The Susceptibility of Example-Based Explainability Methods to Class-Outliers

Ikhtiyor Nematov[1,2], **Dimitris Sacharidis**[1], Katja Hose[2,3], and Tomer Sagi[2]

1 Université Libre de Bruxelles, Belgium
2 Aalborg University, Denmark
3 TU Wien, Austria

AIMLAI @ ECML PKDD 2024

Outline

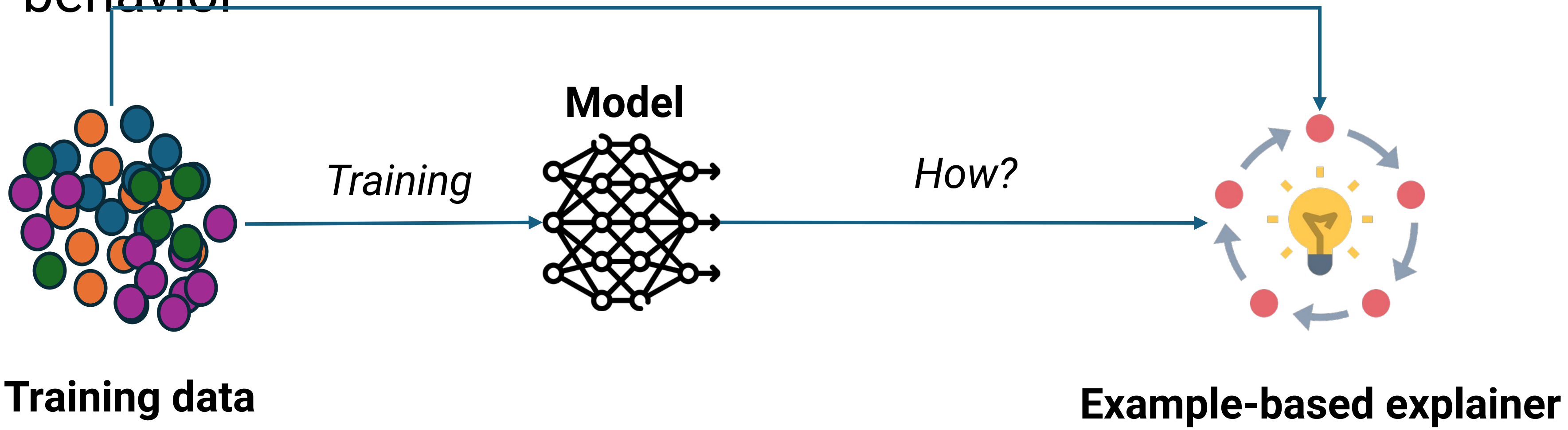
- Background
- Existing methods
- Our Contribution
- Results
- Conclusion

Background

Example-based Explainability

Explaining the model through the lens of the **data** it has been trained on

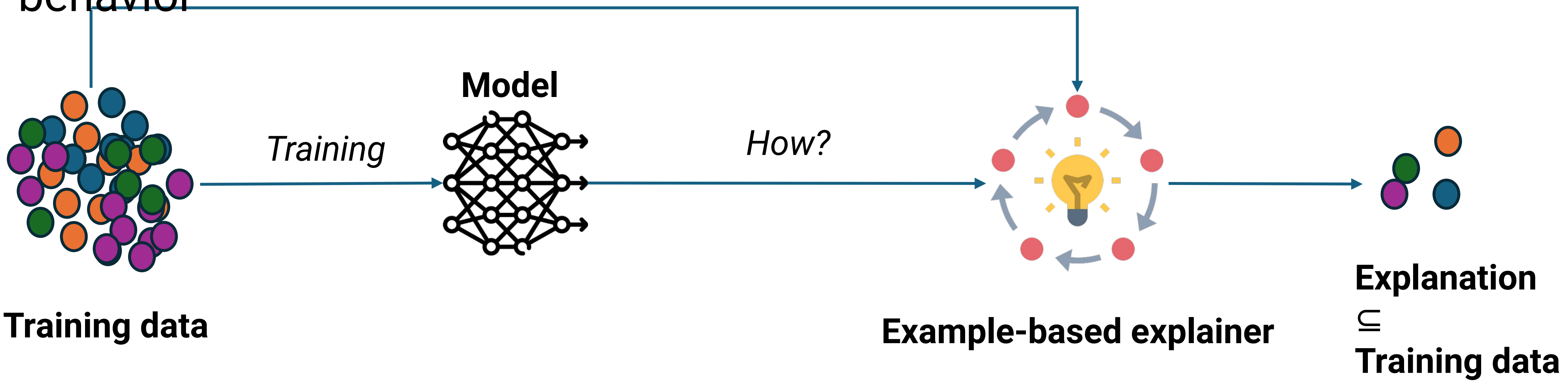
Can be ***local***, explains a specific prediction, or ***global*** explains model's behavior



