Possible Reasons and Solutions to Data Doppelgängers Effects

Abstract

Data doppelgänger effects (DDE) is ubiquitous in biomedical data. It occurs when machine learning model performs falsely well in validation due to the high similarity between training and validation data. In industrial applications of machine learning, either on image, text, or structured data, it is also possible for data doppelgänger effects to occur due to the inconsistency between offline and online data. In this report, we analyse DDE and extract possible causes with case analyses. We propose that DDE is caused by inappropriate selection of samples, limitation of features, and the particularity of biomedical data. Possible and feasible solutions to tackle DDE are also proposed based on careful selection of samples, features, and models, where frameworks based on Learning-to-Rank (L2R) and Contrastive Learning are introduced to avoid DDE.

Introduction

The concept of data doppelgänger (DD) [1] is referred as similarity between independently derived training and validation data, and data doppelgänger effect (DDE) is used to describe its results where the evaluation of machine learning model is inaccurate, i.e., a good performance on validation data doesn't guarantee the same performance on unseen data (test data or real-world data). Several cases are proposed in [1] to demonstrate the pervasiveness of DDE in biomedical data. However, similar phenomena may also happen in real-world scenarios in industry.

Machine learning has been widely used in industrial scenarios, where the recommender system (RS) is one particularly important application. The core idea of RS is to recommend items, such as goods in e-commerce website and short-video in Tik-Tok, based on user history and behaviour. A normal working flow for machine learning engineers mainly contains two parts: offline model development and online deployment. However, due to the pipeline variability, the offline and online environments are likely to differ. The key point here is that offline data/features should be exactly the same as that of online, i.e., the label distribution, types of features should be rigorously identical to online environment. For instance, if a model is trained on a biased offline dataset where 70% of users are ladies, similar features extracted are intensively female-related. Thus, the model is likely to perform poorly in online environment as the actual gender ratio might be 50%:50%, since the model may tend to recommend items popular among females but may not popular among males. As a consequence, the model is biased due to data doppelgänger, as it may seem to be good in terms of accuracy during validating process. In fact, DDE here is originated from the biased offline dataset, resulting in a falsely positive measurement.

Case Analyses

In [1], multiple cases are mentioned particularly in bioinformatics. In the following part, we will conduct case analysis in order to figure out possible reasons for DDE and ways to solve it.

Case 1: Protein Function Prediction

In protein function prediction scenario, proteins with similar structures are assumed to contain similar functions, which is widely known as "structure determines function" [2]. Structure-function relationships arise through the process of natural selection, and it is quite common not only in nature but also in artifacts. For instance, the structure of hammer is specifically designed for its function to pound on something. Therefore, it is reasonable to set it as a prior-knowledge in protein function prediction task, and it is common that data with similar structures and similar functions broadly exist in real-world datasets.

However, a same function may still possible to be conducted by other proteins with different structures, which can lead to errors or limitations on machine learning models. Models trained on such dataset, where there are only similar proteins with similar structures, are likely to perform badly on unseen dataset, i.e., a dataset where proteins with similar functions but different structures broadly exist. This is due to the data-driven principle behind machine learning models. From machine learning models to deep learning models, regardless of the architecture, they are all extracting features from given samples for predictions on unseen data. In protein function prediction, the feature extracted on structure A is likely to be useful for structures similar to A but fails on structure B, even though proteins with structure B function similarly. The DDE occurs due to the false selection of samples; thus model fails to learn from proteins with different structures but similar functions.

Advice 1: Focus on Data Preparation and Feature Selection

The DDE on case 1 is inherently originated from lack of data diversity. It is strongly advised that model should learn from different samples as many as possible. In this case, more proteins with similar functions but different structures should be added to dataset. To step further, the best practice would be a miscellaneous of samples: similar function with similar structure, similar functions with different structure, different functions with similar structure, and different functions with different structure.

Besides, choosing features appropriately can help alleviate DDE. In case 1, the primal reason is the limited information of structure feature, so that it is not distinguishable by using only structure feature. Structure is no longer a deterministic feature since different structure can provide similar function. Therefore, more features that can well represent samples are encouraged to discover. In particular, concentrations should be on the theory and mechanism of why such different structure can perform similar function. Once we can find an essential and more discriminative feature other than structure, the problem can be solved.

The learning process of machine learning models is similar to how human-beings get to know the world. If swans are mainly white, then this creature might be assumed to be white, even if there are grey or black swans but in a small number. To tackle the issue, it is advised to form a prior knowledge with correct ratio of different colour of swans. Besides, another feature like region of birth may be helpful, since swans in region A are likely to be white whereas they tend to be black in region B. Based on our discussions above, it is also feasible to utilise models enhanced with Bayesian methods, since it seems related to prior and posterior, and certainly such framework can be useful. Due to the limitation of length, related discussion is omitted.

Case 2: Protein Activity Prediction

The small variation of structure can substantially impact binding affinity [3], thus for a well-trained model, in theoretical, can surely distinguish between the tiny differences. However, due to the situation in case 1 where training set and validation set is highly similar, the performance of model is inaccurately represented. Here, apart from reasons mentioned in case 1, the DDE is also caused by the sensitivity of protein activity to structure, i.e., the overall structure is highly similar but carries with different activities. As a consequence, we propose to use other framework to model the problem apart from traditional machine learning.

