Leveraging Causal Inference for Detecting Distribution Shift with Application to Metrics Attribution Analysis

Heqiao Ruan[1,2], Jian Tian[1], Lucas Mentch[2,*]

Tiktok. Inc.[1], Department of Statistics, University of Pittsburgh[2]

Motivation

- In industry, huge amounts of data are generated incrementally where it comes in batches - existing batch of data $\mathbf{D}_{exist} = (\mathbf{X}_{exist}, Y_{exist})$ and new batch of data $\mathbf{D}_{new} = (\mathbf{X}_{new}, Y_{new})$. Types of shifts can vary:
  - Covariate Shift: $P(\mathbf{X}_{exist}) \neq P(\mathbf{X}_{new})$ where $P(Y_{exist}|\mathbf{X}_{exist}) = P(Y_{new}|\mathbf{X}_{new})$, only covariates distribution change where the conditional dependence remain the same.
  - Concept Drift: $P(Y_{exist}|\mathbf{X}_{exist}) \neq P(Y_{new}|\mathbf{X}_{new})$ where $P(\mathbf{X}_{exist}) = P(\mathbf{X}_{new})$, only conditional dependence change where the covariate distribution remain the same
  - Distribution Shift: At least one of $P(Y|\mathbf{X})$ and $P(\mathbf{X})$ changes or both changes across existing batch of data and new batch of data.
- In practice, the existence of both types of shifts is prevalent.
- Existing methods [2, 4, 5, 6] only focus on certain type of shift. We aim to develop hypothesis testing procedure efficient in most of scenarios.

Methodology

- We treat the existing batch of data $D_{exist} = (\mathbf{X}_{exist}, Y_{exist})$ as control group with treatment assignment $T = 0$ and the new batch of data $D_{new} = (\mathbf{X}_{new}, Y_{new})$ as treatment group with treatment assignment $T = 1$. Then the variable importance will represent the relative degree of change of $Y|\mathbf{X}$

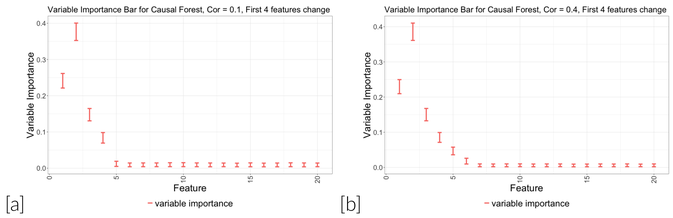

Figure 1. Illustration of the Motivation of Test - the variable importance is proportional to the degree of the change of data-generating process(DGP)

- Existing batch of Data $D_{exist} : Y = X_1 + ... + X_5 + X_6 + ... + X_{10} + \epsilon$ and New batch of Data $D_{new} : Y = 4 * X_1 + 5 * X_2 + 3 * X_3 + 2 * X_4 + X_5 + ... + X_{10} + \epsilon$
- Causal Forest([1]) is adapted, Confidence Interval of Feature Importance is displayed in Figure 1. The VIMP(Variable Importance) is approximately proportional to the degree of change of data-generating process $Y|\mathbf{X}$.
- The methodology can be summarized as a general-purposed three-step procedure, the details can be found in **Algorithm 1**.
  - First fit a causal learner and record the (original)variable importance.
  - Permute the Treatment Assignment between the two batches of data for **B** times, retrain the causal learner and record the permuted variable importance.
  - Perform the hypothesis test based on whether there exist some features whose original variable importance is larger than all of the permuted variable importance.

Algorithm 1 Testing Covariate Shift via Permuting Treatment Assignment - **CFPerm**

BEGIN Start with existing batch of data $\mathcal{D}_{exist} = \{(\mathbf{X}_1, Y_1), ..., (\mathbf{X}_{n_{exist}}, Y_{n_{exist}})\}$ where $(\mathbf{X}_i, Y_i)$ and new batch of data $\mathcal{D}_{new} = \{(\mathbf{X}_{n_{exist}+1}, Y_{n_{exist}+1}), ..., (\mathbf{X}_{n_{exist}+n_{new}}, Y_{n_{exist}+n_{new}})\}$ where $(\mathbf{X}_i, Y_i)$, with each $\mathbf{X}_j = (x_1, ..., x_p)$. Assign (control) treatment $T = 0$ to existing batch of data and $T = 1$ (treatment) to new batch of data. Define hypotheses $H_0 : P(\mathbf{X}_{new}, Y_{new}) = P(\mathbf{X}_{exist}, Y_{exist})$ vs. $H_a : P(\mathbf{X}_{new}, Y_{new}) \neq P(\mathbf{X}_{exist}, Y_{exist})$.