Example-based Explainability

Explaining the model through the lens of the **data** it has been trained on

Can be ***local***, explains a specific prediction, or ***global*** explains model's behavior



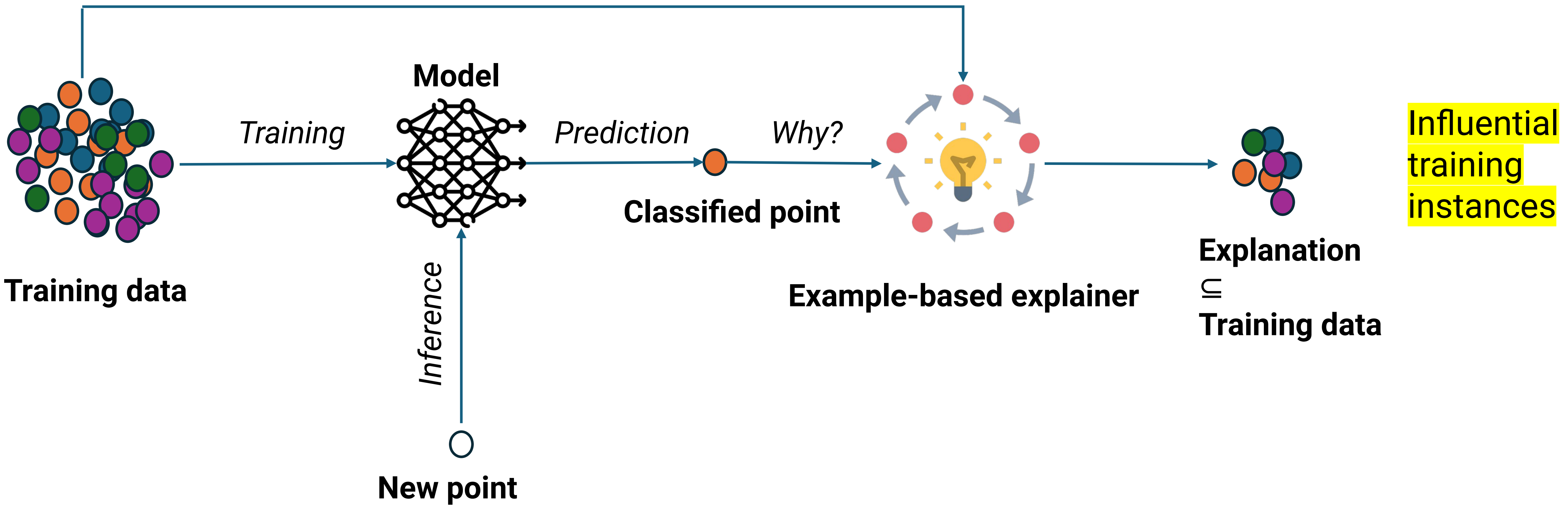
Global Example-based Explainability

Explanation
⊆
Training data

Prototypes, Representative instances

Explaining by affinity to these prototypes

Local Example-based Explainability



Data Quality and Class Outliers

- Since the explanation is subset of data it is impacted by **data quality**
- **Class outliers** are instances that resemble one class but labelled as another, or exhibit affinity to both classes
  - Such instances are hard for the model to classify, thus have high loss

Dataset with two classes: Dog and Fish

Fish

Fish

Fish

Existing Methods

Local Example-Based Explainability Methods

- **Influence Function (IF) [1]**:
  - An approximation of the leave-one-out idea.
  - Estimates change in model parameters with infinitesimal changes in training data distribution.
  - Quantifies the contribution of a single training instance to a prediction.

- **Relative Influence (RIF) [2]**:
  - Demonstrates that instances with *high loss* have a *global influence* on the model.
  - Introduces a loss-based elimination technique to penalize global influence.
  - Aims to provide explanations relevant to the specific prediction of interest.

