Junfan Chen

Review of the doppelgängers effect, consideration of the uniqueness of biomedical data and avoidance of the doppelgängers effect

Abstract In this report, according to the research of Li Rong Wang, Limsoon Wong, and Wilson Wen Bin Goh, we reviewed the widespread existence and discrimination of data doppelgängers, the impact of doppelgängers effect on machine learning models and existing solutions. At the same time, the uniqueness of doppelgängers effect on health and medical science data is discussed. Finally, a feasible scheme for the training process of the non-data doppelgängers model is proposed by analogy with the regularization idea in the logical regression method: model training is carried out by adding penalty factors and combining PPCC values to eliminate the doppelgängers effect in the training process.

1. Introduction

Machine learning methods are gradually being used in the field of drug research with faster speed as well as lower cost[1]. The existing application cases include EXS21546, a new anti-cancer candidate drug, which was discovered through Execientia's' Centaur Chemistry 'artificial intelligence (AI) platform and is being clinically tested(NCT04727138)[2]; the drug combination of melanin and torque discovered by network-based approaches3 (NCT04531748)[3].

For machine learning, the ideal condition for evaluating the performance of the model is that the training data set and the test data set are derived independently. However, even if the ideal assumption of an independently derived training set and test set is satisfied, there will still be similarity across their measurements (data doppelgängers) between the data, resulting in unreliable verification results (doppelgängers effect; confounding ML outcomes)[4]. This is the impact of data doppelgängers on ML models. That is, when model training and cross-validation are based on data doppelgängers, the quality of training will not affect the evaluation of model performance. Although Data Doppelgangers may not necessarily produce a doppelgängers effect, it is extremely necessary to investigate the nature of data doppelgängers, the extent to which the doppelgängers effect affects the ML model and the improvement methods [5]. This report will discuss the uniqueness of doppelganger effects in biomedical data by reviewing existing materials through literature research, discuss the causes of this phenomenon from a quantitative perspective, and propose feasible measures to avoid doppelganger effects in combination with existing personal data science knowledge.

2. The widespread existence and identification of data doppelgängers

Data doppelgangers are a problem that exists widely but lacks in-depth discussion. In one notable case, Cao and Fullwood performed a detailed evaluation of existing chromatin interaction prediction systems[6]. In their work, they found that data doppelgangers advocated the model performance. At the same time, in bioinformatics, especially in protein function prediction, for proteins with similar sequences and functions, the existence of data doppelgängers does not attract attention. However, for the prediction of proteins with different sequences but the same functions, such as two-zone homologs[7], and enzymes that are different in sequence overall but with similar active site resistances[8], the accuracy of prediction cannot be guaranteed. Quantitative structure–activity relationship (QSAR) models are classification and regression ML models trained to predict the biological activities of molecules from their structural properties[9]. And the model assumes that molecules with similar structures will have the same functions, which is true in most cases. But for data doppelgängers, we can only find problems by comparing good and bad models. We can't judge performance simply by a poorly trained model that uses similar data. Therefore, it is necessary to identify data doppelgängers in time to ensure the accuracy of ML model training.

The article of Li Rong Wang et al., first discussed the feasibility of identifying data doppelgängers according to data distribution (order methods (e.g., principal component analysis) or embedding methods (e.g., t-SNE), coupled with scatterplots).[5] However, the data distribution after dimension reduction by the PCA method still has no specificity and cannot be effectively identified. At the same time, dupChecker, identifying duplicate samples by comparing the MD5 fingerprints of their CEL files[10] is not an effective means to identify data doppelgängers because only the leakage situation can be considered.

However, Li Rong Wang et al. also presented a feasible scheme, [5] namely, pair Pearson's correlation coefficient (PPCC),[11] and discussed the significance and rationality of PPCC values in characterizing data doppelgängers, as well as the limitations of being unable to establish the relationship between data doppelgangers and their impact on ML performance. By using the renal cell carcinoma (RCC) proteomics data of Guo et al.,[12] we can quantify the universality of data doppelgängers (half of the samples are PPCC data doppelgängers with at least one other sample).