S1 Fit the causal forest on $(\mathbf{X}, Y, T)$ and record the variable importance (VIMP) for each feature $v_1, ..., v_p$ (VIMP can be user-specified as cover based or variance based)

S2 Permute the treatment assignment T randomly across the data a total of B times (by default $B = 500$). Each time, fit the causal forest on the resulting permuted batches and calculate the variable importance, recorded as $v_{b,1}, ..., v_{b,p}$ for $b = 1, .., B$.

S3 Construct an interval for the variable importance of each of the feature with the lower bound as the smallest value and the higher bound as the largest value across the permutations $I_i = (v_{i,min}, v_{i,max}), i = 1, 2, ..., p$

S4-1 **CFPerm0**: The maximal value of the permuted confidence interval of variable importance for each of the feature is recorded as $I_{max} = \{v_{max,1}, ..., v_{max,p}\}$. Reject H_0 if there exists a feature whose original variable importance is larger than the maximal of the upper bound among all of the features $max(I_{max})$ served as the decision threshold; otherwise, do not reject.

S4-2 **CFPerm1**: The 99% upper quantile for the permuted confidence interval of variable importance in each of the feature is recorded: $I_{99\%} = \{v_{1,99\%}, ..., v_{p,99\%}\}$ Reject H_0 based on whether there exists a feature whose original variable importance is larger than the 95% upper quantile $q_{0.95}(I_{99\%})$ of the recorded feature-wise 99% quantile across the features served as the decision threshold. Otherwise the null hypothesis would not be rejected.

END

Figure 2. CFPerm Methodology for detecting the distribution shift

Validity of the Test

- Type-I error is well-controlled(Figure 3 [a]-[b]).
- The test would achieve higher power in detecting concept drift(Figure 3[c]), covariate shift(Figure 3[d]) compared with the existing methods([2, 4, 5, 6]).
- The performance is highly competitive compared with the existing methods on a variety of **UCI** datasets where various features are adapted to perform the split into two batches of data as displayed in Table 1.

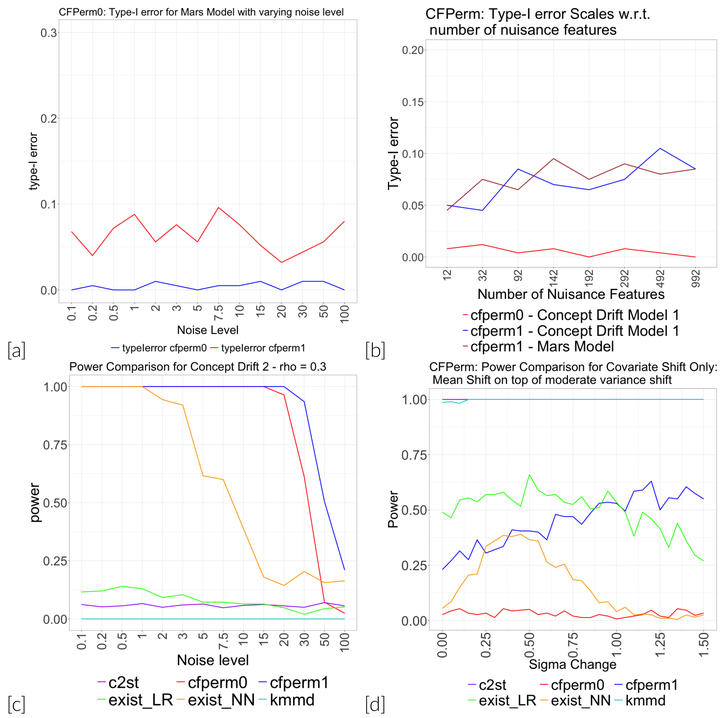

Figure 3. (a) Type-I error in models with different data quality (b) Type-I error w.r.t. different numbers of nuisance noise features (c) Power Comparison with existing methods for Concept Drift (d) Power Comparison with existing methods for Covariate Shift

Dataset - Feature	CFPerm0	CFPerm1	Existing-Max
AToxic - V1	0.035	0.405	0.370
Autoprice - V10	0.135	0.807	0.410
Autoprice - V5	0.205	0.830	0.408
Boston - lstat	0.005	0.727	0.265
Autoprice - V7	0.220	0.920	0.445
Boston - nox	0.085	0.510	0.469
Autoprice - V8	0.260	0.770	0.500
Boston - crim	0.170	0.833	0.370

