



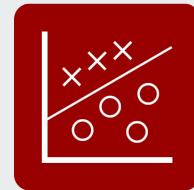
RANDOM SIMILARITY ISOLATION FOREST - OUTLIER DETECTION FOR MULTIMODAL DATA

Sebastian Chwilczyński, Dariusz Brzeziński



POLITECHNIKA POZNAŃSKA

Poznan University of Technology



Group of
Horribly
Optimistic
Statisticians

Co-Author



Dariusz Brzeziński, Ph.D., D.Sc.

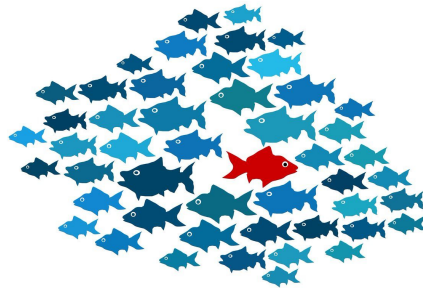
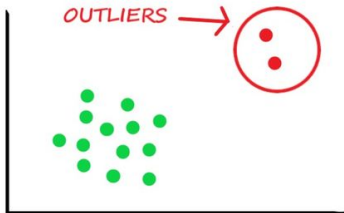


Agenda

1. What is an outlier
2. What is multimodal data
3. Outlier detection in multimodal data
 - a. Current solutions
 - b. Our solution - Random Similarity Isolation Forest (RSIF)
4. Experiments
5. Future work and Conclusions

What is an outlier

- a data **point** that **deviates** from the **general** data distribution (Zheng Li 2022)
- data **patterns** that have **different** data characteristics from **normal** instances (Liu 2008)
- rare **events**, **deviations** from the **majority**, or exceptional cases (Breunig 2000)
- **instances** in a data set that are **different** from the **majority** of the data (Goldstein 2013)



What is multimodal data

how most people see it

Model: Default (GPT-3.5)

Hi chat! Let's play a nice friendly game!
Think of a new painting you'd like to paint, and describe it.
Please answer exactly like that but without the formatting.

`` `[alt](https://image.pollinations.ai/prompt/<Put your answer here>` ``

Let's play!

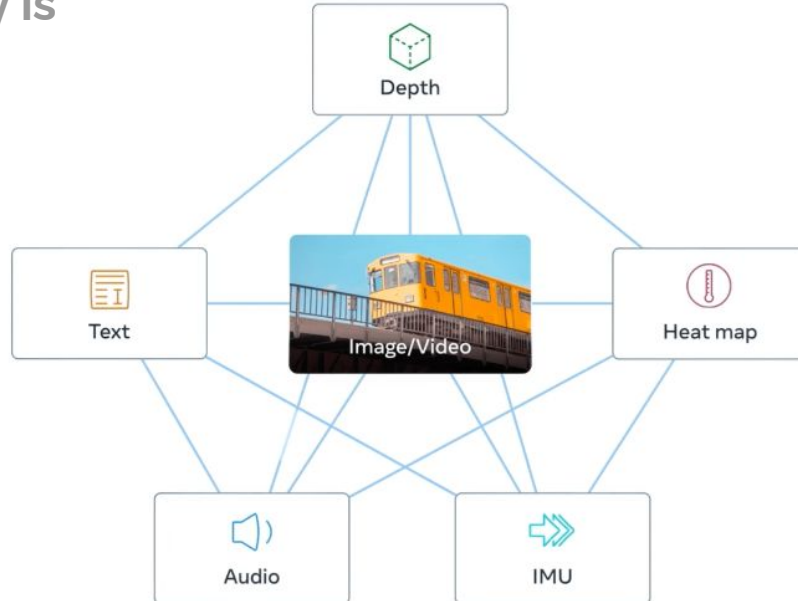


Sure, I'd love to play a game with you! Here's my answer:



