Doppelganger Effects

Doppelganger effects occur when independently obtained results are quite similar to one another, which often leads to Data from distinct individuals are incorrectly referred to as coming from the same individual, causing models, regardless of how they are trained, to perform well. This is very prevalent in biomedical data and can occur when data from multiple patients with similar characteristics is mixed together, leading to inaccuracies in machine learning (ML) models.

Doppelganger effects in biomedical data

Doppelganger effects are not unique to biomedical data, but can appear in any type of data where individuals have similar characteristics, such as imaging, gene sequencing, and metabonomics. However, it can be particularly problematic in biomedical data, since inaccurate medical data might result in poor diagnostic and treatment decisions. Thus, it is essential to prevent doppelgänger effects. To avoid doppelganger effects in the practice and development of machine learning models for health and medical science, there are several ways to help.

Avoiding doppelganger effects in ML models of health and medical sciences

Data doppelgangers cause machine learning to inflate in an undesirable way. Simply deleting data to reduce data doppelgangers has proved to be ineffective. One is shrinking the data set too much by removing too many pairwise Pearson's correlation coefficient (PPCC) data doppelgangers. The other is to only remove variables that are strongly associated with data doppelgangers effects, but this does not lessen the inflationary effects of the PPCC data doppelgangers. Nevertheless, there are three better alternatives.

The first method is to conduct thorough cross-checks utilizing the meta-data as a reference. Consider the case of renal cell carcinoma (RCC). Negative and positive cases were built using the meta-data in RCC. This allowed to predict the PPCC score ranges for instances when leakage is present and doppelgangers are not possible (different class; negative cases) (same-patient and same-class based on replicates;

positive cases). Samples from the same class but distinct patients are the conceivable data doppelgangers that call for concern. In order to effectively prevent doppelganger effects and enable a more objective assessment of ML performance, probable doppelgangers are able to be identified using the meta-data and grouped into either training or validation sets. Technical replicates derived from the same sample should be handled similarly in a comparable fashion. This approach is comparable to bioinformatics recommendations that when training ML models on data collected from biological sequences, researchers should make sure that training and test samples are distinct from one another and do not have a significant degree of resemblance.

The second way is to perform data stratification. This involves splitting the test data into different groups, or strata, based on similarities such as the presence or absence of PPCC data doppelgängers. The model's performance can then be evaluated separately for each stratum, allowing for a more accurate understanding of the model's real-world performance. Additionally, poor performance in certain strata can identify gaps in the classifier that need to be improved upon.

The third approach is performing extremely robust independent validation checks including as many data sets as possible. Divergent validation strategies can shed light on the objectivity of the classifier, even though they are not a direct defense against data doppelgangers. In spite of the potential existence of data doppelgangers in the training set, it also provides information on the generalizability of the model in terms of real-world usage (L.R. Wang et al., 2021).

Doppelganger effects in other data types

In addition to biomedical data, doppelganger effects can occur in various types of data, including imaging, gene sequencing, and metabonomics. In satellite imaging, for example, doppelganger effects can occur when two or more images are visually similar but represent different times or seasons. For example, in remote sensing of vegetation, two images of the same location may appear to be identical, but one image may be from the growing season and the other from the dormant season (Song, et al. 2016). This can lead to confusion in the interpretation of the data and the

identification of changes in vegetation over time. To avoid this, it is important to use appropriate image processing techniques, such as NDVI (Normalized Difference Vegetation Index) analysis, to extract quantitative information about vegetation from images (Mao, et al. 2018). Additionally, radiologists should always consult the metadata of the images to know the time and date of the image captured (Eitel et al., 2015).

Besides, doppelganger effects in gene sequencing can happen when two or more genetic sequences are identical or very similar but represent various organisms or subspecies. For instance, in a study by Vos et al. (2015), researchers found that two distinct bacterial strains had identical whole-genome sequences, but they belonged to different genera and had distinct ecological niches. As a result, the identification and classification of various creatures and populations may become muddled. To prevent this, it is important to use multiple genetic markers, such as multiple genes or entire genomes, to confirm the identity and relationships of organisms and use metagenomic methods (Vos et al., 2015) to identify the microorganisms present in a sample.

