Doppelganger Effects in Biomedical Data Analysis

ZIQI ZHANG, C2338083

February, 2023

Abstract

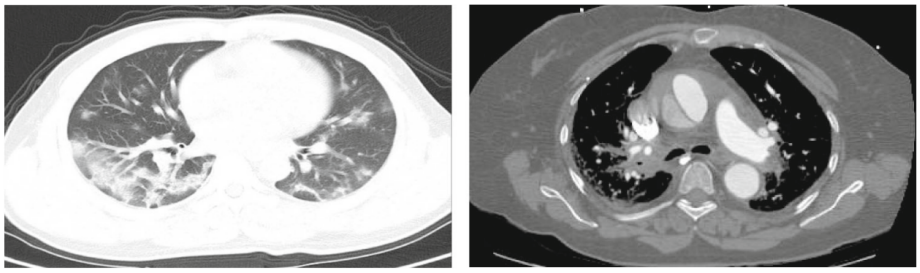
Biomedical data is an important research object in the data analysing field. Using various methods and building models in machine learning (ML), features and trends under the data could be revealed which may greatly deepen people's insight toward physiological processes and mechanisms of diseases. Doppelganger effect is a phenomenon widely existing in biomedical data analysis which may bring confounding results and increase the difficulty of biomedical research. At the same time, data dopperlganger (DD) is hard to be detected during the experiments. In this report, I'm going to discuss the prevalence and uniqueness of Doppelganger effects (DEs) in biomedical data and potential methods to avoid DEs in ML models for healthcare and medical science.

1 The Uniqueness of DEs

We have already known that DEs happen when biomedical data used in training sets and validation sets derived independently show a high degree of similarity. In ML, generation of data similarity always comes from data augmentation methods. It should be noted that this kind of similarity in DEs is different from data augmentation. Data augmentation is a useful way to enlarge the dataset in order to increase the accuracy of models. However, augmentation factors using here could be expressed as matrices and they are obviously not independently with the original data and could not lead DEs. This proved that traditional similarity cannot produce DEs. Thus, we should focus on the traits of biomedical data itself.

One of the most important features of biomedical data science is the concept of omics. Completely objective and unable to be modified, omics describes living things in different aspects and different levels. Thus, biomedical data could not get themselves out of the range of the omics. In other words, they have similarity to some extent regardless of the source of data. For example, proteins which have similar active site similar sequences may have similar functions[1]. Obviously, any two proteins are independent in production, so do their amino acid sequence. If such data appears in the train set and validation set in pair at this time, the model accuracy would be falsely high and lead DEs. Same situations will occur on other levels because of the omics and may also lead DEs.

By proving other similarities could not generate DEs while biomedical data could lead DEs on the premise of deriving independently, we prove the uniqueness of DEs in biomedical data.



(i) COVID-19 positive (ii) COVID-19 negative

Figure 1: Example figures of positive and negative COVID-19 CT images[3]

2 The Revalence of DEs

From the discussion in the early section, DEs in protein sequence have been shown. There are also other examples of DEs in various types of biomedical data which prove the relevance of DEs.

RNA sequence data shows DE in the experiment. Wang et al, 2022[2] used two RNA sequence data sets Haematopoietic and Lymphoid Tissue-Lung and Large Intestine-Upper Aerodigestive Tract in order to identify the DE. Just like what they had done before, they constructed three sample pair types including different classes from different patients (negative pair), same class from different patients (possible doppelganger data) and same class from same patient (data leakage) and calculated pairwise Pearson's correlation coefficient (PPCC) of each pair and drew scatterpoint figures. It's clear that some pairs in second type performance with high PPCC. The introduction of these data pairs actually increases the accuracy of the model and leads DE.

There are also DEs in the image type of biomedical data. In the past 2 years, the world was influenced by the covid-19 epidemic. To study the effect of the virus, scientists used images of chest from various kind of imaging method including CT and MRI. The chest CT image shows that the lung is affected by the virus and appears white flake density shadow (Figure 1). When using those images to train and validate a neural network, the network would no doubt to learn the features like shape, the color and the position of the lung and shadow points from images of different patients. It could be easily find that the characters do not have much difference between different patients no matter whether the patient is healthy or not. But all of them are derived independently. According to the given paper, the data pairs containing same class of lung from different patients may become DD and lead DEs.