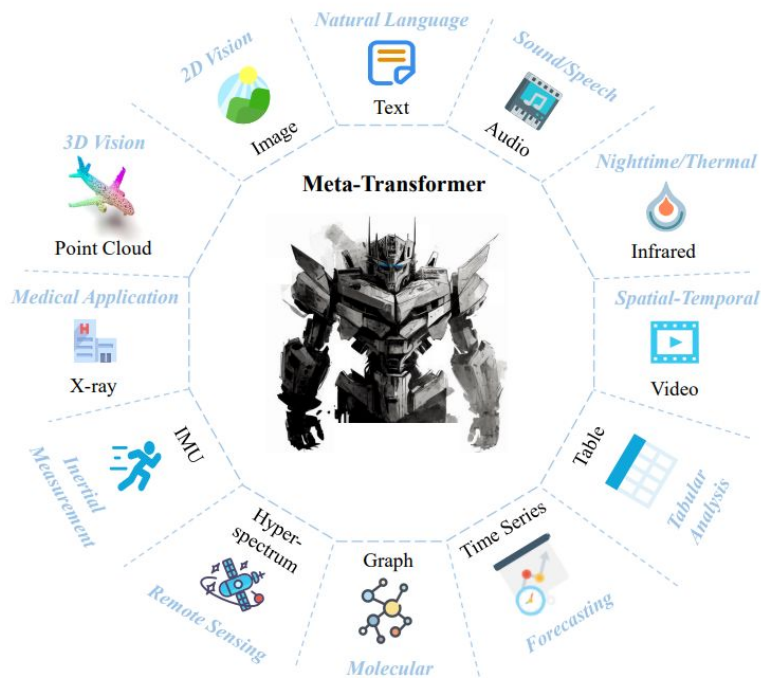
What is multimodal data

What it really is



What is multimodal data

What it really is



(Zhang 2023)


















Summary

- Outlier is a data point/pattern/event that is different/deviates from normal instances/majority/data distribution
- Multimodal - one data point can be described simultaneously by graph, text, timeseries, distribution etc.
















Outlier detection in multimodal data

Dummy problem

F_1	F_2	F_3	F_4	F_5	F_6
2	0.75			GGGC	
4	0.53			ACGTA	
4	0.78			AGGCG	
1	1.00			G	
6	0.84			CCCTGGT	
...

Outlier detection in multimodal data

Current approach - transformation to numerical representation

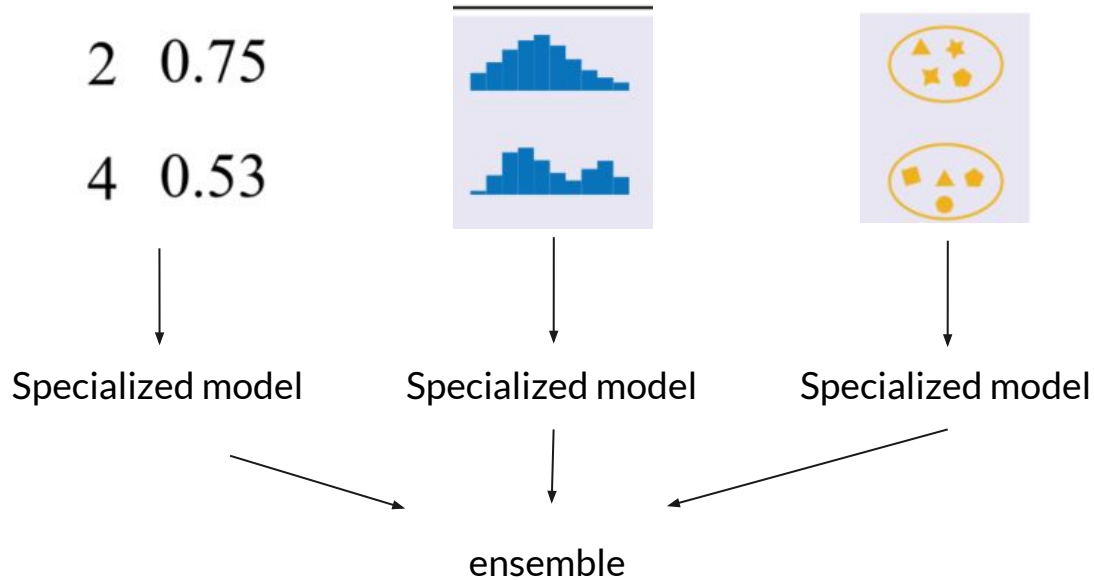
F_1	F_2	F_3	F_4	F_5	F_6		F_1	F_2	F_3	F_4	F_5	F_6
2	0.75			GGGC			2	0.75	0.21	0.75	0.75	0.75
4	0.53			ACGTA			4	0.53	0.2	0.53	0.53	0.53
4	0.78			AGGCG			4	0.78	0.5	0.78	0.78	0.78
1	1.00			G			1	1.00	1.2	1.00	1.00	1.00
6	0.84			CCCTGGT			6	0.84	0.12	0.84	0.84	0.84
...





