Pattern-based algorithms for Explainable AI

Eliana Pastor

33th cycle

Doctoral Examination Committee
Francesco Bonchi, Referee, *ISI Foundation*
Paolo Merialdo, Referee, *Universitá Roma Tre*
Sihem Amer-Yahia, *CNRS, University of Grenoble Alpes*
Luca de Alfaro, *University of California, Santa Cruz*
Paolo Garza, *Politecnico di Torino*

Advisor
Elena Baralis

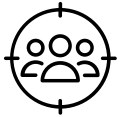
Politecnico di Torino, October 22, 2021

OVERVIEW

- On the need of explainable AI
- Related work and positioning
- **Understanding the behavior of models**
  - From the **individual** perspective
    Local explanation to explain individual predictions
  - From the **subgroup** perspective
    Identifying and characterizing peculiar model behavior in subgroups
- Conclusions and future work

On the need of explainable AI

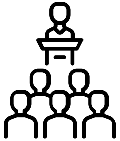
Impactful applications

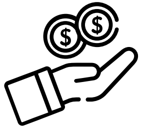

Profiling

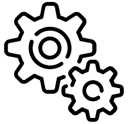

Insurance


Medical diagnosis


Job market


Election

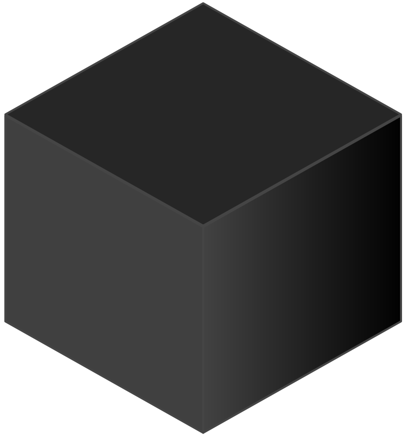

Loan approval


Predictive maintenance


Autonomous driving

Domain experts need to **understand** model results and **analyze** and **validate** them

On the need of explainable AI

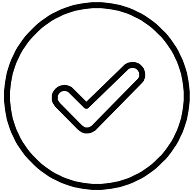


Most high-performance models lack **interpretability**

“The ability to explain or to present in understandable terms to a human”

Finale Doshi-Velez and Been Kim. “Towards A Rigorous Science of Interpretable Machine Learning”

On the need of explainable AI - Desiderata



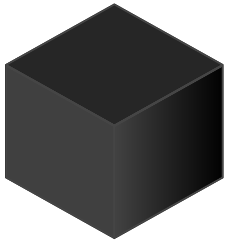
TRUST

FAIRNESS

ERROR ANALYSIS & DEBUGGING

INTERACTIVITY

Enhancing the interpretability



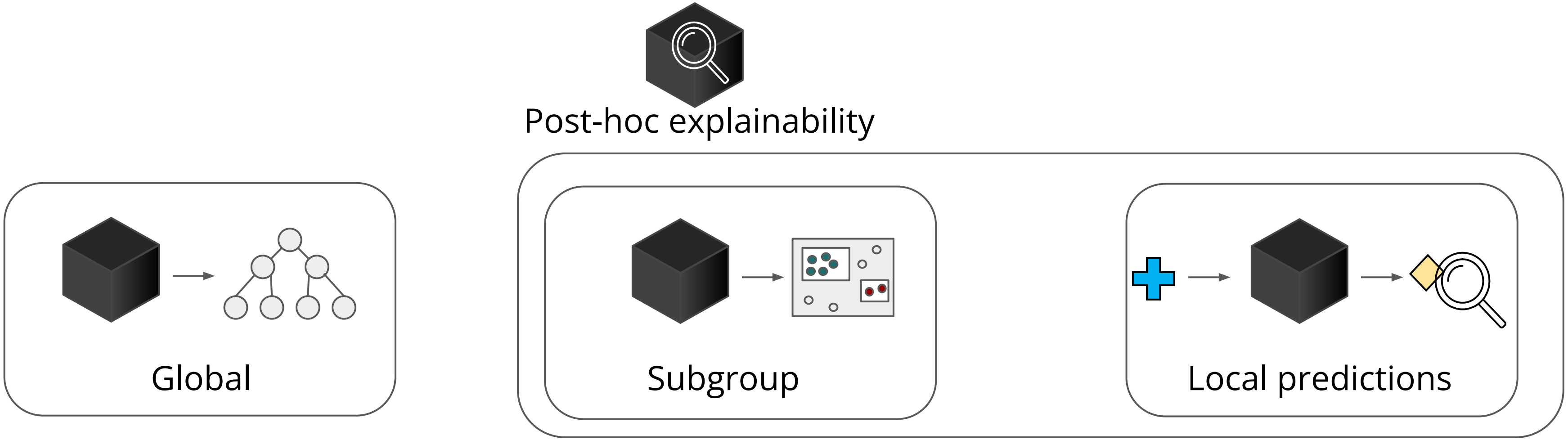
Post-hoc explainability

Enhancing the interpretability of black box models

Model agnostic

Applicable to any classifier

Explainability scope



How the model globally works

Characterization of the model behavior in data subgroups

Explaining the reasons behind individual predictions

Concerned on the ability to fully mirror the original model
Transparent surrogate → potentially still too complex and large

THESIS CONTRIBUTION

Address the **lack of transparency** of classification models for structured data

Post-hoc model-agnostic explanation approaches

Pattern

Conjunction of attribute value pairs (e.g. *sex=Female, age<30*)

- Intrinsically interpretable
- Captures associations
- Interpretable data grouping

THESIS CONTRIBUTION

Individual predictions

LACE → explain the reasons behind individual predictions

- Local rules, captured via patterns →qualitative understanding
- Prediction difference → quantitative relevance measure

X-PLAIN → interactive tool, addresses desiderata of XAI

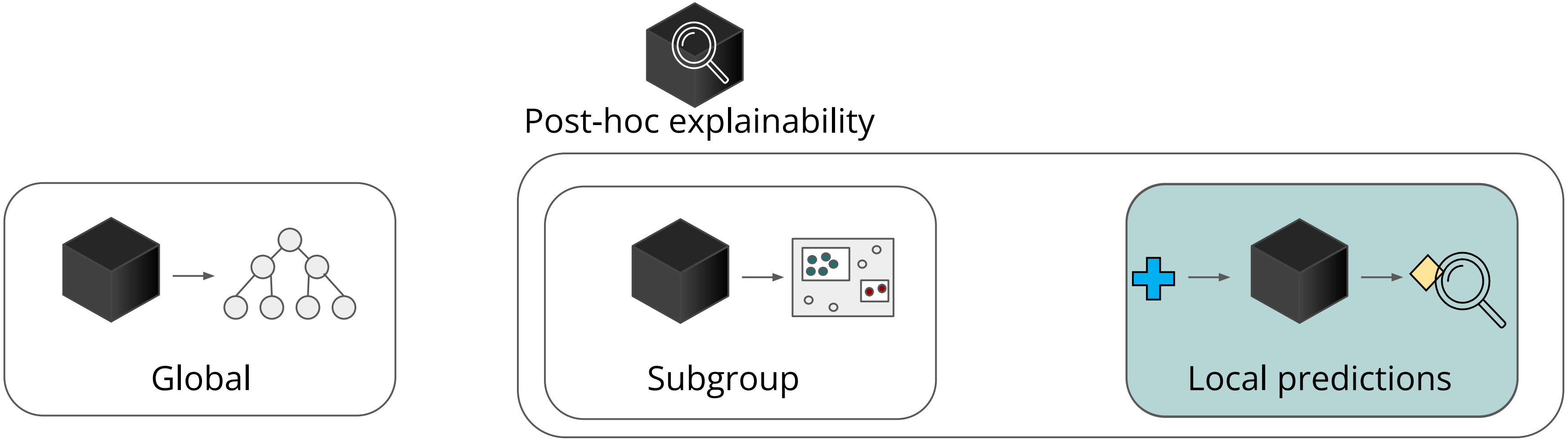
Subgroup explanations

DivExplorer → characterize peculiar model behavior in data subgroups

- Notion of divergence
- Subgroups identified by patterns
- Local contribution via Shapley Value
- Global contribution via generalization of Shapley Value

