Exploiting the Classification Algorithm for Robust Multidimensional Homogeneity Test in Causal Inference

Proposal

Aukkawut Ammartayakun

MA 590 Special Topics: Causal Inference

Spring 2023

1 Introduction

The simple but effective way to quickly estimate whether the average treatment of two treatments are zero or not on the randomized controlled experiment (RCE) is to perform hypothesis testing. Specifically, the fisherian approach where the null hypothesis on the RCE is that we can't tell the difference between the two treatments which can be simplified into the homogeneity test. Traditionally, this can be test with hypothesis testing like the Kolmogorov-Smirnov test. However, those test has a limitation on the dimension of the data. For example, to consider the difference between two text generated models in terms of the students performance after receiving the hint from those two models, it is hard to do so with the traditional hypothesis testing, let alone any other method that does not involves with machine learning.

This work challenges that limitation by looking at this problem from a different perspective. Recall that hypothesis testing is essentially the classification problem of whether one has statistical evidence to reject the null hypothesis or not. Semantically, they both are different approaches but those problems are equivalent in the sense that we can have the classification functions $f : \mathcal{D} \to [0, 1]$ such that $f(x, y)$ for $(x, y) \in \mathcal{X} \supseteq \mathcal{D}$ reflecting the probability of rejecting the null hypothesis $H_0 : X = Y$ for paired data (x, y) in the data space $\mathcal{X}$ which now, the p-value can be estimated from f and the decision can be done with thresholding.

This work will explore the possibility of exploiting the classification algorithm for robust multidimensional homogeneity test (naive way to test the difference between two treatments) with the application in causal inference. First, this work will explore the known method for homogeneity test in high dimensional space like exploiting f-divergence[3], maximum mean discrepancy[2], C2ST[5], etc. Then, the proposed method will be analyzed and tested in comparison to the former methods.

1.1 Hypothesis testing exploitation

As the process of hypothesis testing involves evaluating the test statistics and using that to find the critical region under the distribution. However, that can be simplified, hopefully, into the classification problem that can be exploited with the machine learning model. The transformation of the data into its representation form, i.e., embedding might reveal the underlying causal relation. The use of a machine learning model in substitution for hypothesis testing for causal inference will be explored, along with the analysis of the power of testing using a classification model.

2 Dataset

Food.com[6] recipe and the rating will be used as the dataset. The randomized experiment in this case is to test whether there is the difference in rating in wording of the recipe between the treatment group and the controlled group. The treatment group is the recipe that has the word "healthy" in it, and the controlled group is the recipe that does not have the word "healthy" in it.

The key (derived) variables here are:

1. Description: The description of the recipe
2. Rating: The rating of the recipe
3. Healthy: The indicator of whether the recipe has the word "healthy" in it or not

In this case the healthy variable is an instrument variable to this randomized experiment.

3 Goal

To apply the exploited homogenity test to test whether the result from the treatment group and outcome group are homogenious or not. If the result is not homogenious, then the treatment group has an effect on the outcome group.

4 Proposed Approach

In the similar manner to [5], the proposed method will be to use the classification algorithm to test the homogenity of the treatment group and the outcome group. However, instead of later use t-test on the test statistic that derived from the classification model, what if the model itself return p-value? As discussed on the introduction, the hypothesis testing is essentially the classification problem of whether one has statistical evidence to reject the null hypothesis or not.

In this task, the aggregated embedding which is essentially similar to the sufficient statistic, however, it is just a summary (i.e., the moment) of the all embedding in each group will be used and pass that through the classification model like siamese network[4] which returns p-value as its confidence in rejecting the null hypothesis in homogeneity test.

The text embedding can be generated using the pre-trained model like BERT[1]. Different aggregation will be explored, this includes the mean, median, and the max pooling.

The performance and test power of the proposed method will be compared to the other high dimensional homogenity test like the MMD[2] and the C2ST[5] test.

References

[1] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[2] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research 13*, 25 (2012), 723–773.

[3] Kanamori, T., Suzuki, T., and Sugiyama, M. f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models, 2010.

[4] Koch, G., Zemel, R., and Salakhutdinov, R. Siamese Neural Networks for One-shot Image Recognition. Proceedings of the 32nd International Conference on Machine Learning.

[5] Lopez-Paz, D., and Oquab, M. Revisiting classifier two-sample tests. In *International Conference on Learning Representations* (2017).

[6] Majumder, B. P., Li, S., Ni, J., and McAuley, J. Generating personalized recipes from historical user preferences. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong, China, Nov. 2019), Association for Computational Linguistics, pp. 5976–5982.