View Reviews

Paper ID
1173

Paper Title
Explanation Shift: Detecting and quantifying the decay of predictive performance and fairness on tabular data

Track Name
Safe and Robust AI Track

Reviewer #2

Questions

1. {Summary} Please briefly summarize the main claims/contributions of the paper in your own words. (Please do not include your evaluation of the paper here).
This paper proposes a new measure of distribution shift that is helpful in predicting decay in model predictive performance and fairness across protected groups. The measure is based on changes in model explanations across two samples of data, the original data sample and a small sample of OOD data. The approach is to bootstrap small training datasets for a decay classifier, using the ID and OOD data. The decay classifier is a logistic regression that predicts whether the task model, either a linear regression or boosted trees, declines past a given threshold in terms of accuracy or equal opportunity fairness. Experiments with synthetic data indicate that a Kolmogorov-Smirnov test on the distribution of SHAP explanation scores obtained by applying the task model to old and new data will better indicate whether there has been a distribution shift than the same statistical test directly on the data or prediction distributions. Experiments with real world data show that the decay classifier trained with explanation shift statistics usually achieves slightly better AUC than baselines including another method, ATC, as well as decay classifiers trained on measures of input and prediction shifts.

2. {Strengths and Weaknesses} Please provide a thorough assessment of the strengths and weaknesses of the paper, touching on each of the following dimensions: novelty, quality, clarity, and significance.
Strengths:
- Very important: The paper makes reasonable experimental design choices and sets out to show an interesting hypothesis, that explanation shift is a good predictor of decay in model accuracy and fairness.
- Important: The writing is pretty straightforward, and the paper is overall easy to understand.
- Of minor importance: The theoretical cases and results with synthetic data help demonstrate the core ideas behind the approach, before seeing the experimental results with real data.

Weaknesses:
- Very important: The gains over prior methods are usually pretty small, and ultimately the baselines are not strong enough. I have one question about how exactly Dist. Shift column would be calculated (see questions below), but I think my concern applies regardless. A few subpoints: (1) What is ATC and how does it work? If this is the paper's SOTA baseline, I would expect more description of the method and a more thorough explanation of why it doesn't work at all in this settings while very simple methods like Pred. Shift work decently. (2) Why not measure shift across p(X, predY) using a multidimensional measure of shift, or a decay classifier which has feature interactions (if first computing shift on each feature then passing a feature vector to the decay classifier)? This seems pretty straightforward and I predict it would perform similarly to the explanation shift. Does no prior work consider measuring shift based on statistics of the joint input and prediction distribution?
- Very important: To my understanding, the experiments do not really check for a case of false positives, which I suspect could be common when basing decay predictions off of explanation shift. This is based on my

understanding that (1) the experiments with real world data only include cases of actual distribution shifts, and not any cases of no distribution shift, (2) explanations will change significantly across different regions of a single data distribution, for a nonlinear model, even in the absence of any distribution shift, while model performance might not always vary across these different segments of a single data distribution. I suspect the baselines might avoid false positives in a way that explanation shift does not.
- Of minor importance: Since the theoretical results use linear models, they are not that surprising or deeply interesting. For instance, SHAP will basically exactly recover the linear model, and so of course two linear models trained on different posterior distributions will yield distinct SHAP explanations.

3. {Questions for the Authors} Please carefully describe questions that you would like the authors to answer during the author feedback period. Think of the things where a response from the author may change your opinion, clarify a confusion or address a limitation. Please number your questions.
Comments and Questions:
- One question on the approach first seen in “Quantifying model performance decay on synthetic data.” How many features are given to the logistic regression? Is it one KS test on each feature? Or is a multidimensional KS test across the explanation distributions, to produce a single input feature for the logistic regression? Did you try both?
- I don’t really understand why the paper claims “Change in distribution or explanations do not necessarily imply a model performance decay.” In Table 5, four of the five shifts tested showed changes in accuracy or EOF. Isn’t that evidence that changes in distribution usually imply changes in performance? This point might as well be made in a more precise, probabilistic manner. The univariate testing here seems pretty limited too. Why not use multidimensional Kolmogorov-Smirnov?
- You could try training your decay classifier using all of the shift measurements as inputs, and then see if adding Explanation shift statistic improves performance significantly via an F-test over the two models.

4. {Evaluation: Novelty} How novel are the concepts, problems addressed, or methods introduced in the paper?
Fair: The paper contributes some new ideas or represents incremental advances.