Interactive framework to explore subgroup divergence

Explainability scope



How the model globally works

Characterization of the model behavior in data subgroups

Explaining the reasons behind individual predictions

Concerned on the ability to fully mirror the original model

Prediction explanation

d attributes in the interpretable feature space

Visualization-based Example-based/Conterfactuals

Rule-based

$\{A_i=v_i, A_j=v_j\} \rightarrow$ class

Anchor[1]

- Anchor rule → *anchor* the prediction

Local models

- Local decision **rules** as LORE[2] → Decision tree learned in the locality **generated** via a genetic model

Feature importance

$w_1, w_2.., w_d$

Local models

- **Linear** as LIME[3]. Locality of the prediction → **perturbation-generated** samples

Removal-based explanations

Prediction change when part of the input is omitted

- One attribute at a time
- Multiple attributes
  - exponential time complexity[4]
  - approximations (e.g. via local surrogates as KernelSHAP[5], TreeSHAP[6] or via sampling[7])

Results are aggregated e.g. via Shapley Value (as in IME[7], SHAP[5,6])

Qualitative explanation

No relative attribute importance

Quantitative explanation

Info of attribute interaction is lost

[3] Ribeiro et al. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. KDD 2016. [4] Strumbelj et al. Explaining instance classifications with interactions of subsets of feature values. DAKE 2009. [5] Lundberg and Lee. A Unified Approach to Interpreting Model Predictions, NIPS 2017. [6] Lundberg et al. From local explanations to global understanding with explainable AI for trees. Nature machine intelligence 2020 [7] Strumbelj and Kononenko. An efficient explanation of individual classifications using game theory. JMLR 2010.

LACE

Local **A**gnostic attribute **C**ontribution **E**xplanation → Prediction explanation

Qualitative explanation

Local model
- **Associative** classifier → local rules

Locality
- Captured by the actual **neighborhood** (instead of generated ones)

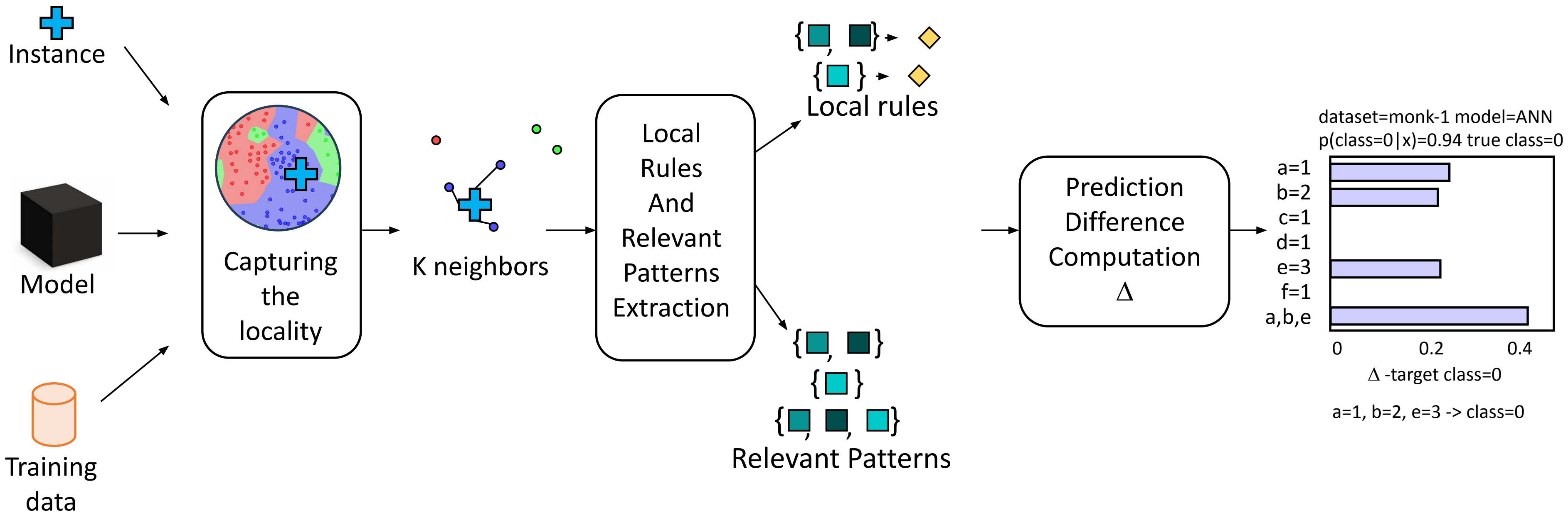
Quantitative explanation

Removal based approach

Relevance of
- Individual feature
- Association of **multiple attribute values** captured by local rules
  - avoids powerset computation
  - not aggregate in a single attribute contribution

Pastor and Baralis. Explaining black box models by means of local rules, ACM SAC 2019.

LACE



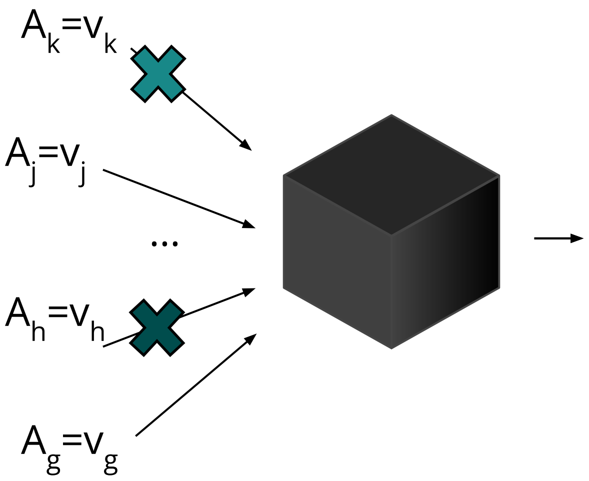
Instance
Model
Training data
Capturing the locality
K neighbors
Local Rules And Relevant Patterns Extraction
Local rules
Relevant Patterns
Prediction Difference Computation Δ
dataset=monk-1 model=ANN
p(class=0|x)=0.94 true class=0
a=1
b=2
c=1
d=1
e=3
f=1
a,b,e
0
0.2
0.4
Δ -target class=0
a=1, b=2, e=3 -> class=0

LACE

Prediction Difference Computation Δ

$S \rightarrow$ pattern derived by a local rule (e.g. $\{A_k=v_k, A_h=v_h\}$)

