UIC CS594 literature review

Causality-based Explanation of Classification Outcomes

Chenjie Li

ACM Reference Format:
Chenjie Li. 2024. UIC CS594 literature review: Causality-based Explanation of Classification Outcomes. In . ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

This paper proposes a novel framework to address the interpretability of black-box classification models using a causality-based approach. The authors introduce the RESP score, an extension of the simplified notion of actual cause and responsibility proposed in [2] which evaluates the influence of features on a classification outcome. The authors compared RESP score with SHAP score, a metric derived from Shapley values, and discussed the capabilities and trade-offs between those 2 measures. The computational challenges with these metrics are analyzed. Based on those challenges, the authors proposed two probabilistic spaces: product and empirical probability spaces. Experiments evaluated the proposed approach using those 2 probablistic spaces on financial risk assessment and fraud detection.

2 THE COUNTER SCORE AND RESP SCORE

2.1 COUNTER SCORE

Lets start with RESP score. The definition of RESP score is adapted from the definition from [2]. Given an entity, the *counterfactual cause* is defined as a feature and a possible value from that feature that would flip the prediction by a classifier. In other words, if we can flip a classifier's prediction by changing the value of a feature, this feature and value together would be called counterfactual cause.

$$\text{COUNTER}(\mathbf{e}^\star, F_i) \stackrel{\text{def}}{=} L(\mathbf{e}^\star) - \mathbf{E}\left[L(\mathbf{e})|\mathbf{e}_{\mathcal{F}-\{F_i\}} = \mathbf{e}^\star_{\mathcal{F}-\{F_i\}}\right]$$

Figure 1: COUNTER score definition for a feature

To better understand COUNTER score of a feature, let's walk through an example I created. Suppose our setting is the loan approval and outcome=1 means the black box model output approval as the outcome and 0 otherwise. And the entity of interest $e^\star$ is T_5. suppose our candidate feature F_i is *place of origin*(POO). Then the set that fulfills $L(e)|e_{\mathcal{F}-\{F_i\}} = e^\star_{\mathcal{F}-\{\mathcal{F}_i\}} = \{T_1, T_2\}$. Out of those 2 tuples, both of them get output as approved, so the $COUNTER(e^\star, \text{“}POO\text{”}) = 1 - 1 = 0$. It is possible for an entity $e^\star$ to have no counterfactual cause. This happens when, changing any single feature to any other value, the modified entity e has the same outcome $L(e^\star) = L(e) = 1$. A pair (F_i, v) is called an actual cause with contingency (Γ, w), where Γ is a set of features and w is a set of values, if (F_i, v) is a counterfactual cause for $e^\star[\Gamma := w]$.

Name	Education l	Credit Score	Place of Origi	Approved	
Alice	Master's	High	China	1	T1
Bob	Master's	High	USA	1	T2
Carol	Master's	Low	USA	0	T3
David	Master's	Low	USA	0	T4
Eve	Master's	High	China	1	T5
Grace	Master's	Low	China	0	T6
Frank	Bachelor's	Middle	USA	0	T7

Figure 2: COUNTER score definition for a feature

RESP is derived from the definition of contingency.

Fix an entity $e^\star$, and a contingency (Γ, w) such that $L(e^\star) = L(e')$, where $e' \stackrel{\text{def}}{=} e^\star[\Gamma := w]$. The RESP-score of a feature F_i w.r.t. to the contingency (Γ, w) is: $RESP(e^\star, F_i, \Gamma, w) \stackrel{\text{def}}{=} \frac{L(e') - \mathbb{E}[L(e'')|e''_{F-\{F_i\}} = e'_{F-\{F_i\}}]}{1+|\Gamma|}$.