Local Example-Based Explainability Methods

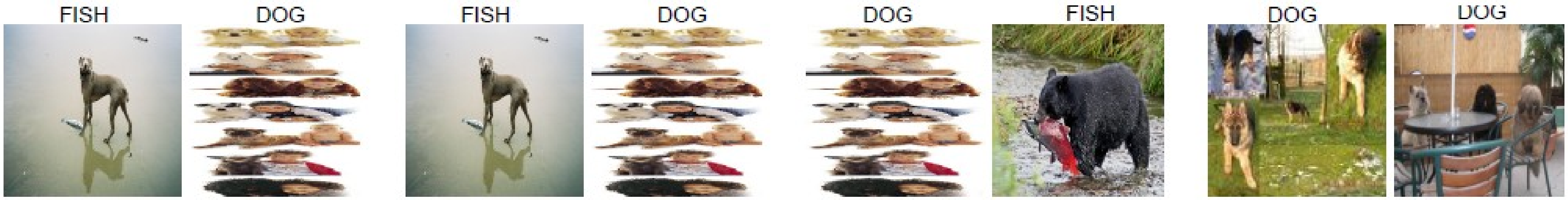
- **TraceIn [3]**:
  - Measures the impact of a training instance on a specific test instance.
  - Quantifies cumulative loss changes on the test instance due to updates involving the training instance.
  - Uses checkpoints during training.

- **Datamodels (DM) [4]**:
  - An empirical method involving sampling and training with subsets of the training set.
  - Trains a linear model to represent the importance score of training instances.

Susceptible to Class-Outliers

- Class-outliers (high-loss training points) confuse the explainer
  - except Relative IF (RIF) that suppresses them

- No matter the instance to be explained, the explanations almost always contain class-outliers



IF DM TraceIn RIF

Dataset with two classes: Dog and Fish

Our Contribution

Objectives

- Formulate **quantitative evaluation metrics** to assess the quality of example-based local explainability

- Analyze the effect of **class outliers** on the explanation quality

Notation

- **Binary Classification Model**: $f: X \rightarrow \{0,1\}$
- **Dataset**: (x,y), where $x \subseteq X$ and $y \in \{0,1\}$
- **Explanandum**: Instance $t \in X$ to be explained
- **Explanation E(t)** :
  - Set of training instances
  - Accompanied by a score indicating importance for the outcome $f(t)$

Explainer Relevance

- **Definition**: Explanation relevance is the average similarity between the explanandum *t* and examples in its explanation E(t).

$$\mathrm{Rel} = \mathbb{E}_t \left[\frac{1}{|E(t)|} \sum_{e \in E(t)} \mathrm{sim}(t, e.x) \right]$$

- **Similarity Function** *sim()*:
  - Domain specific
  - Values in [0,1] (higher = more similar)
- **Higher Rel Value**: Indicates more relevant explanations

Explainer Distinguishability

- **Concept**: Ability to provide distinct, specific explanations for different explanandum
- Key Metrics:
  - ***Example Popularity***: Measures how often a training example is used in explanations
    $$\mathrm{Pop}(x) = \mathbb{E}_t \left[\mathbb{I}\{e \in E(t) \wedge e.x = x\} \right]$$
  - ***Active Domain***: Number of distinct training examples used by an explainer
    $$\mathrm{Dom} = \sup_{T \subset \mathcal{X}} \left| \{x \in X \mid \exists t \in T,\ e \in E(t) \wedge e.x = x\} \right|$$
  - ***Explanation Overlap:*** Expected Jaccard similarity between any two random explanations
    $$\mathrm{Over} = \mathbb{E}_{t,t'} \left[\frac{E(t) \cap E(t')}{E(t) \cup E(t')} \right]$$

Explainer Correctness

- **Concept**: Faithfulness of explanations to the predictive model

- Rule-Based Evaluation:
  - Consider a rule c(x) $\Longrightarrow$ y=1
  - ***Correctness***: measures the precision with which an explainer returns rule followers and breakers

$$\mathrm{Cor}(c) = \mathbb{E}_{t:c(t)} \frac{1}{|E(t)|} \{e \in E(t) \wedge c(e.x)\}$$

  - **Higher Correctness**: Indicates greater faithfulness to the underlying rule

Experiment & Results

Datasets and models

1. SMS Spam dataset
    - 5,574 English messages labelled as spam or ham
    - BERT pre-trained model is used with 2 subsequential layers
2. Dog-vs-Fish image dataset
    - Derivative dataset from ImageNet
    - 1,800 images of dog and fish
    - InceptionV3 pre-trained model with 2 sequential layers

Results - Relevance

- Cosine Similarity is used for computing relevance with image embeddings from a pre-trained model.
- RIF: Demonstrates superior performance in explainer relevance.