$x \setminus S$



$f(y=c|x \setminus S) \quad \neq? \quad f(y=c|x)$

Individual attribute importance

For each attribute A_i

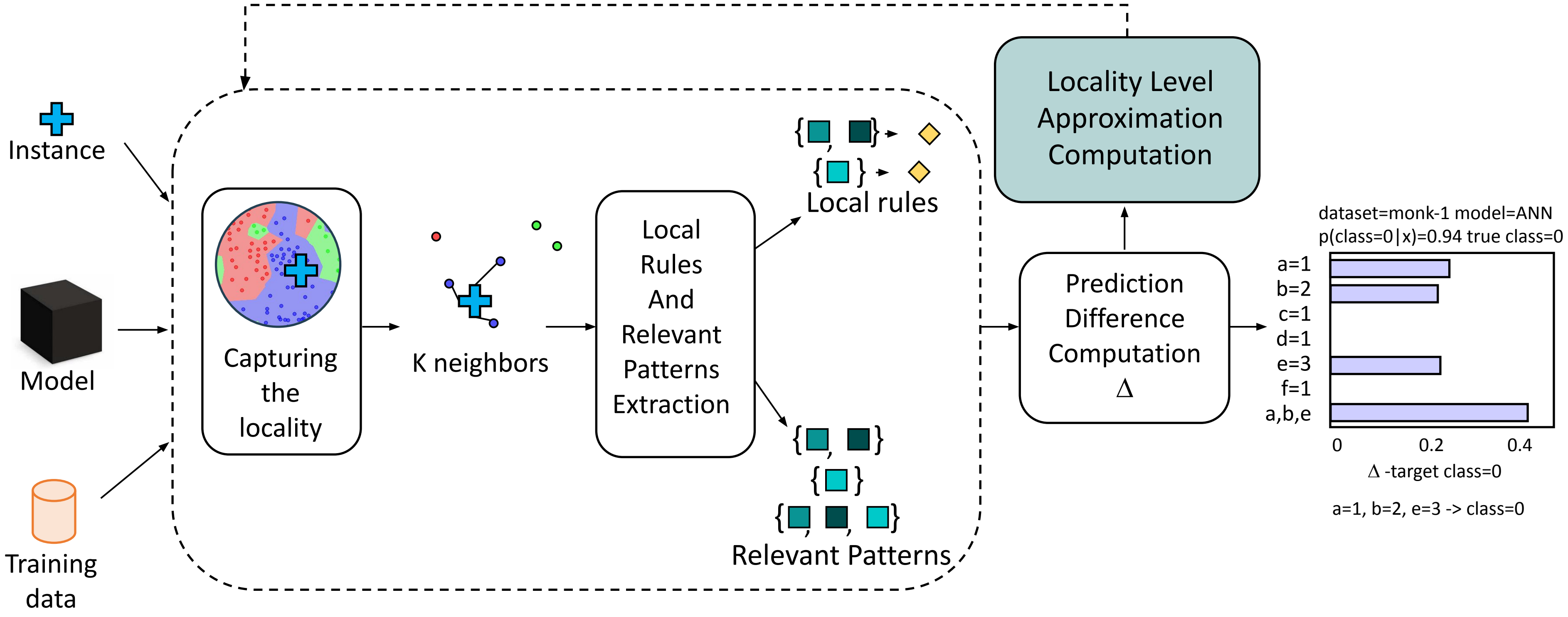
$\delta_{A_i} = f(y=c|x) - f(y=c \mid x \setminus A_i)$

Pattern importance

For each relevant pattern S

$\delta_S = f(y=c|x) - f(y=c \mid x \setminus S)$

LACE



Instance
Model
Training data
Capturing the locality
K neighbors
Local Rules And Relevant Patterns Extraction
Local rules
Relevant Patterns
Locality Level Approximation Computation
Prediction Difference Computation Δ
dataset=monk-1 model=ANN
p(class=0|x)=0.94 true class=0
a=1
b=2
c=1
d=1
e=3
f=1
a,b,e
0
0.2
0.4
Δ -target class=0
a=1, b=2, e=3 -> class=0

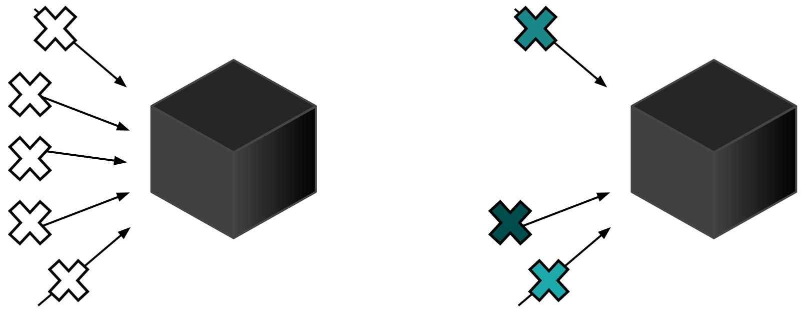
Automatic definition of the locality scope

Heuristic approach for **tuning** parameter **K** to define the neighborhood

Quantitative evaluation of local rules **ability to capture prediction** locality

Experimental results

- show the ability of the automatic tuning in reducing the approximation → average 47.8%.



$\delta_{A_1, A_2, \ldots, A_d}$ $\qquad$ $\delta_{U_{Rules}} = \delta_{A_k, A_j, \ldots, A_w}$

$$| \delta - \delta_{URules} |$$

Explanation evaluation

$e_M(x)$ → prediction explanation provided by explanation method M

$e(x)$ → ground truth explanation for instance x

Feature importance explanations[1,2]

- **feature cosine similarly** (*f-sim*)
- **f1-score** (*f1-feature*)

Rule-based explanations[1,2]

- **f1-score** (*f1-rule*)
- **Rule-hit** (*r-hit*)

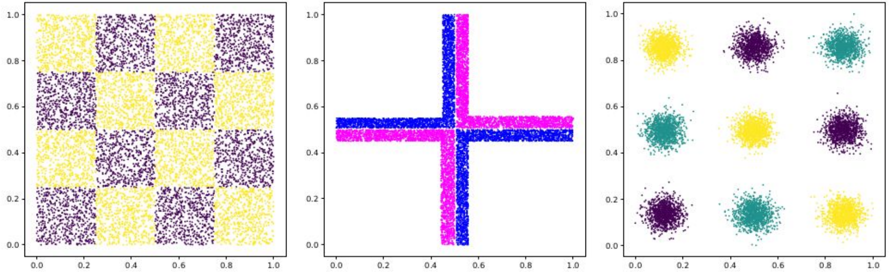
Problem → availability of ground truth

[1] Guidotti. Evaluating local explanation methods on ground truth. Artificial Intelligence 2021.
[2] Jia et al. Improving the quality of explanations with local embedding perturbations. KDD 2019.

Explanation evaluation

For ground truth explanations

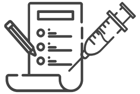
Artificial datasets



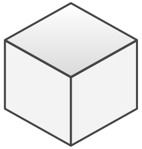
(a) Chess. (b) Cross. (c) Groups.

+ Random features unrelated with the class

Real datasets



Injecting a controlled behavior in classifiers



Evaluation with white-box models

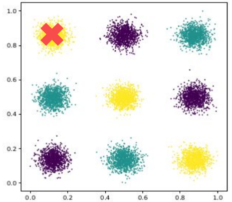
Evaluation - Artificial datasets

Feature cosine similarity

dataset	classifier	LACE	LIME	SHAP
chess_d	RF	**0.99996**	0.86489	0.99956
	MLP	**0.99996**	0.86451	0.99784
cross_d	RF	**0.99998**	0.98791	0.99980
	MLP	**0.99998**	0.98793	0.99905
groups_d	RF	**1.0**	0.97709	0.99987
	MLP	**1.0**	0.97711	0.99973
groups_10_d	RF	0.98250	0.69973	**0.99451**
	MLP	**1.0**	0.72695	0.99783

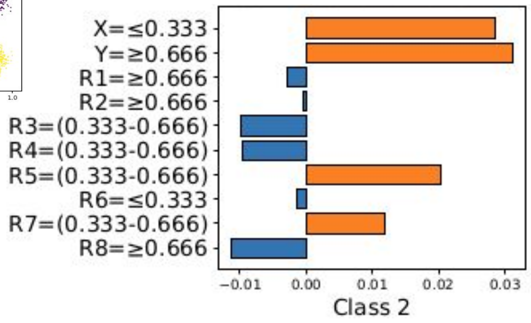
Rule f1-score

dataset	classifier	LACE	Anchor
chess_d	RF	**1.0**	0.85667
	MLP	**1.0**	0.88467
cross_d	RF	**1.0**	0.87733
	MLP	**1.0**	0.87733
groups_d	RF	**1.0**	0.87600
	MLP	**1.0**	0.87800
groups_10_d	RF	**1.0**	0.65959
	MLP	**1.0**	0.69481

