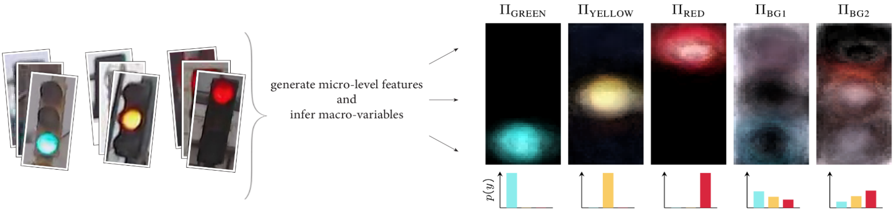
CaFe DBSCAN

A Density-based Clustering Algorithm for Causal Feature Learning



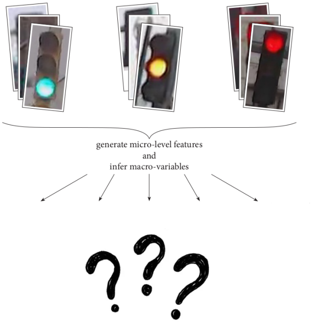
Pascal Weber, Lukas Miklautz, Akshey Kumar,
Moritz Grosse-Wentrup, Claudia Plant

Overview

- New density-based Clustering Algorithm for Causal Feature Learning
  - Standalone algorithm
  - Faster
  - Robust to noise
  - Automatic learns a reasonable number of clusters
- Two related research areas: Clustering and Causality
- Synthetic and real world experiments

Background – Causal Feature Learning

- Infer macro-level variables from micro-level data and their effects
- E.g., which aggregation of pixels are responsible for recognizing the state of a traffic light



Group x and x' together iff.

$$P(Y \mid do(X = x)) = P(Y \mid do(X = x'))$$

where do is the operator for applying interventions.

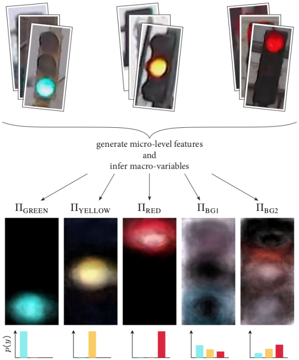
Background – Causal Feature Learning

- Often it is not feasable to intervene on data
  ethical considerations — e.g. medicine
  physical constraints — e.g. weather
  ...
- Instead find *observable* macro-variables

Group x and x' together iff.

$$P(Y \mid X = x) = P(Y \mid X = x')$$

where x, x' is the observed data.



Causal Coarsening Theorem

Theorem

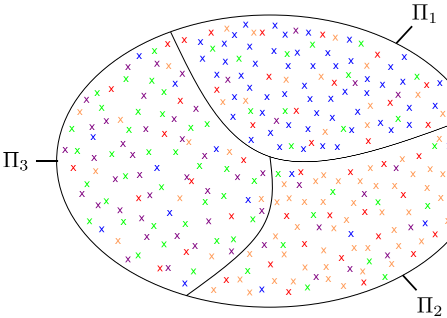
Among all the generative distributions which induce a given observational partition $\Pi^{(o)}$, all except for a subset of Lebesgue measure zero induce a causal partition $\Pi^{(c)}$ that is a coarsening of $\Pi^{(o)}$.

by K. Chaloupka, et. al. in 2015

$\Rightarrow$ We can still merge observational partitions to single causal partitions without loosing causal information with human knowledge or further experiments / observations

$\Rightarrow$ Interventions / Experiments on every observational class instead of on every micro level data combination

Example



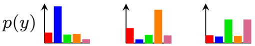
$\mathcal{Y} = \{\ 1,\ \ 2,\ \ 3,\ \ 4,\ \ 5\ \}$

effect states

$\Downarrow$

clusters

$\Pi = \{\ \Pi_1,\ \ \Pi_2,\ \ \Pi_3\ \}$



effect state distributions

State-of-the-art algorithm

2-step algorithm [1]:

1. Learn the conditional probabilities with a classifier first, e.g. MLP
2. Cluster the learned probabilities, e.g. k-Means

Drawbacks:

- Classifier and clustering algorithm rely on different assumptions
- Computationally expensive
- Can’t handle noise and outliers very well

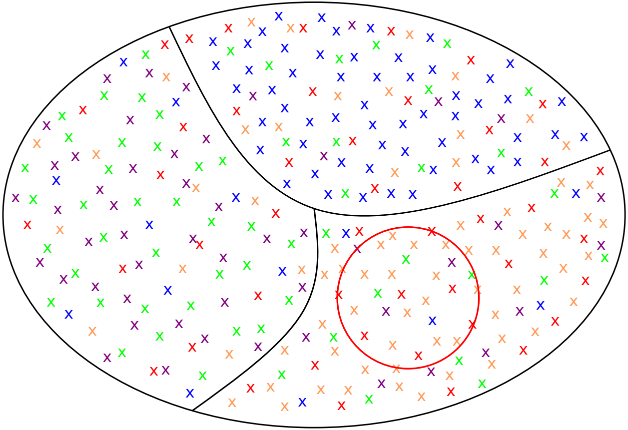
[1] "Visual causal feature learning"by K. Chalupka, P. Perona, and F. Eberhardt (2015)