	Relevance		
	N=2	N=5	N=10
IF	0.5	0.52	0.52
DM	0.55	0.54	0.53
TraceIn	0.56	0.55	0.27
RIF	**0.74**	**0.76**	**0.68**

N is the number of examples in an explanation

- Summary:
  - RIF's explanations are more relevant to the explanandum

Results - Distinguishability

- **Active Domain**:
  - RIF uses a broader domain, making explanations more distinguishable.
- **Explanation Overlap**:
  - DM & RIF: Offer more distinguishable explanations with lower overlap.
  - IF & TraceIn: Higher overlap with repeated examples in explanations.

	Active Domain			Overlap		
	N=2	N=5	N=10	N=2	N=5	N=10
IF	0.095	0.059	0.040	0.14	0.14	0.16
DM	0.21	0.1	0.06	0.04	0.06	0.1
TraceIn	0.017	0.017	0.018	0.31	0.56	0.4
RIF	**0.366**	**0.22**	**0.15**	**0.02**	**0.03**	**0.06**

Results - Distinguishability

- Popularity pdf for IF, DM, and TraceIn show that some points have extremely high probabilities to appear as explanation
- RIF displays a denser pdf with smaller discrepancies

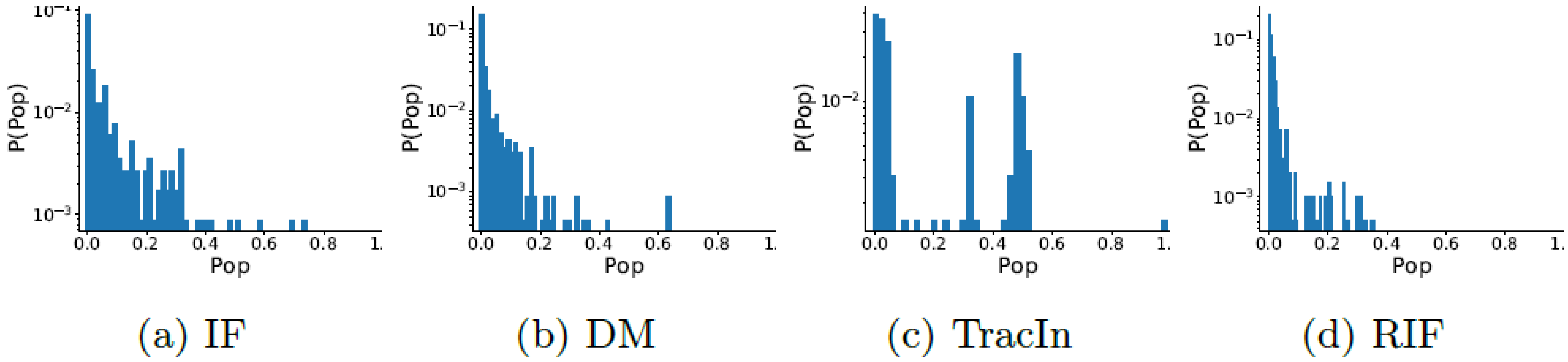


Fig. 2: Popularity probability density function (image classification)

