Explainabilty Metrics and Properties for Counterfactual Explanation Methods

EXTRAAMAS 2022

Vandita Singh Kristijonas Cyras Rafia Inam

09-10 May 2022

Ericsson Research, Sweden Uppsala University, Sweden

Introduction
Background

- Explainable AI(XAI) Methods - Understanding the rationale behind the output provided by AI systems
- Build trustworthy AI systems that comply with regulations
- Obtaining explanations is not enough!

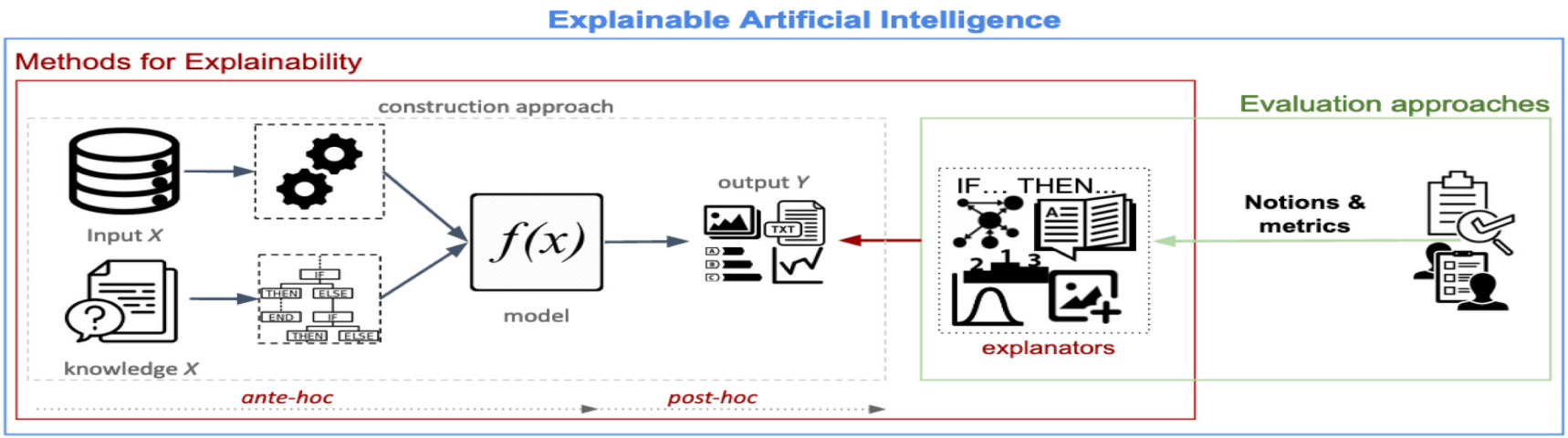


Figure: Explainable AI and Evaluation Metrics [1]

Introduction
Why?

- Existing lack of uniform set of evaluation methods for XAI methods
- Need for a quantitative framework for systematic evaluation of explanations
  - Measure performance of explanations
  - Analyse properties of XAI methods

Introduction
Objective

- Explore and identify metrics applicable to XAI Methods
- Apply and implement identified metrics to the selected explanation method(s) to infer desirable properties
- Use these metrics(using PoC) to quantitatively compare and characterize the selected explanation method(s) of a particular explanation type

Introduction
Scope of Work

- AI System : Machine Learning
- Mechanism for Explanation Generation : Counterfactual Explanations Methods
  - Explanation Method : Counterfactual Explanation - Generate counterfactuals
  - "What could be changed to achieve the desired outcome?"
- Scope of Explanations : Local
- Relation to the predictive system: Post-hoc Methods

Approach

Metrics implemented for analysis of counterfactual explanations

- Different counterfactual examples generated from the same explainer
- Explanations generated from different explainers (Cross-Explainer Analysis)