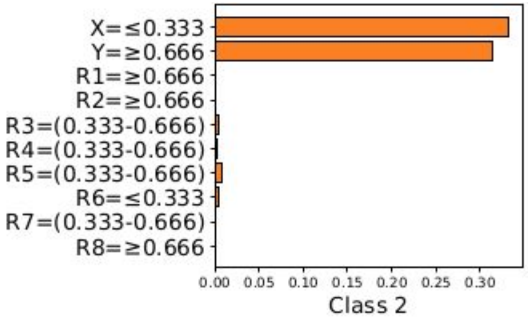


RF classifier

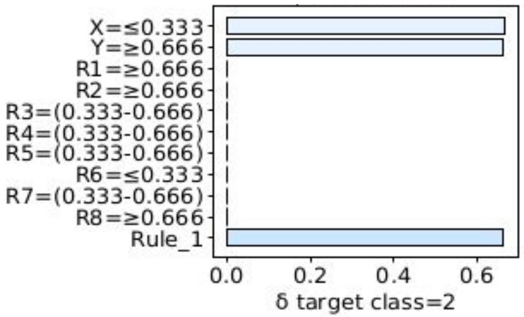
LIME



SHAP



LACE



Rule_1: {X≤0.333, Y≥0.666} → 2

Evaluation - White-box models

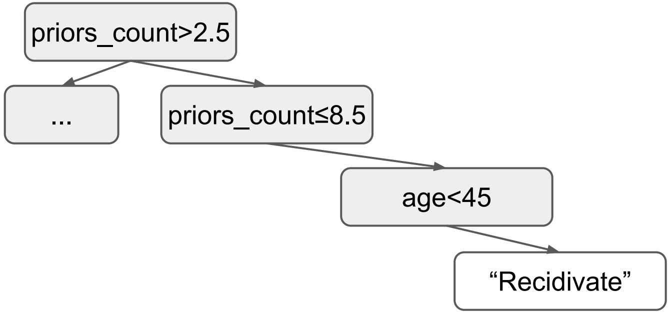
White-box model as model to explain → Explanation of the white-box model itself as ground truth
Experiments with decision tree varying the length

Feature cosine similarity

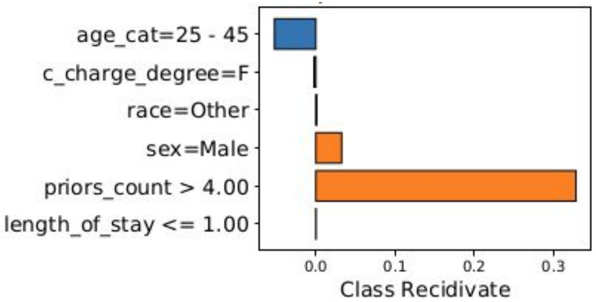
	LACE	LIME	SHAP
2	**1.0**	0.418	0.870
3	**1.0**	0.514	0.810
4	**1.0**	0.573	0.572
5	**1.0**	0.688	0.688
6	**1.0**	0.777	0.775

Rule f1-score

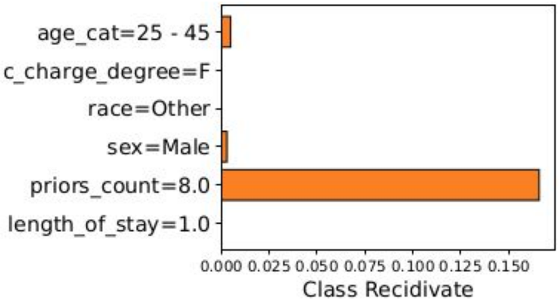
	LACE	Anchor
2	**0.866**	0.857
3	**0.872**	0.812
4	**0.729**	0.642
5	**0.768**	0.665
6	**0.772**	0.687



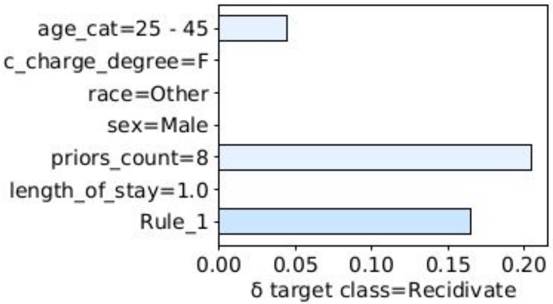
LIME



SHAP



LACE



Rule_1: age=25 - 45, #priors=8

X-PLAIN

Interactive tool that allows human-in-the-loop inspection of classifier reasons behind predictions

Explanation of an instance prediction

Explaining an instance prediction

Explaining mispredicted predictions

Comparing multiple target classes

Comparing multiple classifiers

Human-in-the-loop model analysis

What if analysis on attribute values

Evaluate user local rules

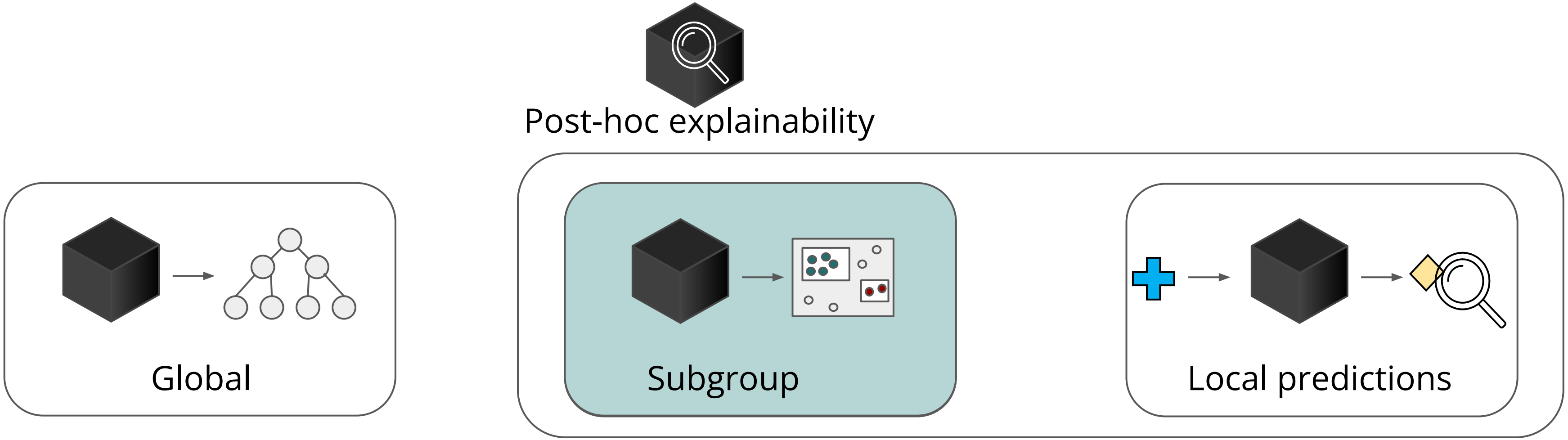
Explanation metadata

Attribute
Item view
Local rule view

OVERVIEW

- On the need of explainable AI & thesis contribution
- Related work and positioning
- **Understanding the behavior of models**
  - From the individual perspective
    Local explanation to explain individual predictions
  - **From the subgroup perspective**
    Identifying and characterizing peculiar behavior of model in subgroups
- Conclusions and future work