Table 1. Comparison of power on a variety of **UCI** datasets

Theoretical Guarantee Given $(\mathbf{X}_{\text{exist},i}, Y_{\text{exist},i}) \overset{\text{iid}}{\sim} P_1$ for $i = 1, \ldots, n_{\text{exist}}$ and $(\mathbf{X}_{\text{new},i}, Y_{\text{new},i}) \overset{\text{iid}}{\sim} P_2$ for $i = 1, \ldots, n_{\text{new}}$ with p covariate in $\mathbf{X}$. The original variable importance $v_1, ..., v_p$ and the permuted variable importance is $\cup_{j=1}^{B}\{v_{1,j}, ..., v_{p,j}\}$ The hypothesis testing procedure for distribution shift is $H_0 : \forall j \;\; v_j \leq \max_{1 \leq i \leq k} max(v_{i,1}, ..., v_{i,B}) \; v.s. \; H_a : \exists j \; s.t. \; v_j > \max_{1 \leq i \leq k} max(v_{i,1}, ..., v_{i,B}) \; \forall j \in \{1, 2, ..., p\}$. With number of permutations $B > \frac{k}{\alpha} + 1$, the type-I error is well-controlled under significance level α.

Flexibility of the Framework

The **CFPerm** procedure is only one of the many under the framework We propose a class of hypothesis testing procedure which can be summarized into a general-purposed three-step procedure displayed in Figure 4. It possesses amazing flexibilities.

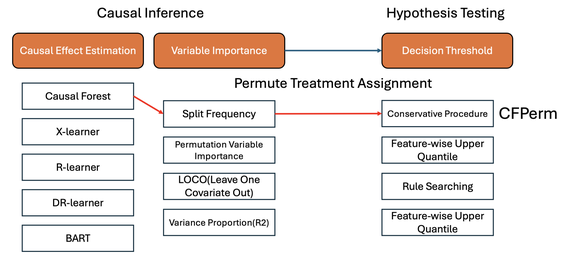

Figure 4. The Diagram for the proposed framework combining **Causal Inference** and **Online Learning**

Use Case: Creator Segmentation

We adapt the procedure to a use-case in business scenario to characterize the change of the business metrics and identify the important indicators for the difference of creators' behavior in **Tiktok** Platform in the June(50 **Metrics** built, served as X) in the number of videos they published(Y) among **Sports Category Creators** and **Education Category Creators**. The metrics that contributes significantly to the distribution shift is displayed in Table 2 with their relative contribution(proportion of variable importance, sum up to 1) and corresponding p-value. The overall p-value with 0.048 is achieved which indicate the distribution shift between the creator's behaviors within these two segments.

Metrics	Variable Importance	p-value
Author Average VV(Video Views) past 30days	9.4%	0.000
Author Ads Value	9.2%	0.002
Author Reward USD Amount past 30days	6.7%	0.002
Author Activity Countries Past 30 Days	5.8%	0.008
Author Special Reward Income	5.4%	0.028
Author Average VV(Video Views) from Japan in past 30 days	5.0%	0.046

Table 2. Metrics that contribute significantly to the change of the creators' behavior

Conclusion & Future Directions

- We propose a general-purposed framework combining causal inference and online learning to detect distribution shift with solid theoretical support.
- The **CFPerm** procedure of hypothesis testing is highly flexible, scalable and much more powerful in detecting distribution shift compared with existing methodologies.
- Significant potential(future) use cases can be adapted include but not limited to **distribution shift testing**, **metrics attribution analysis** and **customer segmentation**.

References

[1] Susan Athey, Julie Tibshirani, and Stefan Wager.
Generalized random forests.
2019.

[2] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola.
A kernel two-sample test.
The Journal of Machine Learning Research, 13(1):723–773, 2012.

[3] Oliver Hines, Karla Diaz-Ordaz, and Stijn Vansteelandt.
Variable importance measures for heterogeneous causal effects.
arXiv preprint arXiv:2204.06030, 2022.

[4] Xiaoyu Hu and Jing Lei.
A two-sample conditional distribution test using conformal prediction and weighted rank sum.
Journal of the American Statistical Association, 119(546):1136–1154, 2024.

[5] David Lopez-Paz and Maxime Oquab.
Revisiting classifier two-sample tests.
arXiv preprint arXiv:1610.06545, 2016.

[6] Jian Yan and Xianyang Zhang.
A nonparametric two-sample conditional distribution test.
arXiv preprint arXiv:2210.08149, 2022.

Presenting Author: Heqiao Ruan Email: [1] :her76@pitt.edu, [2] :jian.tian@bytedance.com [2] :lkm31@pitt.edu