Besides all of the examples above, there is still another fact that exacerbates the prevalence of DEs in biomedical data. That is the scarcity of the data. Constricted by ethics and technology, biomedical data is always hard to obtain. Thus, most of the analysis should depend on the public database. When combining data with other research, re-using of the data which leads hidden duplication inflates the accuracy of ML models[4]. At the same time, biomedical data contain mass spatiotemporal information. Data from the same patient at different times could also leads this kind of similarity.

3 The Emergence and Potential Methods to Avoid DEs

In the ML model, data would be transformed into vectors for training and testing. To evaluate the similarity of the data, the norm of the vectors is always being considered. This idea is often used in clustering analysis. Euclidean Distance, which is 2-norm of the vector, is a distance people like to use in the ML model training by the K-nearest Neighbour method. In the given paper, accuracy of the model increases linear when more DDs are used in the KNN model[1]. Also, accuracy in the Naive Bayes models also shows this kind of trait. The Naive Bayes method considers characters independently with each other and that is the trait of DEs which are led by the similarity of independent data[5].

The given paper[1] and its follow-up work[2][6] detailed discussed using PPCC as the parameter to check whether the data pairs are DD and cause DEs. What makes the study significant is that it detects the DD before the training process and packages the functions in a toolbox using R language. I suppose that other parameters could also be calculated and evaluate the similarity and detect DDs before using them. For sequence data, norm is a good choice as the criterion.

$$d = \|x\|^p$$

For image data, there also have some methods to evaluate the similarity. For example, the histogram distance provides information about the distribution of pixels which is useful in computational vision experiments.

Besides all of the potential parameters and coefficients which may be helpful in DEs detection, there is still another important thing that should be noticed. The origin of the DDs in biomedical data is because of the re-use of public databases. Actually, the regulation of those public data is insufficient. The sources of the data are also various. If there is any chance to build a system which could integrate different types of data and make detailed tags, scientists could therefore extract data in by avoiding hidden duplication to some extent. Patients‘ privacy and data security need to be considered here so it's not a easy task. But we can imagine the help the system will offer for building machine learning models in health and medical data research.

4 Conclusion

To summarized, DE is a unique phenomenon in biomedical data analysis by ML models. However, DEs are widely exist in the biomedical data in various type including medical image, gene sequence, protein sequence etc. because of their own characteristic (omics) so it's hard to totally get rid of the effect of DEs when training and valid the ML models. The author's team of given paper have done a lot in using PPCC to detect DDs in the data sets before the training and validation process of ML models building. There can be still some other parameters or coefficients like norm (defining distance as the level of similarity of data) and cosine coefficient. The similarity of image could also be reflated by histogram. Besides all those potential mathematical methods for avoiding or detecting the DDs in the datasets, we are looking forward to a system which

could manage the various types of data from different sources to reduce the number of hidden duplication of data in the samples.

References

[1] Wang, L.R., Wong, L., and Goh, W.W.B. (2021). How doppelga¨nger effects in biomedical data confound machine learning. Drug Discov. Today 27, 678–685. https://doi.org/10.1016/j.drudis.2021.10.017

[2] Wang, L.R., Choy, X.Y., and Goh, W.W.B. (2022). Doppelga nger spotting in biomedical gene expression data. iScience 25, 104788. https://doi.org/10.1016/j.isci.2022.104788

[3] Garain, A., Basu, A., Giampaolo, F. et al (2021). Detection of COVID-19 from CT scan images: A spiking neural network-based approach. Neural Comput Applic 33, 12591–12604. https://doi.org/10.1007/s00521-021-05910-1

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

[5] Zhou, Z. Machine Learning. Tsinghua University Press. 2016

[6] Wang, L.R., Fan, X., and Goh, W.W.B.(2022). Protocol to identify functional doppelgangers and verify biomedical gene expression data using doppelgangerIdentifier. STAR Protocols 3, 101783. https://doi.org/10.1016/j.xpro.2022.101783