Report on Doppelganger effects in biomedical data

Wang Ruining

January 12, 2022

Introduction

As many researchers start to apply machine learning methods in the biomedical area, huge successes have been achieved in drug discovery and development. ML models not only shortlist better drug categories more efficiently but also repurpose drugs to other diseases to reduce cost for designing new drugs[1].

Well trained classifier is the core of classification models in ML model because the training effect of classifiers directly determine the quality of model. When training ML classifier, the selection of dataset is critical. Inappropriate size of data sets, biased features contained in data sets may lead to unsatisfactory performance on validation or clinical circumstances. As a result, identification of the factors that influencing training effects negatively in ML model is in turn equally vital for system design. One of those factors is the doppelganger effect.

When classifier is trained by data set which has high similarity with the validation data set, it performs better compared with that trained by less similar data set, where the results is unreal because of the existence of data doppelganger. In my opinion, this phenomenon can be comparable to overfitting of the ML model, which shows almost perfect accuracy in data with similar features with training set data but terrible effects in generalization. The models trained by this way is obviously not the ideal result when applied to reality.

Occurrence of the Doppelganger Effects

When a classifier falsely performs well because of the data doppelgangers, no matter the classifier was trained enough or not, the accuracy on data doppelgangers is high. Data doppelganger is key to the misleading, which is currently known as functional doppelganger. Detection of functional doppelganger before training can reduce the doppelganger effects.

Demonstration on the occurrence of the Doppelganger Effects from a quantitative angle. Because the logical methods (*i.e.* PCA) for identification of data doppelgangers is unfeasible[1], measurable ones are quite essential in the detection

session. DupChecker compares MD5 fingerprints in their CEL files to distinguish duplicate samples[2], but it cannot tell when two data sets are occasionally measured similar in independent measurements. Another method is pairwise Pearson's correlation coefficient (PPCC)[3]. Let (x_i, y_j), $i = 1, 2, \ldots, n$, be the sample pairs. The correlation coefficient[4] between x and y is expressed as:

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

where x and y are sample means for x_i and y_i, respectively. The more the value of r_{xy} is, the higher the similarity between data pairs. If the data pairs are justified as PPCC data doppelgangers and are in train and valid data set separately, the inflationary effects occur.

Doppelganger effects are not unique to biomedical data. There normally exists this kind of phenomenon that doppelganger pairs affect the construction of mathematical model. In the field of economics, emotional branding brings 'Doppelganger Brand Image'[5], which may pose a threat to the culture resonance of the brand. When a brand cannot provide identity value to customers, crisis occurs. Thus, the effect in the economic area means danger of being identical with other brands and lose the specific emotional link with target group. What's more, this phenomenon can also be observed easily when huge amount of data is required in ML network or other mathematical analysis method because some repetitive data must lead to better performance in test.

Interesting Data Doppelganger examples in other data types. There are data doppelgangers of different types, such as images, gene sequences. If we take analysis method by L.R. Wang et al.[1] as standard for distinguishing data doppelgangers, the CT images in figure 1 can be viewed as a data doppelganger pair in my recent research.

The area I am currently research in is removal of artifacts in CT images due to metal implantation, and this set of images are from a publicly available data set—SpineWeb[6]. In this experiment, shape and size of metal artifact in images matter, thus metal implants in similar positions with semblable shapes may result in doppelganger effects. The shape and area of artifacts always links with physical shapes of metals, which is the dark areas around light white regions in CT images. These dark regions affect diagnosis and need to be removed effectively. Two images derived from different patients but with similar artifacts by coincidence constitute data doppelganger pair.

Although the visual contents in the images are not similar in general opinion, it may cause doppelganger effects when they are put into deep learning network as train and valid pair.