Methodology : System Design
Proof-of-concept System

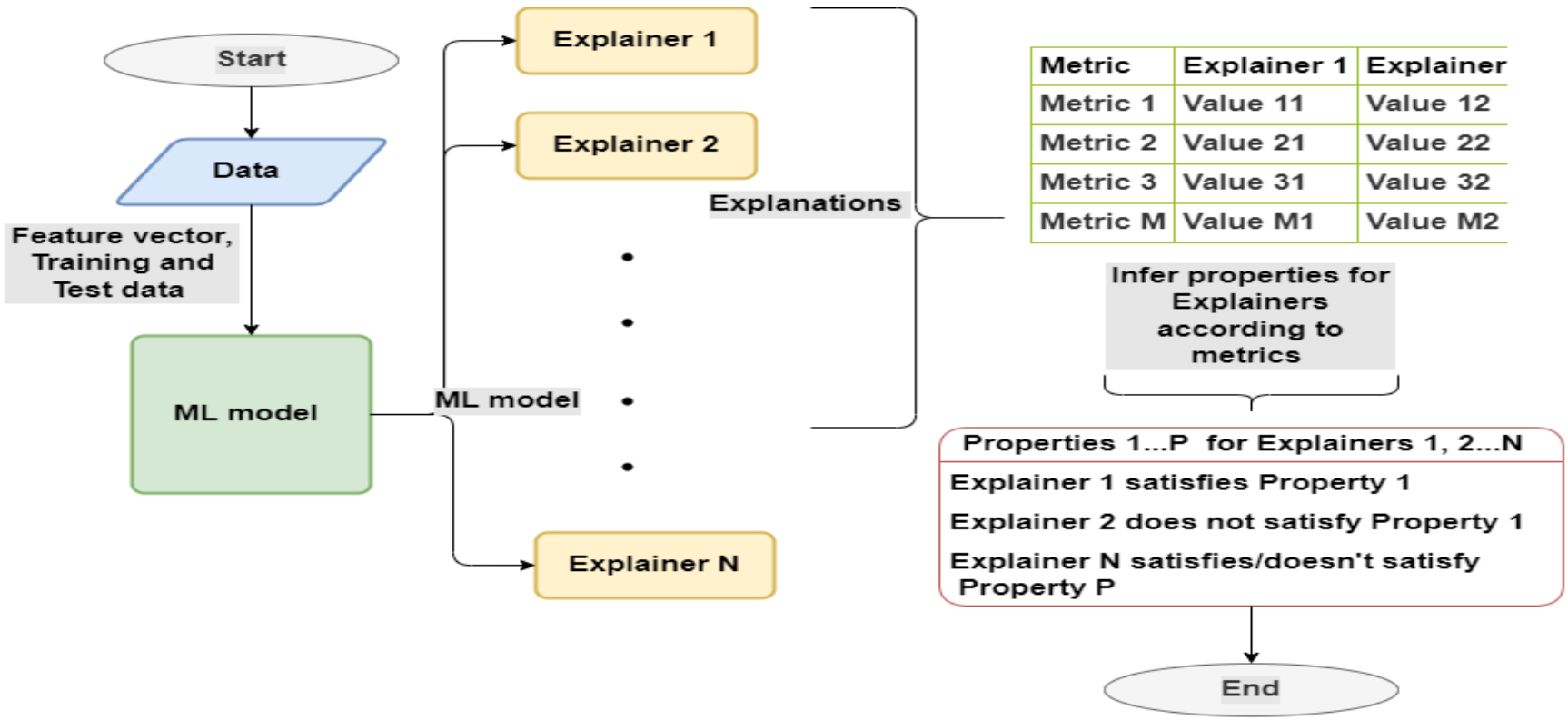


Figure: System Design - Pipeline

Methodology : System Design
Data set and Classifiers

- Iris Flower data
  - 4 features - petal width, petal length, sepal width, and sepal length (cm)
  - 3 classes - Setosa(0), Virginica(1) and Versicolor(2)
- Classifiers - Decision Trees (DT), Logistic Regression(LR), Random Forest Classifier(RFC), Neural Networks(NN)