CaFe DBSCAN

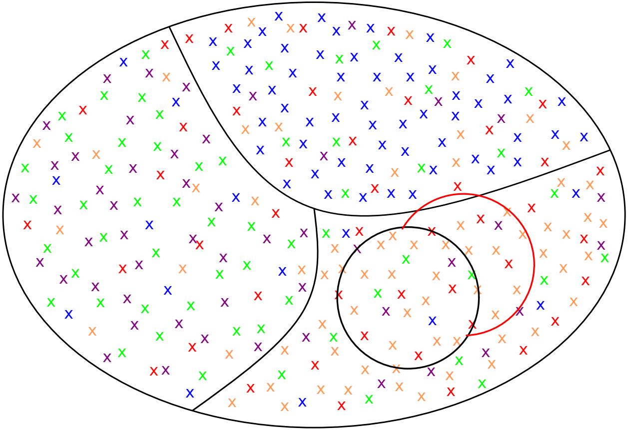
- Density-based clustering algorithm
- Idea: Extend an initial probability region until the probability distribution of neighboring regions changes
- Parameters: ϵ (radius), μ (min_points), and τ (probability distance)
- Returns cluster with their estimated probability distributions

Implementation:
`https://gitlab.cs.univie.ac.at/pascalw777/cafe-dbscan`

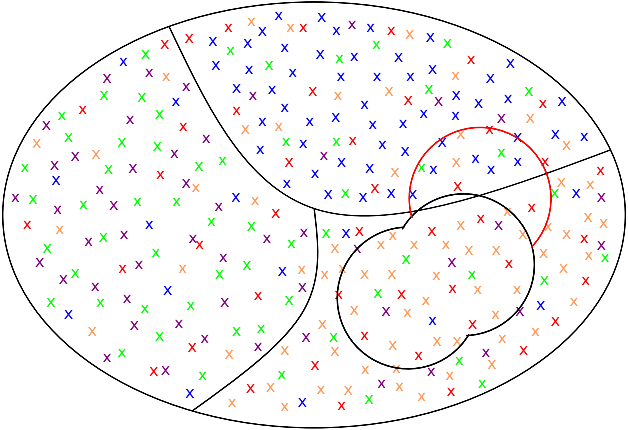
CaFe DBSCAN



CaFe DBSCAN

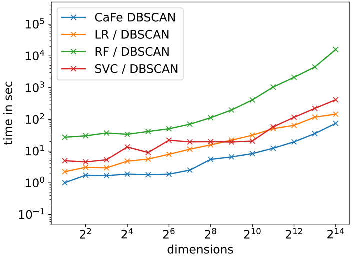


CaFe DBSCAN



Experiments

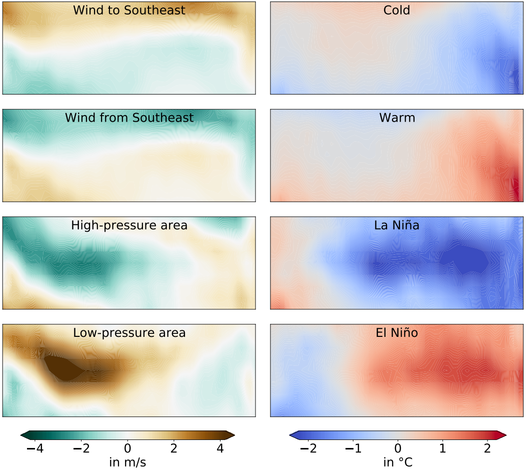
- Runtime experiments
- Hyperparamter sensitivity experiments
- Experiments on synthetic data
- Real world experiments