Advice 2: Focus on modelling and architecture

It is the tiny structural differences that causes discrepancy of activity, and it can be hard to learn such feature inherently, regardless of whether it is well-trained or not. Apart from specifically designing more advanced architectures to capture such differences, a more efficient way is to focus on the relative differences between molecules, where learning-to-rank (L2R) and contrastive learning can be helpful.

L2R [4] is a common approach for modelling ranking problems in information retrieval, particularly in search engine. In searching scenarios, there are massive documents to be ranked, and those ranked ahead will be returned to user when they use search engine with key words inputs. A popular approach in L2R is to construct pairwise sample in form of doc pair input: (doc1, doc2, label), where doc1 and doc2 are documents and label is in {+1, -1} indicating whether doc1 is more likely to be the one user wants. In this case, what model learns are not direct features from a single sample, but a relative difference between two samples. Based on L2R, we can find out which molecule is more likely to demonstrate a certain property or activity. For two proteins with different activities but similar structures, the model can directly learn the difference, so such tiny and imperceptible difference that can lead to entirely different activities are explicitly taken into consideration.

In addition, we shall further introduce contrastive learning, an emerging technique which has been widely focused recently. Contrastive learning belongs to self-supervised learning, and self-supervised learning is a certain type of unsupervised learning. Unsupervised learning is used when explicit label (supervised information) is not available, and a classic approach is clustering where there is no label involved. In comparison, self-supervised learning will use data itself as supervised information to learn representations from samples. Contrastive learning is based on distinguishing differences between samples to learn an appropriate representation of samples on another feature space, in which samples are much easier to be classified.

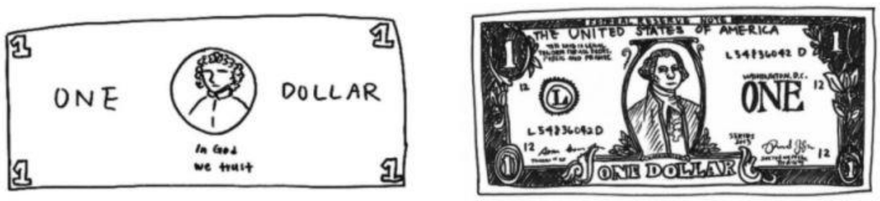

Fig 1. Drawing dollar from memory [5]

The core assumption of contrastive learning is the sufficiency of information from samples themselves. We can't duplicate a dollar even if we've seen dollars for endless times, but the left dollar in Fig 1 is enough to recognise that it's a dollar, thus we don't have to replicate it perfectly for recognition like the one in right part of Fig 1. That indicates: a representation learning algorithm doesn't have to learn all details of samples, but only till when the features are enough to distinguish from other samples. And this is the biggest difference between traditional representation learning methods, such as autoencoder (AE) where the target is to reconstruct the whole sample. Contrastive learning is based on the idea that representations are sufficient when samples can be well recognised if projected to the learnt representation space, rather than accurately re-construct the sample. Its goal is to learn an encoder *f* so that:

$$score\big(f(x), f(x^+)\big) \gg score(f(x), f(x^-))$$

where x^+ is positive sample similar to x, x^- is negative sample dissimilar to x, score is a metric function to measure the similarity between samples.

Based on contrastive learning, we can learn a model to extract feature from samples, representing the samples by projecting them into another feature space, and conduct classification/regression tasks based on the learnt feature space which can be more discriminative than the original feature space of samples.

It is worth mentioning that PCA or t-SNE are also viable approaches based on similar considerations, while contrastive learning here is a more advanced and efficient technique to employ.

Conclusion

Based on our discussion, we have found that there are mainly three possible reasons for occurrence of data doppelgänger effect (DDE): (1) inappropriate selection of samples, (2) limitation of feature expressiveness, and (3) particularity of biomedical data.

DDE is likely to happen when unitary samples are selected for training and validating but tested on diversified data. Therefore, DDE is not unique to biomedical data, but can be common in scenarios where datasets used are biased. Furthermore, we also proposed that limitation of feature can be one particular reason for DDE, where features such as structure no longer provide sufficient information to distinguish different samples. Thus, more expressive features are encouraged to sorted. The last but most important reason for DDE is the sensitivity of biomedical data, especially in case where tiny structure can lead to entirely different properties. To tackle this issue, we introduced L2R and Contrastive Learning framework, two new modelling approaches.

To avoid DDE, it is advised to (1) carefully select samples for dataset construction, (2) discover more distinguishable features, and (3) use frameworks such as L2R or Contrastive Learning to either explicitly learn consequences of tiny structural differences or represent samples in a more discriminative feature space.

Reference

[1] Wang, Li Rong, Limsoon Wong, and Wilson Wen Bin Goh. "How doppelgänger effects in biomedical data confound machine learning." Drug discovery today (2021).
[2] "Structure Determines Function (Interactive Tutorial)." Learnbiology, https://learn-biology.com/ap-biology/module-1-menu/structure-determines-function/.
[3] Chen, Qian, et al. "Enhanced QSAR model performance by integrating structural and gene expression information." Molecules 18.9 (2013): 10789-10801.
[4] Liu, Tie-Yan. "Learning to rank for information retrieval." (2011).
[5] Anand, Ankesh. Https://Ankeshanand.com/Blog/2020/01/26/Contrative-Self-Supervised-Learning.html. 26 Jan. 2020, https://ankeshanand.com/blog/2020/01/26/contrative-self-supervised-learning.html.