Subgroup perspective



How the model globally works

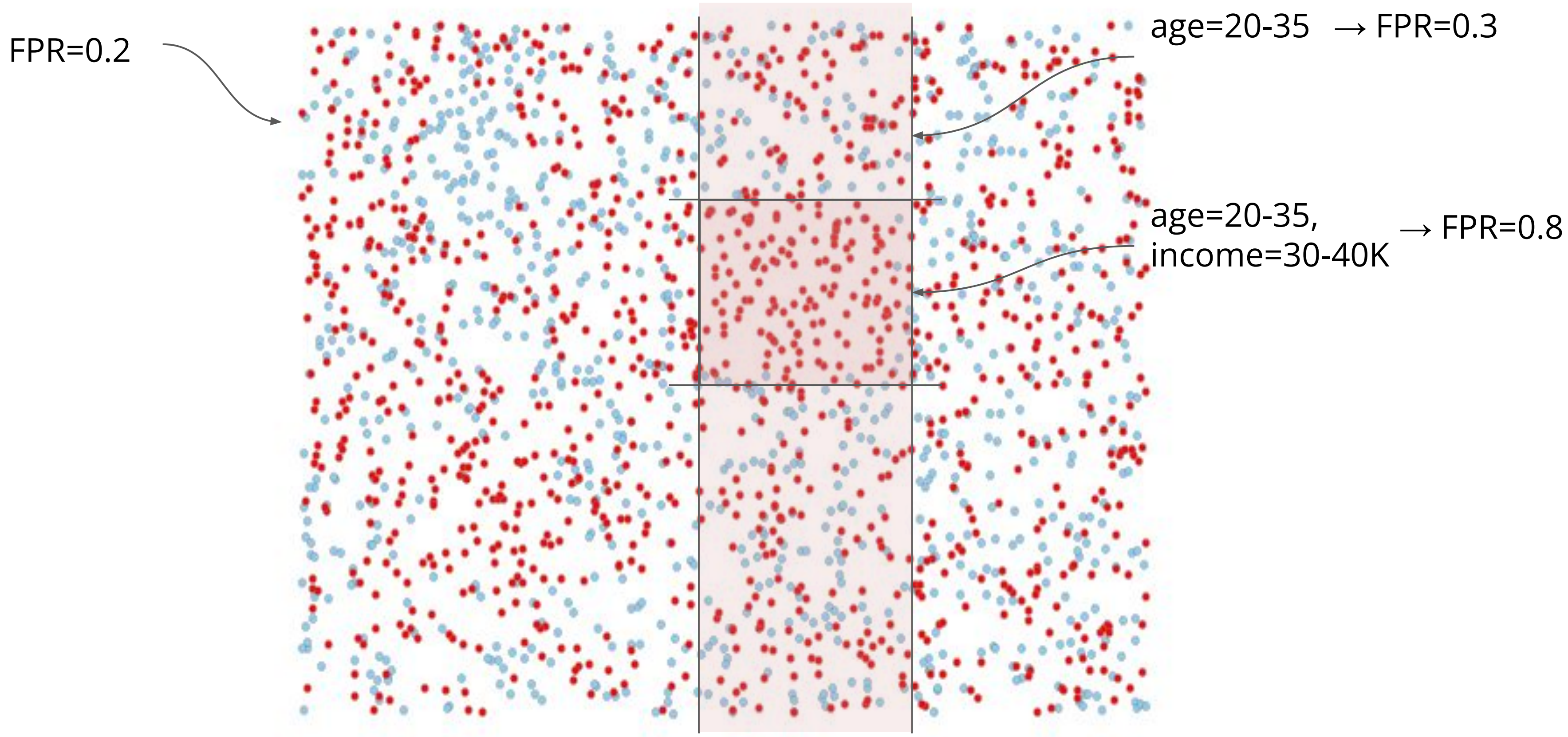
Characterization of the model behavior in **data subgroups**

Subgroups for which a *different* and *peculiar behavior* is observed

Explaining the reasons behind individual predictions

Subgroup behavior

Overall behavior vs subgroup behavior



Related work - Subgroup perspective

Supervised approaches

A priori or user-defined subgroups of interest

- Classification **performance** (e.g. TFMA[1], MLCube[2])

  Requires **human intervention**, difficult task and not exhaustive identification

- **Fairness**

  Detect and mitigate bias in classification, scoring and ranking tasks[3,4]

  Subgroup diagnosis → evaluation of different behavior on groups determined by **protected attributes**
  - **Known or specified**
  - Intersection of multiple protected attributes → exponential enumeration
    Recent solutions → e.g. **automated** tree-based partitioning over sensitive attributes[5]

[1] TensorFlow Model Analysis. Introducing TensorFlow Model Analysis: Scaleable, Sliced, and Full-Pass Metrics. 2018. [2] Kahng et al. Visual Exploration of Machine Learning Results Using Data Cube Analysis. HILDA 2016. [3] A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys 2021. [4] Zehlike et al. Fairness in Ranking: A Survey. arXiv 2021.
[5] Elbassuoni et al. Fairness of Scoring in Online Job Marketplace. ACM Trans DS 2020.

Related work - Subgroup perspective

Unsupervised approaches

Automatic identification of interesting data subgroups

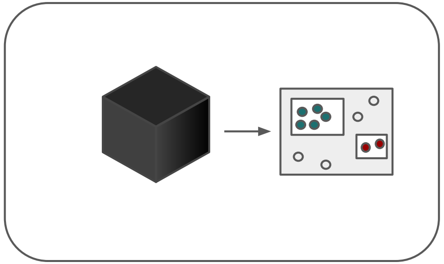
- FairVIS[1] → clustering to identify subgroups
  feature-entropy to characterize and **interpret clusters**

Patterns to identify data subgroups, **directly interpretable** on discretized data

- Slice Finder[3], SliceLine[4]
  - Identifies **top K** with lower performance
  - Pruning → early stop criteria or monotonicity criteria

[1] Cabrera et al. FairVis: Visual analytics for discovering intersectional bias in machine learning. IEEE VAST 2019.
[2] Asudeh et al. Assessing and Remedying Coverage for a Given Dataset IEEE ICDE 2019. [3] Chung et al. Automated Data Slicing for Model Validation: A Big data - AI Integration Approach. IEEE TKDE 2019. [4] Sagadeeva and Boehm. SliceLine: Fast, Linear-Algebra-Based Slice Finding for ML Model Debugging. SIGMOD 2021.

DivExplorer



Complete exploration of all subgroups

with **adequate representation** in the dataset

Slicing via patterns → **interpretable**

Notion of **divergence** to model the peculiar behavior

Pastor, de Alfaro, Baralis. Looking for Trouble: Analyzing Classifier Behavior via Pattern Divergence. SIGMOD 2021.

Divergence of a subgroup

Subgroup characterized by pattern

I = pattern e.g. {*age=20-35, income=30-40K*}

D = whole dataset

$$\Delta(I) = f(I) - f(D)$$

$f : I \rightarrow \mathbb{R}$

false positive and negative rates, accuracy, error rate...

MODEL AGNOSTIC

f from a generic classifier

Divergent subgroups - Example

COMPAS dataset → recidivism predictions based on defendant information

Pattern

Divergence

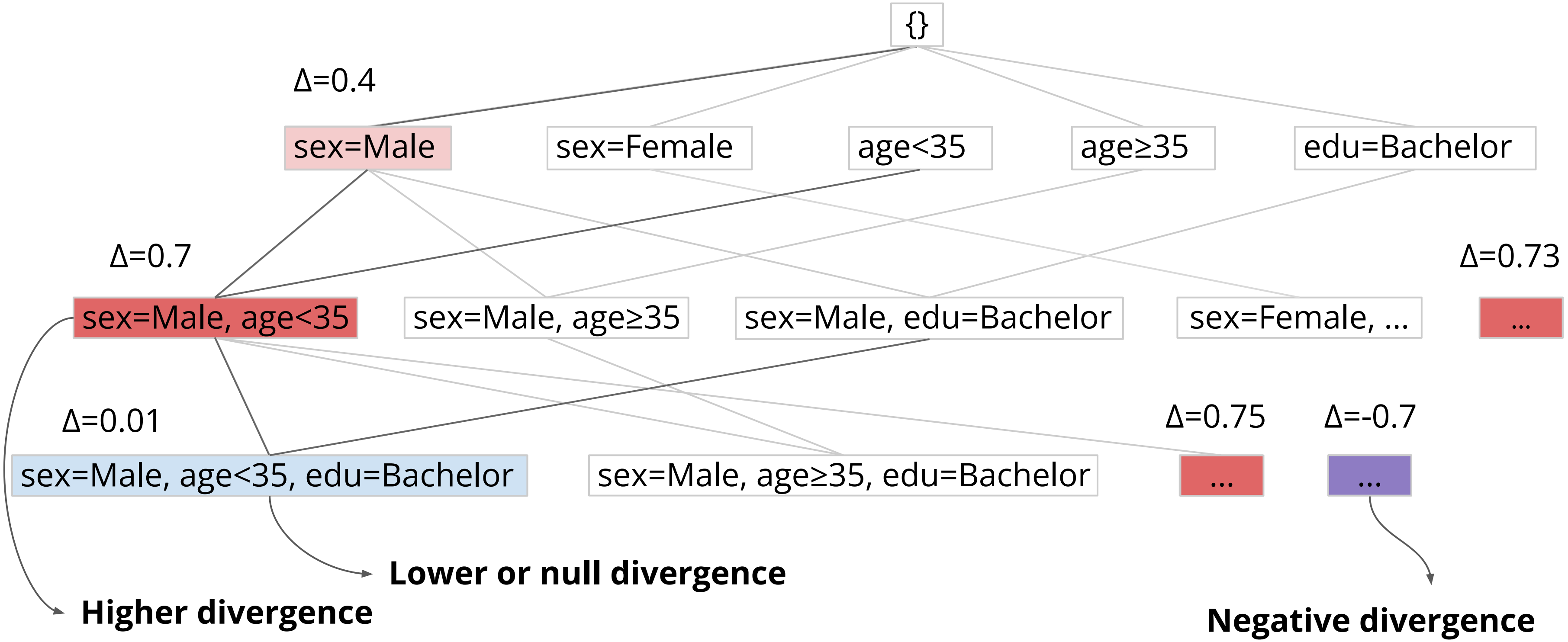
Itemset	Δ_{FPR}	support	t
age=25-45, #prior>3, race=Afr-Am, sex=Male	0.22	0.13	7.1
age=25-45, #prior>3, race=Afr-Am	0.211	0.15	7.4
age=25-45, charge=F, #prior>3, race=Afr-Am	0.202	0.11	6.2

