' contains composite and replicate samples due to the prolific nature of the dataset (Wang et al., 2021). It includes a predominance of negative cases, uncompromising sample pairs, and permissible doppelgangers formed from the identified dataset. The 'validation accuracy for all properly trained models' within this particular dataset lies at 10%, suggesting that the prominence of doppelgangers within the data is significantly higher at the outset (Wang et al., 2021).

Methods of avoiding or checking for doppelganger effects

It is not surprising that data doppelgangers cause a significant upheaval of the biomedical data process, which leads to a necessity of appropriate discovery and avoidance methods. Enforcing doppelganger discovery in training or validation sets is unhelpful as a solution as it does not eliminate the issue, nor does it avoid doppelgangers previously utilized in data sets.

Best practice suggests that there should be 'more comprehensive and rigorous assessment strategies, based on the particular context of the data being analyzed' (Wang et al., 2021). The use of systems and programs such as dupChecker, doppelgangR, and PPCC is useful regarding large datasets, but their usefulness and appropriateness fall short when it comes to smaller datasets. The reduction of data to an unusable size by implementing PPCC renders the data wholly unhelpful; the doppelgangers are required to form a complete dataset.

Methods to reduce doppelgangers must not negate the size and value of the data. However, they must also adequately remove doppelgangers that will significantly negatively affect the data. Data trimming is one such method that could be used, which involves 'removing variables contributing strongly toward data doppelgangers effects' (Wang et al., 2021). However, due to the highly complex nature of the doppelganger effect, it is necessary to form a complex solution, and simply removing or reducing variables is not conducive to a productive dataset. The need becomes one of both removing or reducing variables and also discovering and eliminating the doppelgangers as they arise, but first analyzing whether the removal of the doppelgangers will negatively affect the subsequent dataset. The issue and solution are therefore both understandably complex. The solution resides in meta-data, which can be used to 'perform careful cross-checks' and to construct 'negative and positive cases' (Wang et al., 2021). The value of meta-data in the solution to the doppelganger effect cannot be understated, as it allows for substantial anticipation and much-needed analysis before removal without negating the value of the dataset itself. Doppelgangers can be appropriately sorted between training and validation datasets by using the information gathered from meta-data. The nature of meta-data is algorithmic, allowing for a more thorough and objective evaluation of ML performance.

Reference

Cheng, F., Rao, S., & Mehra, R. (2020). COVID-19 treatment: Combining anti-inflammatory and antiviral therapeutics using a network-based approach. *Cleveland Clinic Journal Of Medicine*. https://doi.org/10.3949/ccjm.87a.ccc037

Richardson, P., Griffin, I., Tucker, C., Smith, D., Oechsle, O., & Phelan, A. et al. (2020). Baricitinib as potential treatment for 2019-nCoV acute respiratory disease. *The Lancet*, *395*(10223), e30-e31. https://doi.org/10.1016/s0140-6736(20)30304-4

Waldron, L., Riester, M., Ramos, M., Parmigiani, G., & Birrer, M. (2016). The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles. *Journal Of The National Cancer Institute*, *108*(11), djw146. https://doi.org/10.1093/jnci/djw146

Wang, L., Wong, L., & Goh, W. (2021). How doppelgänger effects in biomedical data confound machine learning. *Drug Discovery Today*. https://doi.org/10.1016/j.drudis.2021.10.017