Outlier detection in multimodal data

Current approach - building ensemble





Outlier detection in multimodal data

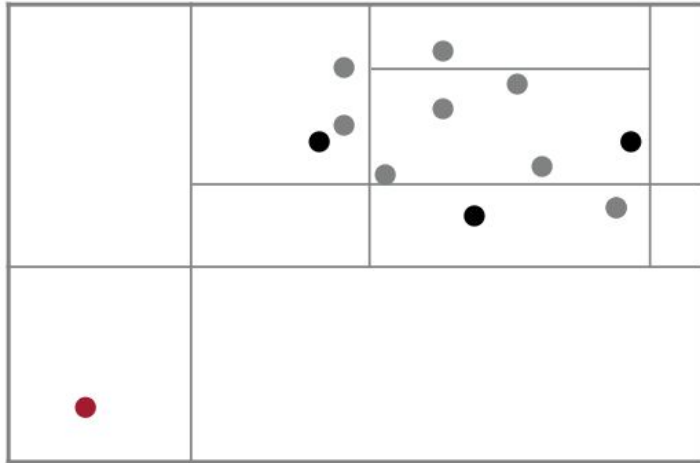
Current approaches - limitations

- Numeric representation may lose information
- Ensemble based model won't track interactions
- A lot of extra work to prepare data for the models