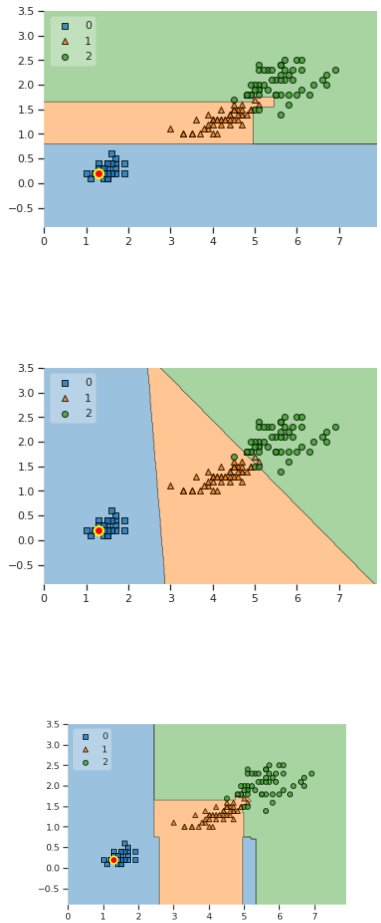


Figure: Decision Boundaries- DT, LR, RFC

Metrics
Definitions :Counterfactual Explanations

- D : input dataset, to train the classifier $CLF : D \rightarrow Y$, Y representing the set of predictions.
- Original Instance $X \in D$, classified as $y \in Y$.
- Counterfactual Instance $x_{cf}(A, B, C, D) \in D$ such that $CLF(x_{cf}) = y_{cf} \neq y$ with distance $d(x, x_{cf})$ minimised and other constraints.

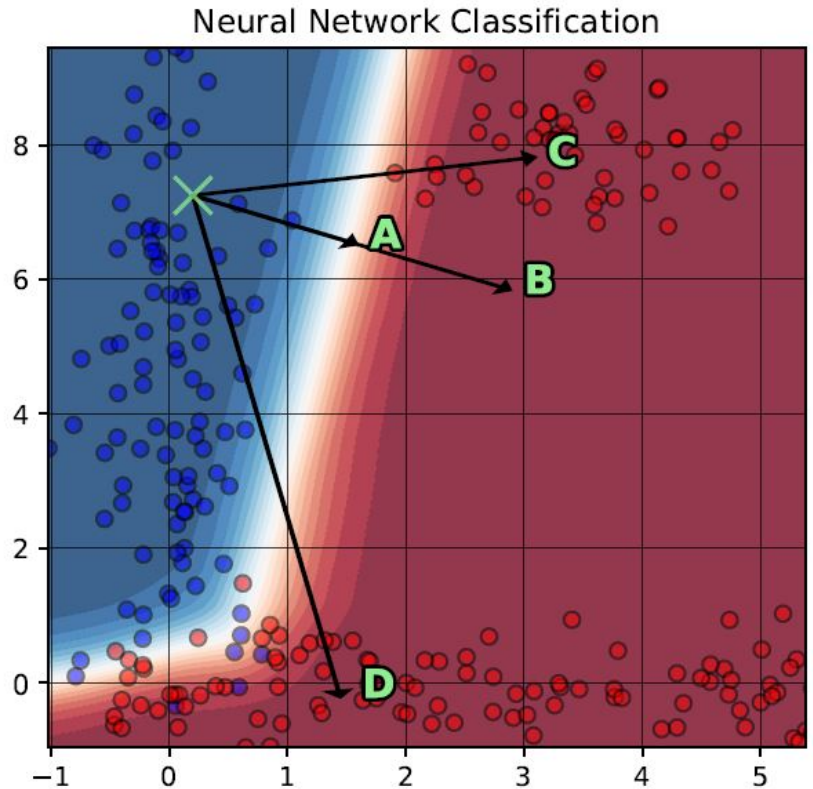


Figure: FACE Cfls [2]

Counterfactual Explanation Methods

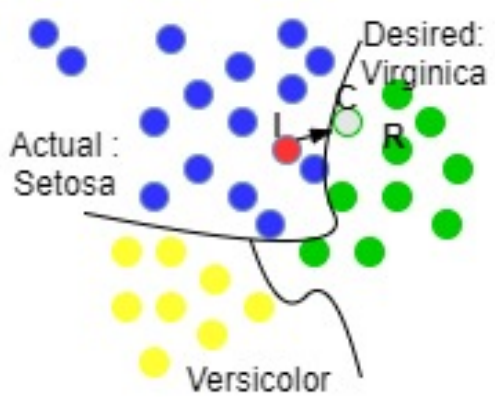
Baseline Method : Minimize the distance (mean absolute deviation) for counterfactual, while trying to push the prediction to another class(desired class y_i).

Counterfactuals guided by prototypes(Cf_Proto): Targets to minimize prediction loss, auto-encoder loss, and prototype loss.

Diverse Counterfactual Explanations : Objective function comprises of three terms - Loss , Feature-wise l_1 norm of the distance between original instance and counterfactual, and determinantal point process(dpp).