3 SHAP SCORE

SHAP-score, based on Shapley values from cooperative game theory [3], quantifies the contribution of a feature by averaging its impact across all possible feature subsets. The definition of SHAP for a feature F_i with respect to an entity e^* is as follows:

$$SHAP(e^\star, F_i) = \frac{1}{n!}\sum_\pi \left(E[L(e)|e_{\pi \le F_i} = e^\star_{\pi \le F_i}] - E[L(e)|e_{\pi < F_i} = e^\star_{\pi < F_i}]\right)$$

Here, π is a permutation of all features, $\pi \le F_i$ includes F_i and all preceding features in the permutation, and $\pi < F_i$ includes all preceding features except F_i. The conditional expectations represent the model's outcome averaged over the remaining unknown features.

Intuitively, SHAP evaluates how much the addition of a particular feature changes the predicted outcome, averaged across all possible feature orderings. The formula ensures that the contribution of each feature is weighted equally across all scenarios.

Despite its utility, calculating SHAP scores is computationally expensive due to the factorial number of permutations required. Efficient approximations or alternative solutions need to be proposed in order to solve this problem.

4 PROBABILITY SPACES

Both RESP and SHAP scores rely on defining a probability space that describes how the data behaves. These probability spaces provide the foundation for calculating conditional expectations, which are essentially the source of the cost of the computation. The challenge lies in balancing computational efficiency with realistic representation. Two specific spaces, the **product space** and the **empirical distribution**, are introduced to address these needs, each with distinct properties.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

4.1 Product Space

The product space assumes feature independence, which enables a simplified RESP-score calculation process. It defines the probability of an entity as the product of the marginal probabilities of its features. For instance, given features $F_1, F_2, \ldots, F_n$, the probability of an entity $e = (x_1, x_2, \ldots, x_n)$ is:

$$P(e) = \prod_{i=1}^{n} P(F_i = x_i) \tag{1}$$

This assumption simplifies RESP calculation since the required conditional expectations can be calculated directly from marginal distributions. However, it oversimplifies the real data distributions since we all know in real world data it often contains correlations among features.

4.2 Empirical Distribution

The empirical distribution captures observed data correlations by directly using the dataset without generalizing to unobserved entities. It defines probabilities based solely on the frequency of entities in the dataset. Since the "data at hand" is quite limited when enumerating the subsets when we calculate the SHAP value, the authors observed that there's opportunities to do "early stopping" since certain cohorts will not be present in this distribution. While this makes the SHAP computation feasible, it is clearly not the real RESP score value.

5 EXPERIMENTS AND ANALYSIS

Experiments were conducted on two datasets: the FICO dataset (credit risk assessment) and Kaggle's credit card fraud detection dataset.

5.1 FICO Dataset

The FICO dataset was used to compare RESP and SHAP with a well-established white-box explanation, the FICO-explanation from Duke University [1]. This dataset emphasizes explainability in financial decision-making, making it crucial to evaluate how well RESP and SHAP align with the white-box standard.

The FICO-explanation from Duke is essentially a "white box" logistic regression model trained on the dataset. By leveraging a model with interpretable parameters, the authors ensured that both methods could be evaluated against a ground truth derived from model coefficients and FICO's hierarchical structure.

Two of the experiment results on FICO dataset are shown in in Figure 3 and Figure 4.

In Figure 3, while both FICO-explanation and RESP ranks the MMR7 at the top, SHAP actually didn't give it any weights. The reasoning provided by the authors is the following: given the formula of the SHAP, the term $\frac{l!(n-l-1)!}{n!}$ are head and tail heavy, making the features that distributed evenly in both classes suffer in terms of its importance. MMR7 happens to fall into this category.

In contrast, in Figure 4, 'ExternalRiskEstimate' is ranked high in RESP and COUNTER. However it got 0 importance in FICO-explanation. The reason is two-fold: 1.The way FICO scores are sorted hierarchically and it can make some abrupt cutoffs which can lead to such drastic difference (either very important or not important at all) which also results in the fact that FICO-scores are less diverse. 2.Model makes predictions on single cases and ignored the rest of the data

