Label-Descriptive Patterns and Their Application to Characterizing Classification Errors

Michael Hedderich*
Jonas Fischer*
Dietrich Klakow
Jilles Vreeken

SIC Saarland Informatics Campus

max planck institut informatik

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

*equal contribution

To err is human (and models do it, too)

Blackbox Classifier

Task, e.g.
Visual Question Answering

(image by Andrea Lightfoot https://unsplash.com/photos/GX6be6LLIR4)

To err is human (and models do it, too)

Task, e.g.
Visual Question Answering

Blackbox Classifier

Instance	Correct Prediction?
Are there many ducks playing?	✓
How many ducks are in the picture?	✗
What are the ducks eating?	✗
Do you see ducks in the puddle?	✓
How many roosters are in the puddle?	✗

(image by Andrea Lightfoot https://unsplash.com/photos/GX6be6LLIR4)

To err is human (and models do it, too)

Task, e.g.
Visual Question Answering

Blackbox Classifier

Instance	Correct Prediction?
Are there many ducks playing?	✓
How many ducks are in the picture?	✗
What are the ducks eating?	✗
Do you see ducks in the puddle?	✓
How many roosters are in the puddle?	✗

Why?

Characterize classification *errors* with *patterns*

- global over all instances
- non-redundant
- easy to interpret
- actionable

(image by Andrea Lightfoot https://unsplash.com/photos/GX6be6LLIR4)

Minimum Description Length Principle

Length in bits of sending data and model

$$\operatorname{argmin}_{M \in \mathcal{M}} \mathrm{L}(\mathrm{D}, \mathrm{M})$$

Model M over model class $\mathcal{M}$

Minimum Description Length Principle

Method

Data is given by instances of correct and wrong classifications with corresponding label

Model is composed of label specific patterns

$$\text{argmin}_{M \in \mathcal{M}} \, \mathrm{L}(\mathrm{D}, \mathrm{M})$$

D

D^-	How many ducks are in the picture?	✗
	How many roosters can you see?	✗
	What colour is the water?	✗
	What color are the ducks?	✗
D^+	Do you see a rooster in the puddle?	✓
	Are there many ducks playing?	✓

M

How and many
What and (color xor colour)

Minimum Description Length Principle

Data is given by instances of correct and wrong classifications with corresponding label

Model is composed of label specific patterns

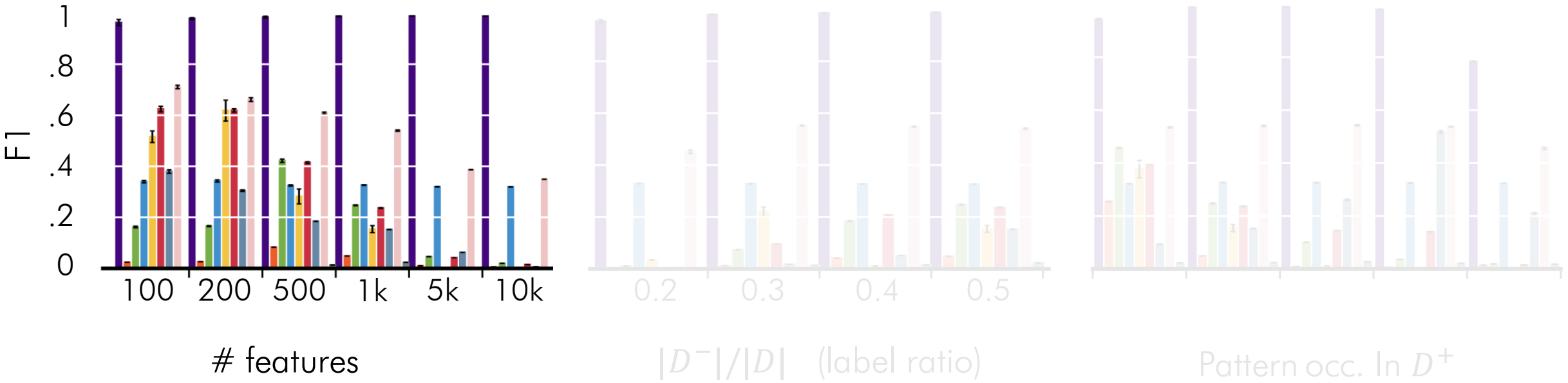
$$\operatorname{argmin}_{M \in \mathcal{M}} \mathrm{L}(\mathrm{D}, \mathrm{M})$$

For efficient search in practice: PREMISE

- iteratively explores the pattern space in bottom-up fashion
- uses word embeddings to explore mutually exclusive patterns
- employs statistical testing for filtering spurious patterns