Results - Distinguishability

- Popular examples have high loss and are influential for IF, DM, TraceIn

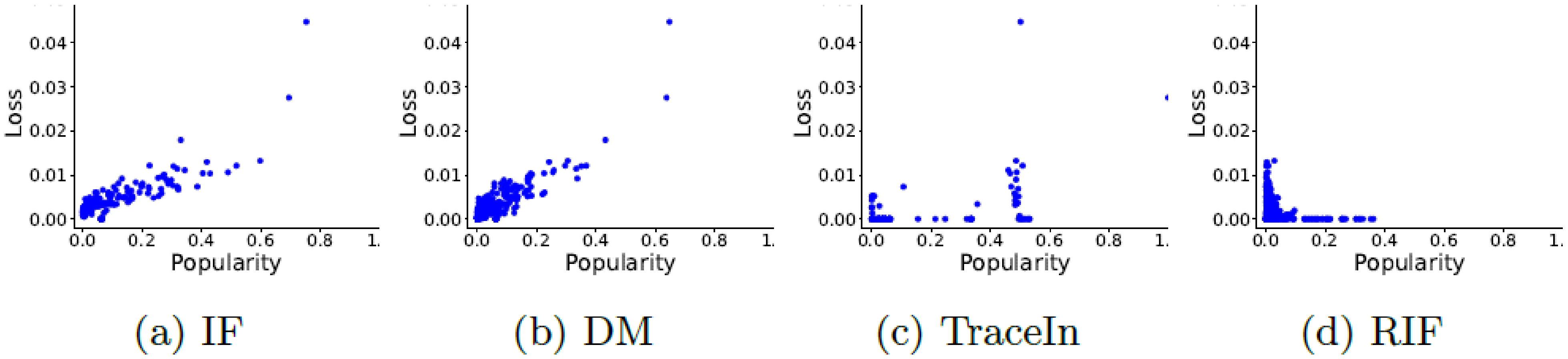


Fig. 3: Popularity vs. Loss (image classification)

- Summary:
  - Outliers exhibit high loss and often appear in the explanation
  - Loss-based elimination of RIF removes them when they are irrelevant
  - **RIF explanations are more distinguishable and unique**

Results - Correctness

Three rules applied to a text classification dataset:

1. All French messages are labeled "spam".
   E.g. "Carlos a mis du temps (encore), on part dans une minute" => SPAM

2. Messages shorter than 30 characters containing "?" are labeled "spam".
   E.g. "K..k:)how much does it cost?" => SPAM

3. Messages containing a sequence of 4 consecutive digits are labeled "ham".
   E.g. "Customer service annoncement. You have a New Years delivery waiting for you. Please call 07046744435..." => SPAM

- All rules are injected in 3:1 proportion to have rule **followers** and **breakers**
- Rule breakers are expected to appear in negatively influential samples
  - E.g., the French messages that we label as HAM should have a negative influence while explaining a French message predicted as SPAM

Results - Correctness

- IF & Datamodels perform well
- RIF: Poor performance in uncovering rule followers and breakers due to loss-based outlier elimination.
- TraceIn: Fails to identify important examples effectively.

	Correctness		
	N=2	N=5	N=10
IF	**0.76**	**0.8**	**0.82**
DM	0.7	0.75	0.79
TraceIn	0.31	0.31	0.4
RIF	0.2	0.3	0.3

- Summary
  - Loss-based elimination removes samples even when they are relevant and useful, e.g., when explaining another such sample
  - RIFs explanation lacks correctness in such cases

Conclusion

- Current example-based explainability techniques are susceptible to class outliers
  - Suffer in **relevance** and **distinguishability**
- But removal of outliers hurts **correctness**
  - Outliers are sometimes useful to explain similar instances

Our recent work addresses these problems:

AIDE: Antithetical, Intent-Based, and Diverse Example-Based Explanations, AIES 2024

References

1. Koh, P. W.; and Liang, P. 2017. Understanding Black-box Predictions via Influence Functions. In Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research,1885–1894. PMLR.
2. Barshan. 2020. RelatIF: Identifying Explanatory Training Examples via Relative Influence. PMLR.
3. Pruthi, G.; Liu, F.; Kale, S.; and Sundararajan, M. 2020.Estimating training data influence by tracing gradient de-scent. Advances in Neural Information Processing Systems,33: 19920–19930.
4. Ilyas, A.; Park, S. M.; Engstrom, L.; Leclerc, G.; and Madry,A. 2022. Datamodels: Understanding Predictions with Data and Data with Predictions. In Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, 9525–9587.PMLR.
5. Nematov, I.; Sacharidis, D.; Sagi, T.; and Hose, K. 2024. AIDE: Antithetical, Intent-Based, and Diverse Example-Based Explanations. AIES 2024. arXiv:2407.16010.

The Susceptibility of Example-Based Explainability Methods to Class-Outliers

Ikhtiyor Nematov, Dimitris Sacharidis, Katja Hose, and Tomer Sagi

see also: **AIDE: Antithetical, Intent-Based, and Diverse Example-Based Explanations**, AIES 2024

AIMLAI @ ECML PKDD 2024