3. Discussion on the uniqueness of doppelgängers effect

From the perspective of the data generation approach, I think the doppelgängers effect is unique in biomedical data since biomedical data can produce data doppelgängers due to the similarity of the action mechanism, after meeting the premise of the mutual independence of the training data set and the test data set derivation path. However, due to the use of the data doppelgängers training model, the doppelgängers effect appears. As mentioned above, the similarity of protein transcription and translation leads to structural similarity, which makes the model predict the functions of proteins with similar structures and functions with the doppelgängers effect. In fact, the model performance is not the same. Such doppelgängers effect caused by the similarity of biological action mechanism is a feature that no other field has.

However, for the application of the ML method, the doppelgängers effect is not the only biomedical data. One of the most representative applications is face recognition. In the research of Evgeny Smirnov et al.,[13] we can use the doppelgängers effect (referred to as "doppelgänger mining" in the article) to sample similar-looking identities ("avatars") together to generate better small batches and insert avatar mining into the process of face representation learning, which significantly improves the discrimination ability to learn face representation. Similarly, if the ML method is applied to the research of brand doppelgängers[14] data in the future, there will also be a doppelgängers effect caused by human factors (commercial activities). Therefore, from the perspective of ML research, the doppelgängers effect is not unique to biomedical data.

4. Confounding effects of PPCC data doppelgängers and amelioration of it

According to the research of Li Rong Wang et al., it can be known that there is an obvious dosage-effect between the number of PPCC data doppelgängers and the significance of the doppelgängers effect, leading to the emergence of an inflationary effect.[5] Especially in KNN and the naive Bayesian model, there is an approximately linear relationship: when there are many data doppelgängers, the higher prediction accuracy of the model will be easier to obtain. However, this result actually lacks the generalization of less similar data. Moreover, when all data doppelgängers are put into the training set, it will lead to insufficient data required for model training and poor results; when data doppelgängers are constrained in either the training set or verification set, it will lead to a winner-take-all situation. The above two schemes are not what we need.

Similarly, according to existing research, there are several investigated methods to improve the doppelgängers effect: by dividing training and test data based on a single chromosome (rather than taking all chromosomes together), and using different cell types to generate training evaluation pairs, to establish good field practices/standards, but this method is difficult to complete due to the lack of good benchmark data; use PPCC outlier detection packet, doppelgangR, to remove PPCC data doppelgängers to mitigate its impact.[15,16] However, this method does not work on PPCC data with a high proportion of small data sets, such as RCC, because deleting PPCC data doppelgängers will reduce the size of unavailable data, which makes this method infeasible.

Li Rong Wang et al. also proposed three methods to improve data doppelgängers[5]: using metadata for cross-check, but this method can only perform as an indicator of the existence of data doppelgängers, and cannot substantially solve the problem of doppelgängers effect in model training; data stratification, which is a feasible solution, can be based on data stratum, corresponding to different categories of groups in the real world (diseases, populations, etc.), and modelling and forecasting on each group, which has practical significance, but may be challenging to use for non-hierarchical data; the use of divergent data sets can exchange the expansion of data volume for the reduction of similarity, but it is not practical for small-scale data itself and may lead to an increase in computational expense.