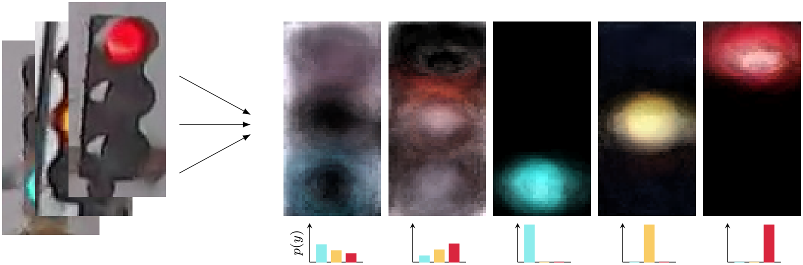
Experiments on synthetic data

Method	DS1: Separated Circles ($n = 3,496$, $d = 2$)				DS2: Adjacent Circles ($n = 4,416$, $d = 2$)				DS3: Rectangles / Circles ($n = 9,343$, $d = 2$)			
	Clf. acc.	k	NMI	ARI	Clf. acc.	k	NMI	ARI	Clf. acc.	k	NMI	ARI
Ground Truth	1.00	6(+1)	1.00	1.00	1.00	7(+1)	1.00	1.00	1.00	7(+1)	1.00	1.00
CaFe DBSCAN	-	6(+1)	$\underline{0.97}$	0.97	-	7(+1)	$\underline{0.94}$	$\underline{0.94}$	-	7(+1)	$\underline{0.86}$	$\underline{0.85}$
CaFe DBSCAN$^+$	-	6(+1)	**0.98**	**0.99**	-	7(+1)	**0.95**	**0.95**	-	7(+1)	**0.91**	**0.91**
LR / k-Means	0.33	6*	0.66	0.45	0.50	7*	0.79	0.74	0.30	7*	0.40	0.18
RF / k-Means	0.33	6*	0.66	0.46	0.50	7*	0.68	0.54	0.32	7*	0.78	0.70
SVC / k-Means	0.33	6*	0.31	0.13	0.50	7*	0.59	0.40	0.27	7*	0.38	0.15
MLP / k-Means	0.33	6*	0.65	0.41	0.52	7*	0.76	0.65	0.34	7*	0.79	0.67
LR / DBSCAN	0.33	6(+1)	**0.98**	$\underline{0.98}$	0.50	5(+1)	0.86	0.70	0.30	10(+1)	0.60	0.37
RF / DBSCAN	0.33	6(+1)	0.74	0.60	0.50	45(+1)	0.69	0.50	0.32	10(+1)	0.77	0.70
SVC / DBSCAN	0.33	5(+1)	0.51	0.28	0.50	4(+1)	0.64	0.42	0.27	5(+1)	0.59	0.33
MLP / DBSCAN	0.33	9(+1)	0.86	0.77	0.52	13(+1)	0.87	0.83	0.34	20(+1)	0.77	0.66
LR / HDBSCAN	0.33	6(+1)	0.90	0.83	0.50	5(+1)	0.87	0.71	0.30	27(+1)	0.52	0.25
RF / HDBSCAN	0.33	6(+1)	0.72	0.59	0.50	17(+1)	0.65	0.36	0.32	21(+1)	0.60	0.32
SVC / HDBSCAN	0.33	20(+1)	0.69	0.46	0.50	9(+1)	0.84	0.69	0.27	9(+1)	0.51	0.23
MLP / HDBSCAN	0.33	5(+1)	0.88	0.81	0.52	12(+1)	0.87	0.87	0.34	13(+1)	0.78	0.70

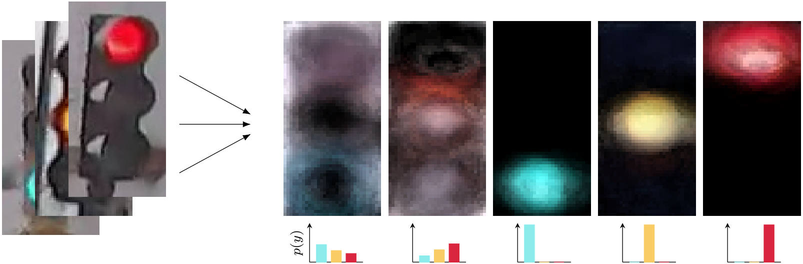
Real world example — El Niño



Real world example — Traffic Lights



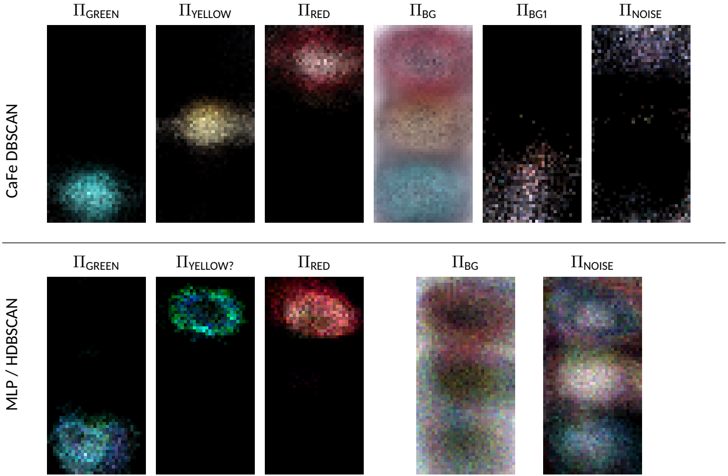
Real world example — Traffic Lights



observational class: $P(Y \mid X = x) = P(Y \mid X = x').$ (2)

causal class: $P(Y \mid do(X = x)) = P(Y \mid do(X = x')).$ (1)

Real world example — Traffic Lights (with noise)



Summary

- Introduced the idea of density-based clustering of conditional probabilities
- We showed in various experiments that our algorithm outperforms the previous state-of-the-art approach for CFL in
    - clustering quality
    - speed
    - robustness to noise

Further work:

- Adapt algorithm to other variants of density based clustering algorithms, e.g. HDBSCAN
- Usage in interpretable machine learning

Thank you for your attention!

Questions?