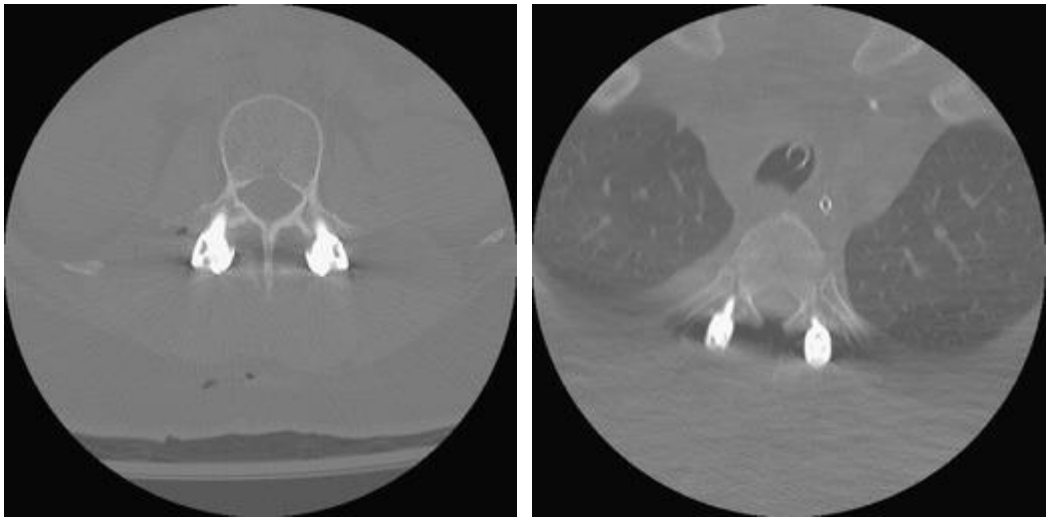

Fig. 1. Data Doppelganger pair in spine CT images

Moreover, in the natural language processing field, this effect is also considerable. When the input of training set comes from one person A and someone else with similar biological features with A provides samples in the validation set, Doppelganger Effect happens.

Ways to avoid Doppelganger Effects

Recent years, researchers pay more attention to Doppelganger Effects and proposed some methods to avoid data doppelgangers and sequential effects brought by these data. There are several methods introduced:

- Cao and Fullwood[7] proposed a method based on contextual data. This approach separates training and valid data by considering individual chromosomes derived from different cells, which avoid the interference of data doppelganger. However, this requires sufficient prediction of prior knowledge and benchmark knowledge, and thus making it difficult to realize.
- The second assumption is the removal of data doppelgangers. By PPCC calculational identification of these data, it can be achieved to remove them in one data set. But this reduces the data size and may even lead to unbalanced data sets or just too small size of data in some data sets with large proportions of data doppelgangers.
- There are many variables in one sample of data, and L.R. Wang et al.[1] attempted to extract related variables from data doppelgangers so that they can be used in both training and valid sets again. But what they found was after removing those variables, the effects also exist. This verifies the similarities lie not only in the variable level.

Proposed approaches to the Doppelganger Effects

Data doppelgangers appear and influence the evaluation of model, so avoiding negative effects cause by these data is important. A direct method is to divide valid data into several groups (more than ten groups), and test separately. Once the results are obtained, two with best performance are removed and take the average score of the remaining groups. But this method only considers better performance as doppelganger effects and is less rigorous.

Another proposed approach is the detection of data doppelgangers. We can train a machine learning network to identify the level of similarity in functional doppelgangers. Two coders constitute this network. The coders produce input data pair into hidden layer, which are represented as two sequences of codes of this pair, separately. And the second coder then generates predicted assemble value of this pair of data, higher score means more possibility of being data doppelganger. By feeding with known data doppelgangers and accepted or set similarity values, data pairs can be checked. When score of one pair exceeds certain threshold, this will be viewed as data doppelganger pair.

Conclusions

As far as I am concerned, the reason why data doppelgangers interfere the evaluation effects is because of the way of the mechanism behind machine learning network. Take brain as an example. When we prepare and practice well on certain questions, it is easier for us to get high score on questions that are similar with our prepared ones. ML network simulates neural links in brain, so it tends to perform better on familiar area and knowledges. Overcoming this effect is vital in the area of ML and AI because this will provide possibility for the enhancement of evaluation in systems.

References

1. Li Rong Wang, Limsoon Wong, Wilson Wen Bin Goh, ‘How doppelgänger effects in biomedical data confound machine learning’, Drug Discovery Today, 2021, ISSN 1359-6446, https://doi.org/10.1016/j.drudis.2021.10.017.
2. Q. Sheng, Y. Shyr, X. Chen, DupChecker: a bioconductor package for checking high-throughput genomic data redundancy in meta-analysis, BMC Bioinform 15 (2014) 323.
3. L. Waldron, M. Riester, M. Ramos, G. Parmigiani, M. Birrer, The Doppelgänger effect: hidden duplicates in databases of transcriptome profiles, J Natl Cancer Inst 108 (2016) djw146.

4. Wu, B (Wu, Berlin), el. Innovative Correlation Coefficient Measurement with Fuzzy Data[J]. MATHEMATICAL PROBLEMS IN ENGINEERING, 2016, ISSN: 1024-123X.
5. Thompson, C. J., Rindfleisch, A. and Arsel, Z., "Emotional Branding and the Strategic Value of the Doppelgänger Brand Image," Journal of Marketing, Vol. 70, No. 1, 2006, pp. 50-64.
6. http://spineweb.digitalimaginggroup.ca/
7. Cho, Hyuk Jun, Kim, Sung-Geun, and Kang, Juyoung, “An Empirical Analysis of Doppelgänger Brand Image Effects: Focused on the Internet Community,” The Journal of Information Systems, vol. 26, no. 1, pp. 21–51, Mar. 2017.