Subgroup frequency

Statistical significance

Pattern generation

DivExplorer



DivExplorer - Divergent pattern exploration

Automatic subgroup identification

SUPPORT-BASED PRUNING

We consider only itemsets above a support threshold

Avoid statistical fluctuations of $\Delta(I)$

GENERAL APPROACH

Using the notion of outcome function

EFFICIENT ALGORITHM

Effective integration into algorithms for frequent pattern mining

Outcome function

$$o : X \to \mathbb{R} \cup \{\perp\}$$

$$o(x) = \begin{cases} 1 & \text{if } p(x) = \mathrm{T} \wedge t(x) = \mathrm{F} \\ 0 & \text{if } p(x) = \mathrm{F} \wedge t(x) = \mathrm{F} \\ \perp & \text{if } t(x) = \mathrm{T} \end{cases}$$

e.g. for FPR

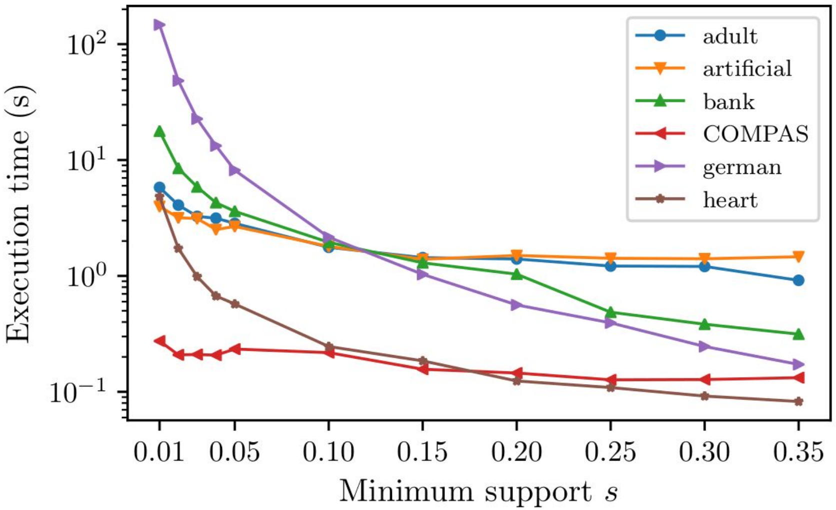
$$o(I) = E\{o(x) \mid x \models I, o(x) \neq \perp\}$$

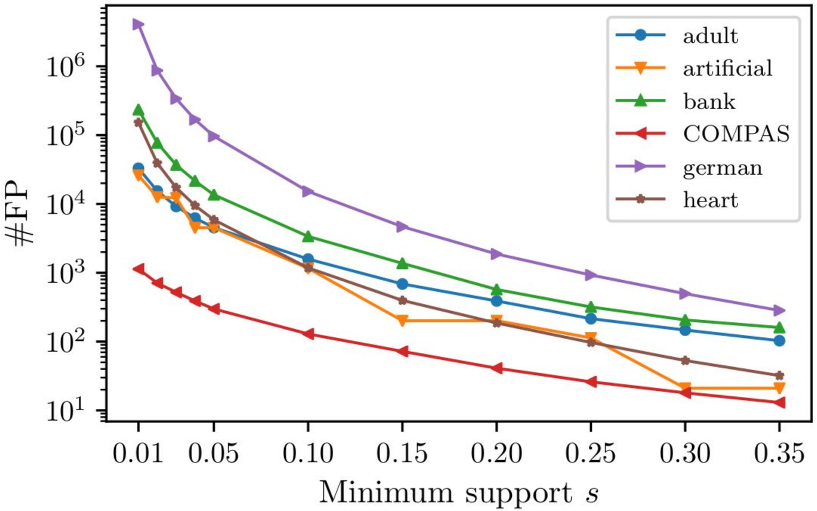
Divergence expressed as

$$\Delta_o(I) = o(I) - o(\emptyset)$$

Efficient integration into the process pattern extraction by tallying the sum and the count of $\{o(x) \mid x \models I, o(x) \neq \perp\}$

Efficiency of DivExplorer





Why **COMPLETE EXPLORATION** of patterns with adequate representation?

- **Complete characterization of the model behavior**

  Analysis of divergence of all adequately represented patterns

- Evaluation of local contribution to subgroup divergence

- Evaluation of global contribution to divergence

Why **COMPLETE EXPLORATION** of patterns with adequate representation?

- Complete characterization of the model behavior

  Analysis of divergence of all adequately represented patterns

- **Evaluation of local contribution to subgroup divergence**

- Evaluation of global contribution to divergence

Contributions of items to divergence

Itemset	Δ_{FPR}
age=25-45, #prior>3, race=Afr-Am, sex=Male	0.22

What is the contribution of each item?

Contributions of items to divergence

Shapley value

Given

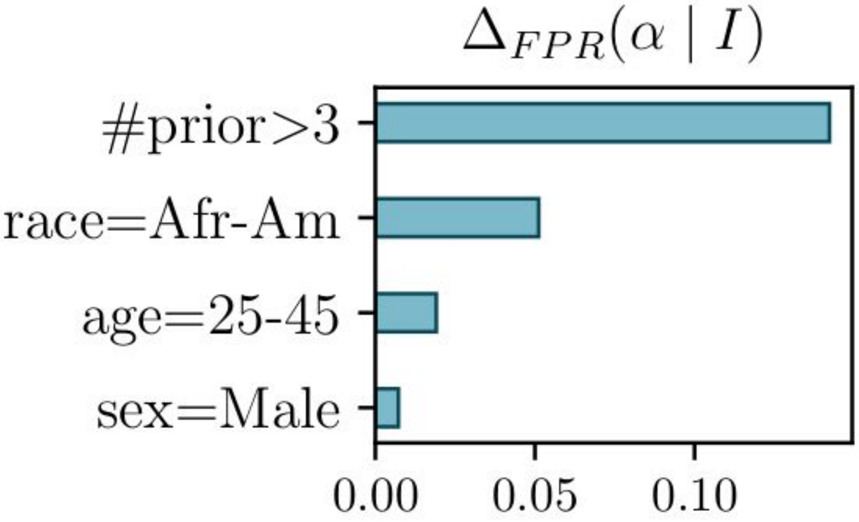
- Team of	$\rightarrow$	pattern I
- N players	$\rightarrow$	items in I
- Value $v(1,2,..N)$ of the team of N players	$\rightarrow$	divergence $\Delta(I)$
- Score of each subset $v(J)$ $\forall J \subseteq I$	$\rightarrow$	If I is frequent, all subsets $J \subseteq I$ are frequent $\rightarrow$ all $\Delta(J)$ are **already available**
estimate the **contribution** of each player to $v(1,2,..N)$	$\rightarrow$	contribution of $\alpha \in I$ to $\Delta(I)$

Contribution of item α in I:

$$\Delta(\alpha \mid I) = \sum_{J \subseteq I \setminus \{\alpha\}} \frac{|J|!(|I| - |J| - 1)!}{|I|!} [\Delta(J \cup \alpha) - \Delta(J)]$$

Contributions of items to divergence

Itemset	Δ_{FPR}
age=25-45, #prior>3, race=Afr-Am, sex=Male	0.22



Why **COMPLETE EXPLORATION** of patterns with adequate representation?