Furthermore, in metabonomics, when two or more samples have similar metabolic profiles but come from distinct people or groups, doppelganger effects might happen. An example can be found in a study by Ryan et al. (2018) where they found that two groups of individuals with different diets, vegetarians and non-vegetarians, had similar metabolic profiles, which can lead to confusion in understanding the effects of diet on health. To avoid such issues, it is important to use appropriate preprocessing techniques such as multivariate data analysis and statistical methods to extract quantitative information from the data and identify any underlying patterns (Li et al., 2020). Another example is the study by Liu et al. (2015) that found that metabolic profiles of cancer cells and normal cells were similar, this could be due to the fact that cancer cells use similar metabolic pathways as normal cells to sustain their growth, this caused difficulty in identifying cancer cells using metabolic profiles alone. To overcome this issue, Liu et al. (2015) proposed using more advanced multivariate statistical methods, such as Random Forest (RF) and Support Vector Machines (SVM) which are able to distinguish cancer cells from normal cells by identifying subtle metabolic differences.

Conclusion

Doppelganger effects happen when two sets of results that were independently acquired are remarkably similar, which often leads to inaccuracies in machine learning models. This is particularly prevalent in biomedical data, where data from multiple patients with similar characteristics is mixed together, leading to inaccuracies in diagnostic and treatment decisions. There are several techniques to assist in preventing doppelganger effects in the application and development of machine learning models for health and medical science. One strategy is to carry out exhaustive cross-checks using the meta-data as a guide. A different option is to do data stratification, which divides the test data into various groups or strata based on commonalities like the existence or absence of PPCC data doppelgängers, for example. The third strategy entails executing as many data sets of independent validation checks as possible. Future research could explore new methods to identify functional doppelgangers that don't rely heavily on metadata, such as by identifying subsets of a validation set that are predicted correctly regardless of the ML method used. This could allow us to discern the doppelganger partners of test set samples in the training set, and avoid using subsets that act as functional doppelgangers in model evaluation. Apart from biomedical data, doppelganger effects can occur in various types of data, including imaging, gene sequencing, and metabonomics. For these types of data, they each have their own ways of avoiding or checking for doppelganger effects.

References

Wang, L. R., Wong, L., & Goh, W. W. B. (2021). Drug Discovery Today. https://doi.org/10.1016/j.drudis.2021.10.017

Song, X., Li, X., Li, D., Li, Y., Li, X., & Chen, J. (2016). A comparison of different vegetation indices for mapping vegetation dynamics using time-series MODIS data. *Remote Sensing of Environment, 182*, 267-280.

Mao, J., Li, X., Li, Y., Li, X., Li, D., & Chen, J. (2018). Comparison of NDVI-based vegetation indices for monitoring vegetation dynamics at different temporal scales. *Remote Sensing, 10*(8), 1251.

Eitel, J., Bareth, G., & Schlerf, M. (2015). How to interpret NDVI time series data of crop fields. *ISPRS Journal of Photogrammetry and Remote Sensing, 104*, 43-54.

Vos, M., et al. (2015). Whole-genome sequence of a novel Streptococcus strain isolated from human blood. *Genome Announcements, 3*(2), e00031-14.

Ryan, A., Murphy, K., O'Sullivan, A., Humphreys, K., & O'Riordan, E. (2018). A metabonomic comparison of the diets of vegetarians and non-vegetarians. *Journal of Proteomics, 156*, 1-11.

Li, Y., Liu, X., & Li, X. (2020). Metabonomic data analysis: a review of methods and applications. *Analytica Chimica Acta, 1098*, 1-20.

Liu, Y., Chen, X., Wang, J., Liu, Y., & Li, X. (2015). Metabolic profiling reveals the similarity of cancer and normal cells. *PloS one, 10*(8), e0134517.