Investigation and Comparison of Evaluation Methods of Model-Agnostic Explainable AI Models

Vanisha Oedayrajsingh Varma
v.oedayrajsinghvarma@student.tudelft.nl

Supervisor: Chhagan Lal
Responsible Professor: Mauro Conti

TU Delft

1 BACKGROUND

- Trust in AI is important in high-stakes applications, such as in medicine, criminal justice, transport and finance.
- However, black-box machine learning (ML) algorithms often have low transparency and interpretability.
- Explainable AI (XAI) models provide explanations for a model's decisions and predictions.
- This way, common black-box issues such as trust, accountability, and transparency can addressed.
- The more users trust AI, the more it can be used in critical environments.
- XAI can also be used to debug and improve ML models.
- Quality of explanations is an important part of increasing trust.
- However, explanations are often not systematically assessed and there is no consensus on how to evaluate them.
- Research into evaluation methods is still low.

2 RESEARCH QUESTION

- XAI can be categorized into model-specific and model-agnostic techniques.
- This research focusses on model-agnostic techniques.
- Identify common trends in the evaluation of state-of-the-art model-agnostic XAI models, and identify any missing or undervalued evaluation methods.

3 MODEL-AGNOSTIC XAI

Model	Explanations based on
LIME	Local interpretable approximations of complex models, identifies the local important variables
SHAP	Assigning an importance value to each feature for a specific prediction
Anchors	If-then rules called 'anchors', if an anchor holds, the prediction is (almost) always the same
PASTLE	Feature importance combined with pivots
CASTLE	Feature importance combined with clusters

Model	Application-grounded evaluation	Human-grounded evaluation	Functionality-grounded evaluation	Real-world task	Proxy task	Human-centred metrics	Functionality-grounded metrics
LIME	✓	✓	✓		✓	Trust Understanding Task performance Task time length Ability to detect errors/Bias detection	Fidelity Level of (dis)agreement 1 unseen
SHAP		✓	✓		✓	1 unseen	Accuracy Fidelity Computational efficiency (unseen)
Anchors	✓		✓		✓	Understanding Task time length	Fidelity
PASTLE	✓				✓	Understanding 1 unseen	
CASTLE	✓		✓		✓	Understanding	Fidelity Computational efficiency (unseen)

Evaluation approach	Both model-agnostic & model-specific	Model-specific only
Human-centred	Trust, Appropriate Trust, Satisfaction, Understanding, Task time length, Task performance, Ability to detect errors/Bias detection, Physiological indicators, Preference	-
Functionality-grounded	Fidelity, Accuracy, Level of (dis)agreement, Reliability, Privacy, Agreement, Monotonicity, Non-sensitivity, Effective complexity, Consistency, Validity, Proximity, Sparsity, Diversity, Closeness, Feasibility, Identity, Separability, Novelty, Representativeness	Implementation invariance, Continuity, Selectivity

4 TAXONOMY & ANALYSIS

- Human-centred evaluation:
  - Application-grounded: Evaluation with domain experts or end users.
  - Human-grounded: Evaluation with lay humans.
- Functionality-gounded evaluation: Formal properties of the explainer are measured as proxies for explanation quality.
- *Understanding* and *fidelity* are the most commonly tested metrics.
- There were some metrics that could not be classified.
- Only proxy tasks were used to evaluate the models.

5 CONCLUSION & FUTURE WORK

- There is a lack of variety in the metrics that are assessed.
- Metrics that are not often assessed in user-studies but deserve more attention are *(appropriate) trust*, *task time length*, and *task performance*.
- There is a lack of evaluation using real-world tasks.
- There is a lack of standardisation in measurement methods and units of measurement.
- Evaluate the state-of-the-art models on unmeasured or rarely measured human-centred metrics.
- Evaluate the state-of-the-art models on unmeasured or rarely measured functionality-grounded metrics.
- Evaluate LIME, SHAP, PASTLE, and CASTLE using the functionality-grounded metrics specific for feature attribution methods, such as monotonicity, non-sensitivity, and effective complexity.
- Assess the benefits and potential future use of the unseen methods and metrics found in the analysis.
- Clearly define and standardise measurement methods for all found metrics.
- Evaluate the state-of-the-art model-agnostic XAI models using real-world tasks.