5. {Evaluation: Quality} Is the paper technically sound?
Poor: The paper has major technical flaws. For example, the proof of the main theorem is incorrect or the experimental evaluation is flawed and fails to adequately support the main claims.

6. {Evaluation: Significance} How do you rate the likely impact of the paper on the AI research community?
Poor: The paper is likely to have minimal impact on SRAI or is out of scope of the SRAI special track.

7. {Evaluation: Clarity} Is the paper well-organized and clearly written?
Fair: The paper is somewhat clear, but some important details are missing or unclear.

8. (Evaluation: Reproducibility) Are the results (e.g., theorems, experimental results) in the paper easily reproducible? (It may help to consult the paper’s reproducibility checklist.)checklist.)
Poor: key details (e.g., proof sketches, experimental setup) are incomplete/unclear, or key resources (e.g., proofs, code, data) are unavailable.

11. (OVERALL EVALUATION) Please provide your overall evaluation of the paper, carefully weighing the reasons to accept and the reasons to reject the paper.
Reject: For instance, a paper with poor quality, inadequate reproducibility, incompletely addressed ethical considerations.

Reviewer #3

Questions

1. {Summary} Please briefly summarize the main claims/contributions of the paper in your own words. (Please do not include your evaluation of the paper here).

This paper proposes a novel method to detect and quantify the decay of predictive performance and fairness, called explanation shift. The mathematical analysis of synthetic examples and experimental evaluation of real-world tabular data are provided for supporting the author's claims.

2. {Strengths and Weaknesses} Please provide a thorough assessment of the strengths and weaknesses of the paper, touching on each of the following dimensions: novelty, quality, clarity, and significance.

Pros:

1. The problem of detecting and quantifying the model degradation is timely and important in the ML community. Also, the motivation to detect performance shifting via explanation shift on statistical features is of novelty.

2. The author considers three shifting scenarios which can case model degradation, where each scenario is provided with a solution and the corresponding theoretical analysis.

3. The data, code, and experimental settings are provided for reproducibility.

Cons:

1. A flowchart or a well-organized algorithm should be presented to show the progress of the developed method.

2. The paper provides three model performance degradation scenarios: i) multivariate shift; ii) posterior shift; iii) uninformative features. However, one of my concern is that, can the three mentioned scenarios occur in a hybrid way? Under this situation, what will be the challenge for the developed method?

3. The experiment on real-world datasets on contains the US income data, which might not be strong enough to illustrate the superior of the provided method.

3. {Questions for the Authors} Please carefully describe questions that you would like the authors to answer during the author feedback period. Think of the things where a response from the author may change your opinion, clarify a confusion or address a limitation. Please number your questions.

Please also refer the comments in the strengths and weaknesses sections, and I have the following questions:

1. As mentioned in the conclusion section, the paper makes strong assumptions that the types of shift can be known prior the method application, is this assumption valid? If yes, please provide more references or proofs.

2. Is the provided method able to handle a hybrid scenario of all three shifting mentioned in the paper? An ablation study is encouraged to be performed.

3. What would be the computation cost of using Shapley value for the proposed explanation shift method?

4. {Evaluation: Novelty} How novel are the concepts, problems addressed, or methods introduced in the paper?

Fair: The paper contributes some new ideas or represents incremental advances.

5. {Evaluation: Quality} Is the paper technically sound?

Good: The paper appears to be technically sound. The proofs, if applicable, appear to be correct, but I have not carefully checked the details. The experimental evaluation, if applicable, is adequate, and the results convincingly support the main claims.

6. {Evaluation: Significance} How do you rate the likely impact of the paper on the AI research community?

Fair: The paper is likely to have modest impact within a subfield of SRAI.

7. {Evaluation: Clarity} Is the paper well-organized and clearly written?

Good: The paper is well organized but the presentation has minor details that could be improved.

8. (Evaluation: Reproducibility) Are the results (e.g., theorems, experimental results) in the paper easily reproducible? (It may help to consult the paper's reproducibility checklist.)checklist.)
Good: key resources (e.g., proofs, code, data) are available and sufficient details (e.g., proofs, experimental setup) are described such that an expert should be able to reproduce the main results.

11. (OVERALL EVALUATION) Please provide your overall evaluation of the paper, carefully weighing the reasons to accept and the reasons to reject the paper.
Borderline reject: Technically solid paper where reasons to reject, e.g., poor novelty, outweigh reasons to accept, e.g. good quality. Please use sparingly.

14. I acknowledge that I have read the author's rebuttal (if applicable) and made changes to my review as needed.
Agreement accepted