5. Recommended feasible measures to solve the doppelgängers effect in the machine learning model for health and medical science

According to the above, we can identify data doppelgängers from a quantitative perspective. At the same time, PPCC is a specific indicator to refer to the similarity of data. Combined with the idea of regularization in logical regression, in the regression class classification prediction model, PPCC data can be normalized to obtain data samples with approximate 0-1 distribution (the concept of PPCC range of valid case has been proposed in the article, which can be used as the basis for normalization). In the data samples obtained, when the normalized PPCC value is closer to 0, it is considered that the similarity of data pairs is lower. On this basis, we can add a penalty factor (an infinite constant relative to the model training process; similar to the λ), and multiply the normalized PPCC value as an item of the model training cost function. At the same time, according to the idea that the optimal model has the minimum cost function in machine learning, the optimal model can be obtained only when the normalized PPCC value tends to zero. This idea can minimize the doppelgängers effect in the model training stage. Similarly, in the KNN model, we can also add a penalty term about the normalized PPCC value to the distance function, so that the condition of minimum distance needs to ensure the lowest similarity, and also can avoid the doppelgängers effect to some extent. However, such a method may ignore the practical significance of the functional similarity brought by the data doppelgängers themselves (for example, proteins with similar structures do come from the same ancestor and indeed have similar functions). Therefore, the penalty factor can achieve better training results only when it is used in clear data without a positive correlation of similarity.

Reference

1. Zhou, Y., Wang, F., Tang, J., Nussinov, R. & Cheng, F. Artificial intelligence in COVID-19 drug repurposing. *The Lancet Digital Health* **2**, e667–e676 (2020).
2. Savage, N. Tapping into the drug discovery potential of AI. *Biopharm Deal* d43747-021-00045–7 (2021) doi:10.1038/d43747-021-00045-7.
3. Cheng, F., Rao, S. & Mehra, R. COVID-19 treatment: Combining anti-inflammatory and antiviral therapeutics using a network-based approach. *CCJM* ccjom;ccjm.87a.ccc037v1 (2020) doi:10.3949/ccjm.87a.ccc037.
4. Ho, S. Y., Phua, K., Wong, L. & Bin Goh, W. W. Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability. *Patterns* **1**, 100129 (2020).
5. Wang, L. R., Wong, L. & Goh, W. W. B. How doppelgänger effects in biomedical data confound machine learning. *Drug Discovery Today* **27**, 678–685 (2022).
6. Cao, F. & Fullwood, M. J. Inflated performance measures in enhancer–promoter interaction-prediction methods. *Nat Genet* **51**, 1196–1198 (2019).
7. Wass, M. N. & Sternberg, M. J. E. ConFunc—functional annotation in the twilight zone. *Bioinformatics* **24**, 798–806 (2008).
8. Friedberg, I. Automated protein function prediction--the genomic challenge. *Briefings in Bioinformatics* **7**, 225–242 (2006).
9. Paul, D. *et al.* Artificial intelligence in drug discovery and development. *Drug Discovery Today* **26**, 80–93 (2021).
10. Sheng, Q., Shyr, Y. & Chen, X. DupChecker: a bioconductor package for checking high-throughput genomic data redundancy in meta-analysis. *BMC Bioinformatics* **15**, 323 (2014).
11. Waldron, L., Riester, M., Ramos, M., Parmigiani, G. & Birrer, M. The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles. *JNCI J Natl Cancer Inst* **108**, djw146 (2016).
12. Guo, T. *et al.* Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. *Nat Med* **21**, 407–413 (2015).
13. Smirnov, E., Melnikov, A., Novoselov, S., Luckyanets, E. & Lavrentyeva, G. Doppelganger Mining for Face Representation Learning. in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)* 1916–1923 (IEEE, 2017). doi:10.1109/ICCVW.2017.226.
14. Ruvio, A., Gavish, Y. & Shoham, A. Consumer's doppelganger: A role model perspective on intentional consumer mimicry: Consumer's doppelganger. *J. Consumer Behav.* **12**, 60–69 (2013).
15. Lakiotaki, K., Vorniotakis, N., Tsagris, M., Georgakopoulos, G. & Tsamardinos, I. BioDataome: a collection of uniformly preprocessed and automatically annotated datasets for data-driven biology. *Database* **2018**, (2018).
16. Ma, S. *et al.* Continuity of transcriptomes among colorectal cancer subtypes based on meta-analysis. *Genome Biol* **19**, 142 (2018).