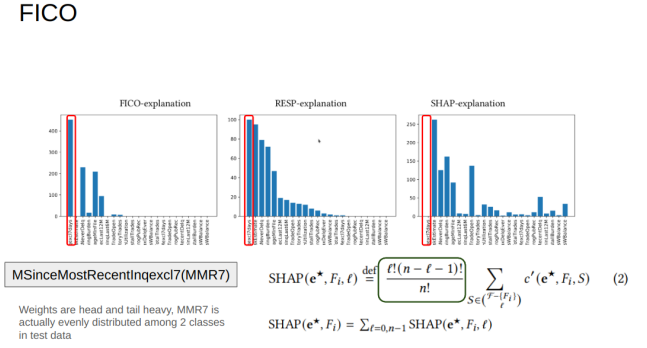

Figure 3: MMR7 difference from different scores

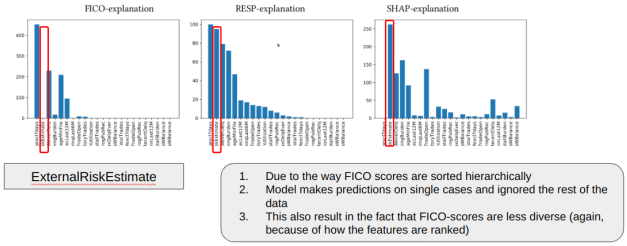

Figure 4: ExternalRiskEstimate difference from different scores

5.2 Kaggle Credit Card Fraud Dataset

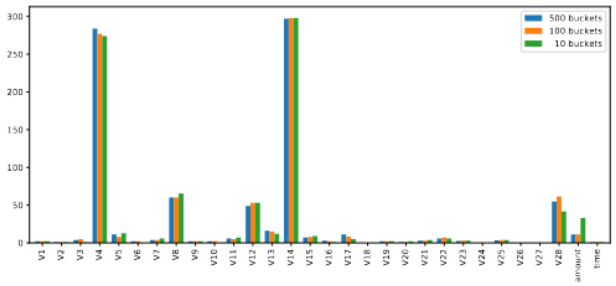

Figure 5: RESP bucketized results

In Figure 5 and Figure 6, the impact of bucketization on feature importance is studied. RESP explanations, calculated using the product space, display insensitivity to bucket size changes. This consistency arises because RESP relies on predefined independence assumptions, resulting in its causal explanations remain stable regardless of the feature distribution within buckets.

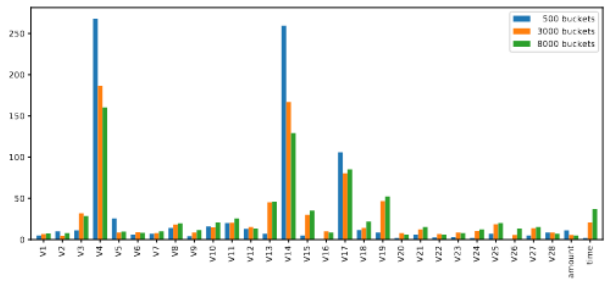

Figure 6: SHAP bucketized results

In contrast, SHAP explanations are sensitive to bucket size adjustments, which directly influence how feature values are grouped and analyzed. Smaller buckets increase granularity, allowing SHAP to more precisely capture subtle variations in feature contributions. However, this also increases computational complexity and introduces noise in cases with limited data, as fewer examples fit within each bucket. Larger buckets, while reducing noise, risk oversimplifying the contributions of features by merging diverse values into broader categories.

6 CONCLUSION

This paper explored causality-based explanations for black-box models. RESP and SHAP, each address critical aspects of interpretability but also have its limitations especially when it comes to computational cost. Future research should continue to study the optimization/approximation strategies.

REFERENCES

[1] Chaofan Chen, Kangcheng Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang, and Tong Wang. 2018. An interpretable model with globally consistent explanations for credit risk. *arXiv preprint arXiv:1811.12615* (2018).

[2] Alexandra Meliou, Wolfgang Gatterbauer, Katherine F Moore, and Dan Suciu. 2010. The complexity of causality and responsibility for query answers and non-answers. *arXiv preprint arXiv:1009.2021* (2010).

[3] Lloyd S Shapley. 1953. A value for n-person games. *Contribution to the Theory of Games* 2 (1953).