- Complete characterization of the model behavior

  Analysis of divergence of all adequately represented patterns

- Evaluation of local contribution to subgroup divergence

- **Evaluation of global contribution to divergence**

Global divergence

Global Shapley Value

A generalization of Shapley value that accounts for:

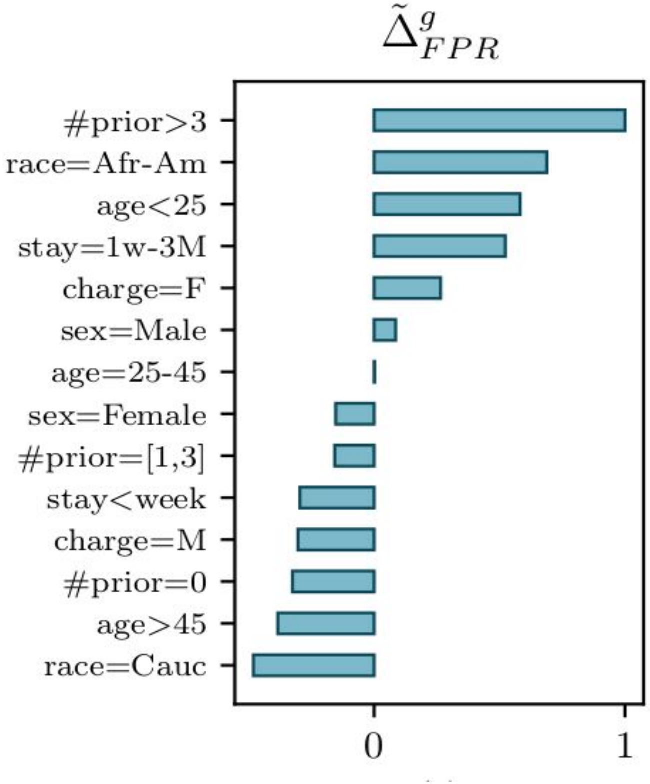
- Incompatible items (e.g. {*age<25*, *age>45*})
- Minimum support threshold

$$\widetilde{\Delta}^g(I, s) = \sum_{B \subseteq A \setminus \mathrm{attr}(I)} \frac{|B|!(|A| - |B| - |I|)!}{|A|! \prod_{b \in B \cup \mathrm{attr}(I)} m_b} \sum_{J : J \cup I \in \mathcal{I}^{\star}_{B \cup \mathrm{attr}(I)}} [\Delta(J \cup I) - \Delta(J)]$$

normalization factor, where m_b is # attribute values of b

set of frequent itemsets with attributes BUattr(I)

Global divergence - COMPAS



User study

- We inject controlled bias in a dataset (COMPAS)
- We produce diagnostics with DivExplorer, Slice Finder, LIME
- Can users figure out where the bias is? We count:
  - Full HITS: Users find bias
  - Partial HITS: Users find some items associated with bias, but not all

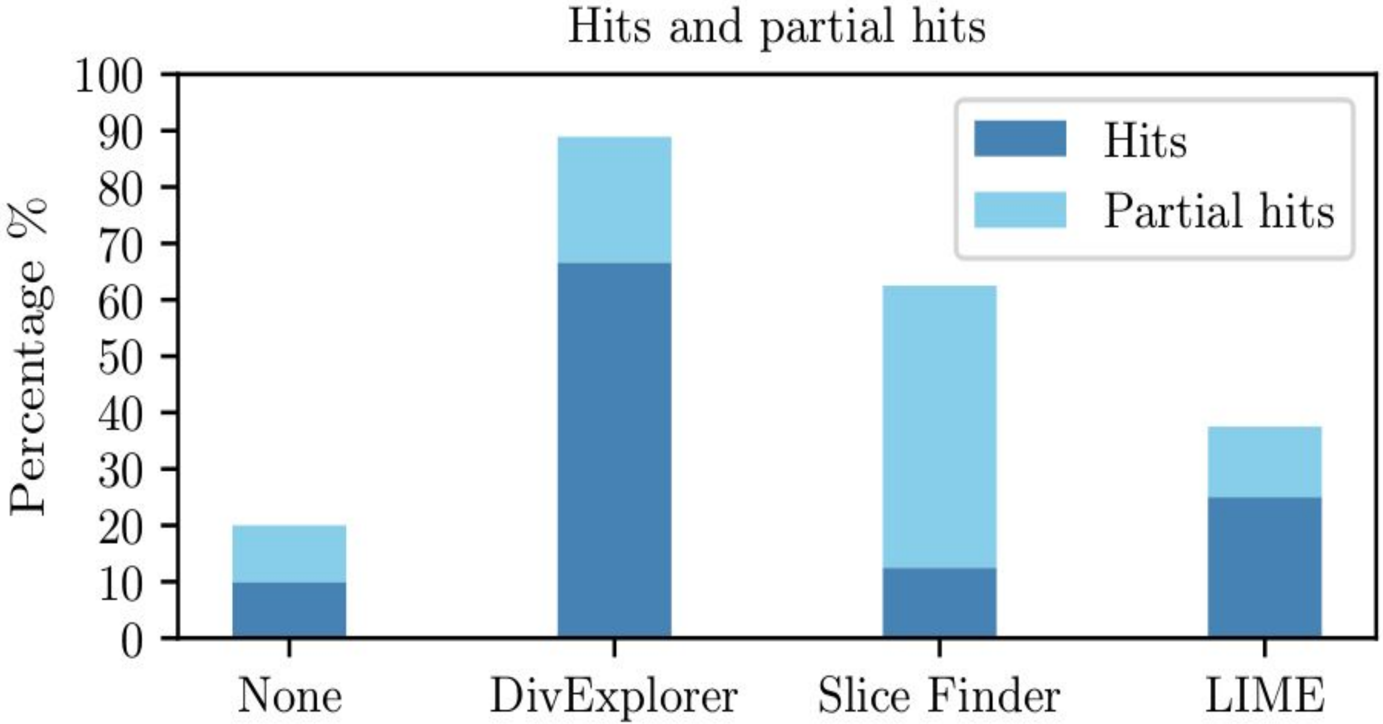
CONTROLLED EXPERIMENT

COMPARISON

USER TARGET

HIT RATE

User study


Hits and partial hits
Percentage %
100
90
80
70
60
50
40
30
20
10
0
Hits
Partial hits
None
DivExplorer
Slice Finder
LIME

www.divexplorer.org

Adjustments: Prune Redundancy by 0 Prune | Show Corrective Values | Reset

Search for specific records here | Clear | Search | Edit Columns

Support	Itemset	Δ_fpr	t_fp	Δ_fnr	t_fn	Δ_error	t_error	Δ_acc	FPR	FNR
0.13	(race=Afr-Am, #prior=>3, sex=Male, age=25-45)	0.22	7.1	-0.228	10.1	0.058	3.2	-0.058	0.308	0.47
0.1	(charge=F, race=Afr-Am, age=25-45, #prior=>3, sex=Male)	0.217	6.0	-0.248	9.8	0.046	2.2	-0.046	0.306	0.45
0.06	(stay=1w-3M, #prior=>3, sex=Male)	0.216	4.9	-0.174	5.7	0.099	3.8	-0.099	0.305	0.525
0.15	(race=Afr-Am, #prior=>3, age=25-45)	0.211	7.4	-0.226	10.4	0.055	3.1	-0.055	0.299	0.472
0.07	(stay=1w-3M, #prior=>3)	0.207	5.1	-0.183	6.3	0.089	3.7	-0.089	0.295	0.515

Globals Computation: Compute Global FPR Values | Compute Global FNR Values | Compute Global Error Values

<< < 1 > >>

Pastor, Gavgavian, Baralis, de Alfaro. How Divergent Is Your Data? *Demo Track.* VLDB 2021.