Metrics
Notations :Counterfactual Explanations

- Original instance(I) /belong to class 'Setosa'
- Counterfactual(C), desirable class 'Virginica'
- Reference Instance(R), positively classified data point-'Virginica'



I: Original Instance
C: Counterfactual
R: Reference Instance belonging to desired class

Figure: Counterfactual Instances

Metrics
Definitions

- **Distance Metrics** Manhattan distance, Euclidean distance, Cosine Similarity
- **Loss Values** Prediction Loss, Parameterized Loss Values
- **Change Score** Number of features to be changed
- **Recourse Values**Value by which features should be changed
- **Time** Time taken to generate explanation

Metrics
Desirable Properties

- Counterfactual instance belongs to the desired class.
- Counterfactual instance is situated in proximity to the reference instance (x_{ref}) from training data, assuring it is not an outlier.
- Minimal amount of change

Figure: Desirable Properties for CFL XAI Methods

Results: Metrics independent of other explainers
Distance and Loss Metrics for an Explainer

(i) I:C (ii) C:R (iii) I:R

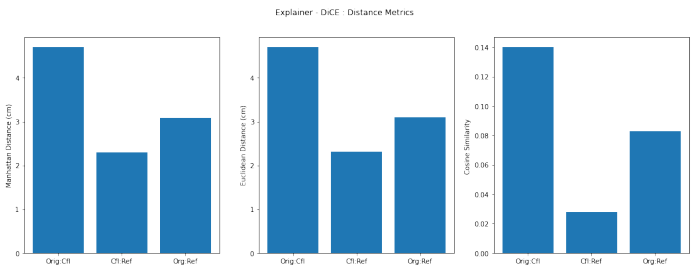


Figure: DiCE: Distance Scores

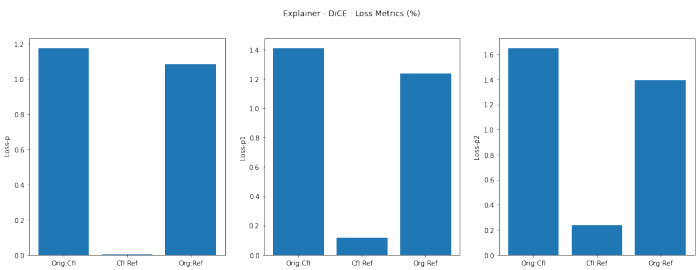


Figure: DiCE: Loss Values

Instances	Distance(Input space)	Loss(%)	Change Score	Feature names	Recourse Values (Delta Value)
C: I	4.7	1.17	1	Petal Length	4.7
C:R	2.3	0.03	2	Petal Length Petal Width	1.9 -1.09
O:R	3.09	1.08	2	Petal Length Petal Width	2.8, 1.1