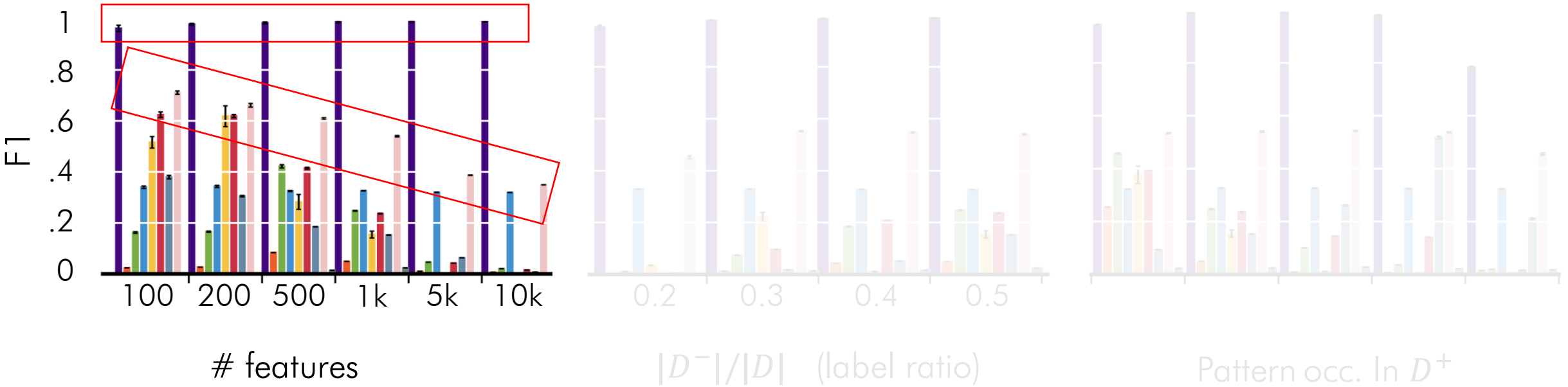
Synthetic Data With Known Ground Truth

Results



Synthetic Data With Known Ground Truth

Results

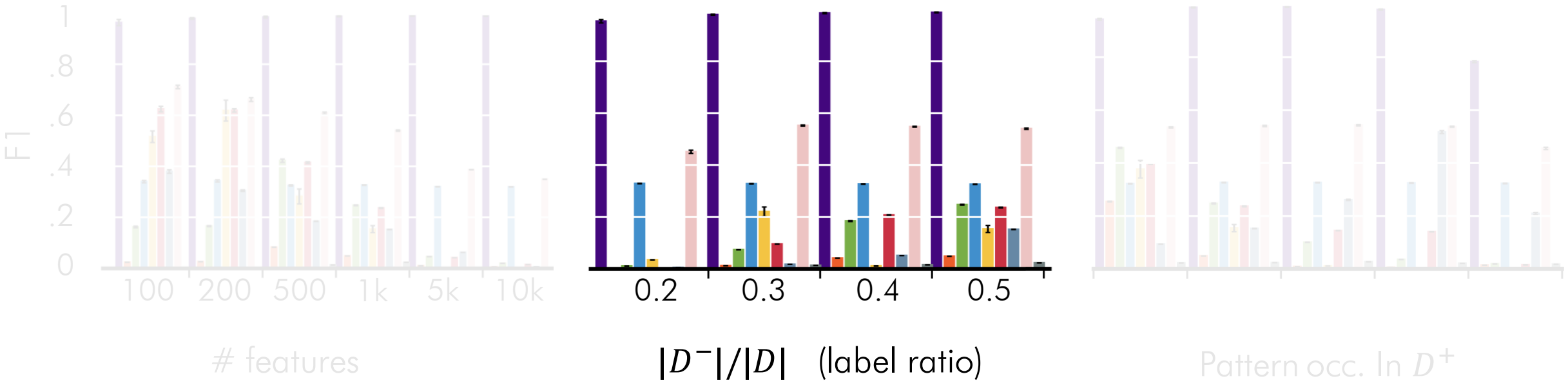


Current methods discover more spurious patterns with increase in features.