Generalization of divergence

Notion of divergence → to inspect the behavior of a generic model or instance property in subgroups

$$o : X \to \mathbb{R} \cup \{\perp\}$$

Attribute

Continuous

$$o(x) = a(x)$$

Discrete

$$o(x) = \begin{cases} 1 & a(x) = v \\ 0 & a(x) \neq v \end{cases}$$

Scoring

$$o(x) = w(x)$$

Ranking

$i(x)$ → rank position

$$o(x) = \gamma(i(x))$$

Top K

$$\gamma(i) = \begin{cases} 1 & i \leq k \\ 0 & i > k \end{cases}$$

Relation rank and benefit

$$\gamma(i) = i^{\alpha}$$

Ranking Divergence

Law School dataset →Ranking based on student normalized first-year average grade,

$$\gamma(i) = i^{-0.1}$$

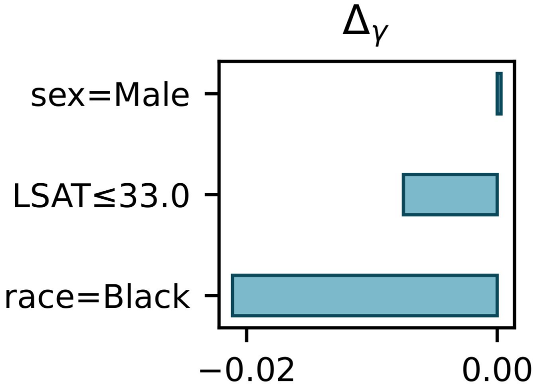
Higher
in the ranking

Itemset	Sup	Δ_γ	t
LSAT>41.0, UGPA>3.5, race=White, sex=Female	0.03	0.0206	8.7
LSAT>41.0, UGPA>3.5, race=White	0.07	0.0196	13.0
LSAT>41.0, UGPA>3.5, race=White, sex=Male	0.04	0.0189	9.9
LSAT≤ 33.0, race=Black, sex=Male	0.02	-0.0283	25.6
LSAT≤ 33.0, UGPA≤ 3.0, race=Black, sex=Male	0.01	-0.0280	21.0
LSAT≤ 33.0, UGPA≤ 3.0, race=Black	0.03	-0.0278	31.4

Lower
in the ranking

Ranking Divergence

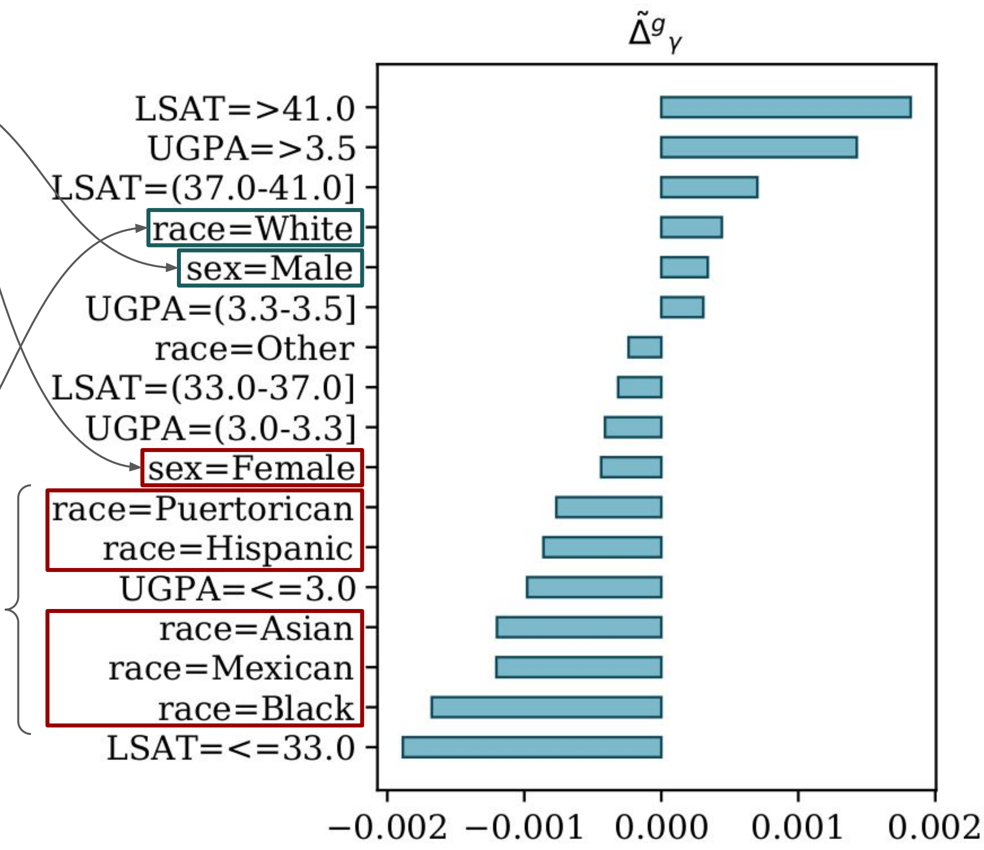
Itemset	Sup	Δ_γ	t
LSAT$\leq$ 33.0, race=Black, sex=Male	0.02	-0.0283	25.6
LSAT$\leq$ 33.0, UGPA$\leq$ 3.0, race=Black, sex=Male	0.01	-0.0280	21.0
LSAT$\leq$ 33.0, UGPA$\leq$ 3.0, race=Black	0.03	-0.0278	31.4



Ranking Divergence

Role of gender

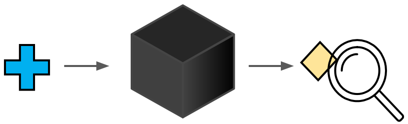
Role of race



Conclusions

Post-hoc explanation approaches to **enhance** the **interpretability** of classification models

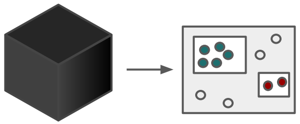
Pattern → Intrinsic interpretability and ability to capture associations and group data



From the **individual** perspective

Local explanations to explain individual predictions

Qualitative and **quantitative** understanding

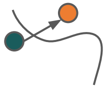


From the **subgroup** perspective

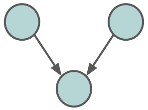
Identifying and characterizing peculiar model behavior in subgroups

- **Automatic identification** of divergent subgroups
- **Exploration** of lattice of **patterns** and their divergence
- **Contribution** of items to divergence

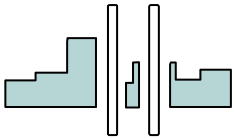
Future work



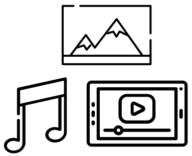
Conterfactual explanations



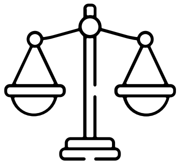
Causal reasoning



Discretization



Unstructured data



Fairness

List of publications

- Pastor, de Alfaro, Baralis. Looking for Trouble: Analyzing Classifier Behavior via Pattern Divergence. SIGMOD 2021.
- Pastor, de Alfaro, Baralis. Identifying Biased Subgroups in Ranking and Classification. Workshop on Responsible AI@ACM KDD 2021.
- Pastor, Gavgavian, Baralis, de Alfaro. How Divergent Is Your Data? *Demo Track.* VLDB 2021.
- Pastor and Baralis. Bring Your Own Data to X-PLAIN. *Demo Track.* ACM SIGMOD 2020.
- Pastor and Baralis. Explaining black box models by means of local rules. ACM SAC 2019.
- Pastor and Baralis. Enhancing Interpretability of Black Box Models by means of Local Rules. ACM womENcourage 2019.
- Pastor. Deriving Local Internal Logic for Black Box Models. SEBD 2018.

- Giordano, Giobergia, Pastor, La Macchia, Cerquitelli, Baralis, Mellia, Tricarico. Data-Driven Strategies for Predictive Maintenance: Lesson Learned from an Automotive Use Case. Computers in Industry 2021 *(to appear)*.
- Giordano, Pastor, Giobergia, Cerquitelli, Baralis, Mellia, Neri, Tricarico. Dissecting a data-driven prognostic pipeline: A powertrain use case. Expert Systems with Applications 2021.
- Apiletti and Pastor. Correlating espresso quality with coffee-machine parameters by means of association rule mining. Electronics 2020.
- Apiletti, Pastor, Callà, Baralis. Evaluating espresso coffee quality by means of time-series feature engineering. EDBT/ICDT Workshops 2020.

- Baralis, Garza, Pastor. A density-based preprocessing technique to scale out clustering. IEEE Big Data 2018.

- Attanasio and Pastor. PoliTeam@ AMI: Improving Sentence Embedding Similarity with Misogyny Lexicons for Automatic Misogyny Identification in Italian Tweets. EVALITA 2020.

Thank you for your attention