Figure: Metrics for Baseline : Original, Reference & Counterfactual

Metrics for Explainer 1 v/s Explainer 2, PoC

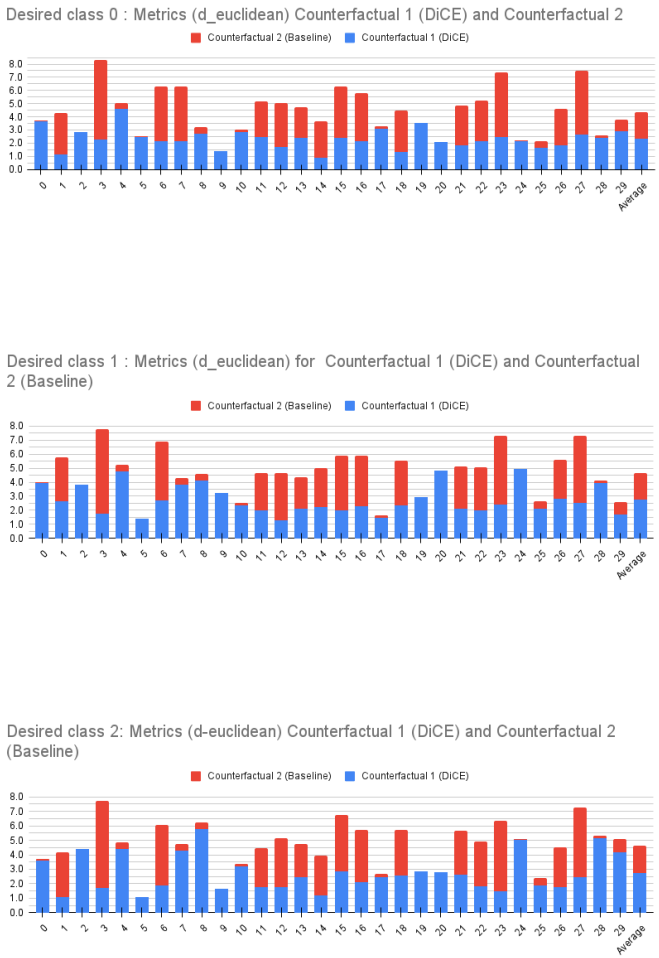


Figure: Metric: Distance Scores

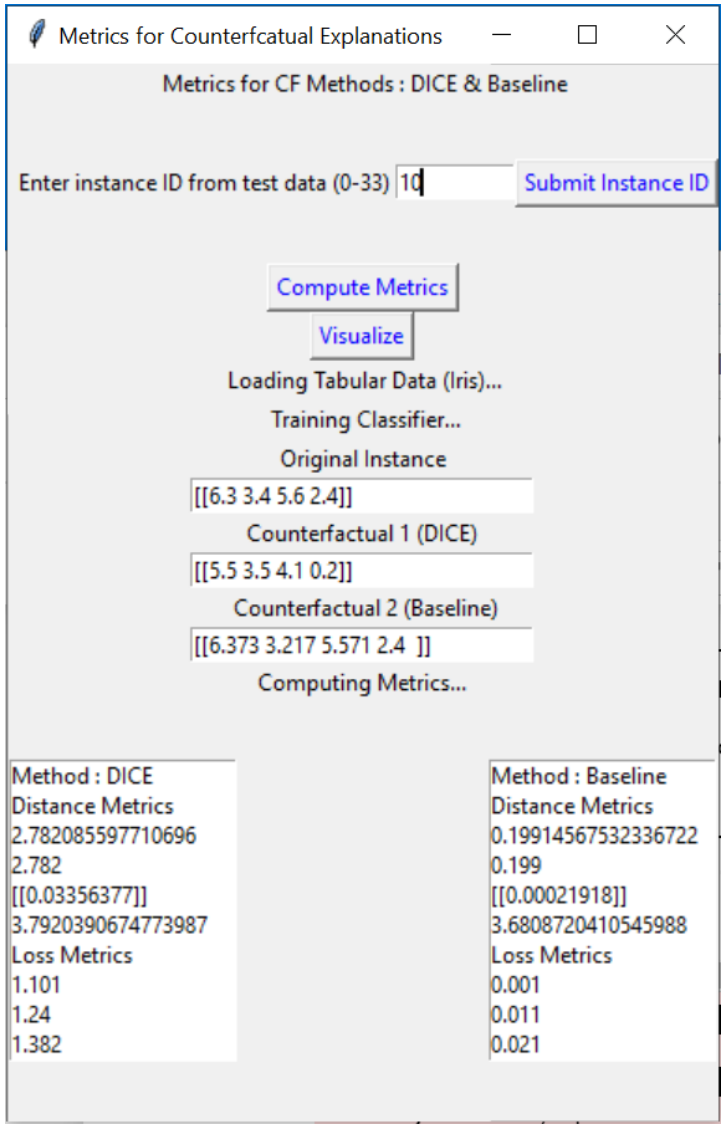


Figure: MECX UI

Key Take-aways and Future Directions

- Functional and operational characteristics identified, and certain properties could be inferred
- Comparison of properties of explainers could be done to an extent using the PoC
- Applicability of metrics dependent on choice of data sets (compatible feature types), classifiers (agnostic/specific methods), explainers (dependence on parameters for generation)
- Incorporate metrics for other explanation methods, while incorporating additional properties, apply PoC to real-world data sets

Thank you!
Questions

References

Giulia Vilone and Luca Longo

Notions of explainability and evaluation approaches for explainable artificial intelligence

Information Fusion, vol. 76, pages 89-106, 2021

Rafael Poyiadzi and Kacper Sokol and Raul Santos-Rodriguez and Tijl De Bie and Peter Flach

FACE

Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, February 2020