PREMISE TREE RIPPER SUBGROUP GRAB CLASSY SPUMANTE CORTANA C-SALT

Synthetic Data With Known Ground Truth

Results

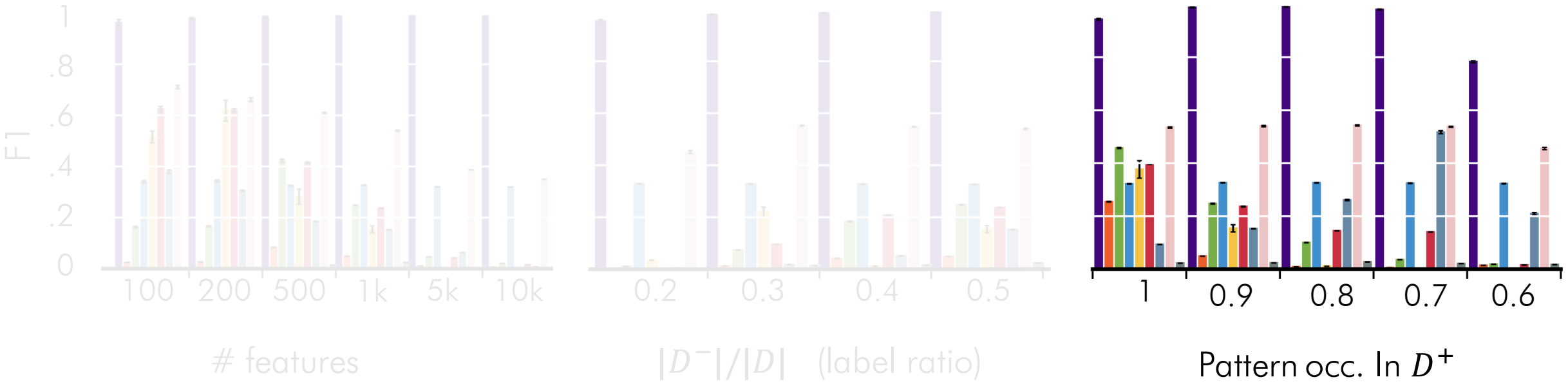


More extreme label ratios are an issue for many methods.

PREMISE TREE RIPPER SUBGROUP GRAB CLASSY SPUMANTE CORTANA C-SALT

Synthetic Data With Known Ground Truth

Results



Many more experiments in the paper!

PREMISE TREE RIPPER SUBGROUP GRAB CLASSY SPUMANTE CORTANA C-SALT

Visual Question Answering (VQA)

Results

LXMERT (Tan & Bansal, 2019) 70% VQA accuracy

(image by Andrea Lightfoot https://unsplash.com/photos/GX6be6LLIR4)

Visual Question Answering (VQA)

Results

LXMERT (Tan & Bansal, 2019) 70% VQA accuracy

How and many

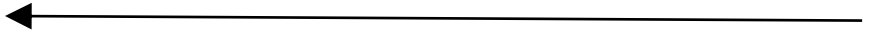
hanging and from

(kind xor sort) and of

(would xor could xor might xor can) and you

number

letter xor letters



Basic math (counting,…)

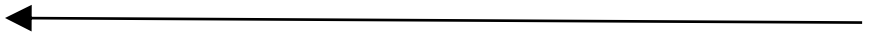
More experiments on VQA and how to improve an NER classifier in our paper!

Visual Question Answering (VQA)

Results

LXMERT (Tan & Bansal, 2019) 70% VQA accuracy

How and many
hanging and from
(kind xor sort) and of
(would xor could xor might xor can) and you
number
letter xor letters



Spatial reasoning

More experiments on VQA and how to improve an NER classifier in our paper!

Visual Question Answering (VQA)

Results

LXMERT (Tan & Bansal, 2019) 70% VQA accuracy

How and many

hanging and from

(kind xor sort) and of

(would xor could xor might xor can) and you

number

letter xor letters

Ontology relationships

More experiments on VQA and how to improve an NER classifier in our paper!

Visual Question Answering (VQA)

Results

LXMERT (Tan & Bansal, 2019) 70% VQA accuracy

How and many

hanging and from

(kind xor sort) and of

(would xor could xor might xor can) and you ← Soft questions

number

letter xor letters

More experiments on VQA and how to improve an NER classifier in our paper!

Visual Question Answering (VQA)

Results

LXMERT (Tan & Bansal, 2019) 70% VQA accuracy

How and many

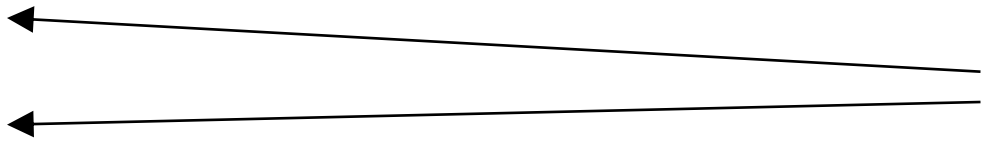
hanging and from

(kind xor sort) and of

(would xor could xor might xor can) and you

number

letter xor letters



Recognition subtask

More experiments on VQA and how to improve an NER classifier in our paper!

Discussion and Conclusion

Summary

Goal:
Discover *easy-to-interpret* patterns that characterize misclassifications.

Method:
Premise, based on Minimum Description Length Principle.

Results:
Consistently discovers non-redundant and descriptive patterns, where state-of-the-art fails. Patterns are easy to understand and to act upon.

Outlook:
Scale to more features to investigate patterns of neuron activations characterizing misclassifications.

https://github.com/uds-lsv/premise
PyPremise

Discussion and Conclusion

Summary

Goal:

Discover *easy-to-interpret* patterns that characterize misclassifications.

Method:

Premise, based on Minimum Description Length Principle.

Results:

Consistently discovers non-redundant and descriptive patterns, where state-of-the-art fails. Patterns are easy to understand and to act upon.

Outlook:

Scale to more features to investigate patterns of neuron activations characterizing misclassifications.

Analyze your own classifier! →

https://github.com/uds-lsv/premise

PyPremise