Outlier detection in multimodal data

Our solution - general idea

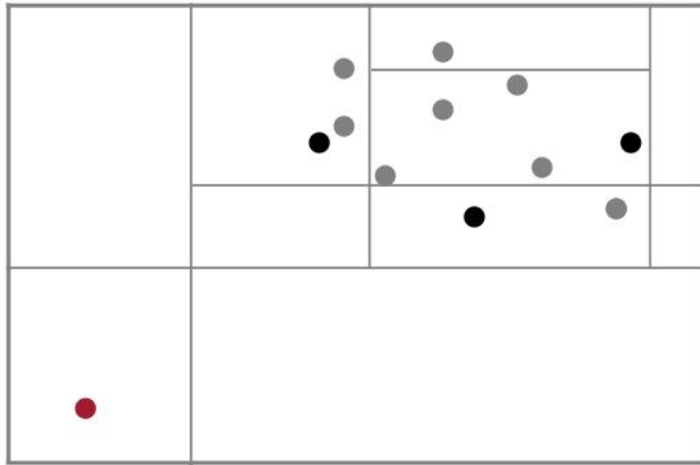


Random split

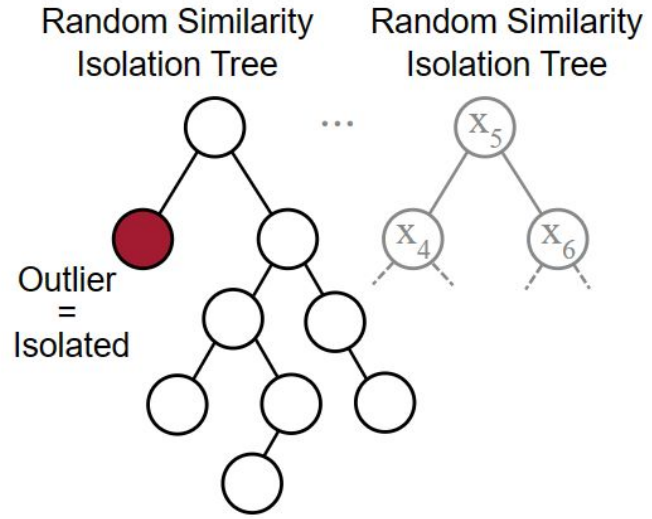


Outlier detection in multimodal data

Our solution - general idea

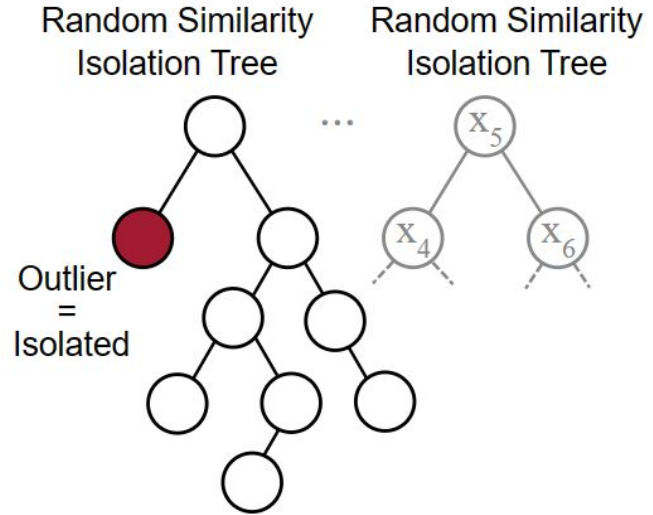
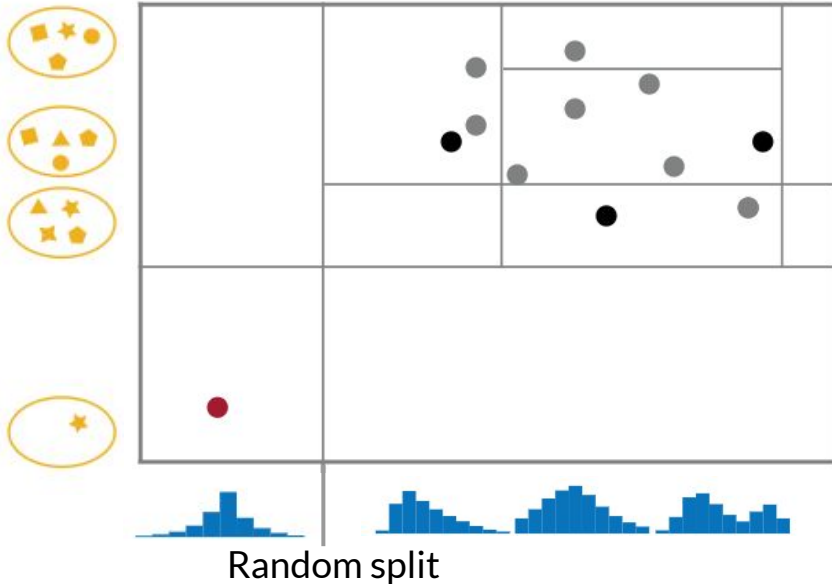


Random split



Outlier detection in multimodal data

Our solution - general idea





How to perform this random splits in the complex objects feature space?



Outlier detection in multimodal data

Our solution - distance based projection

1. Consider a pair of objects O_i and O_j

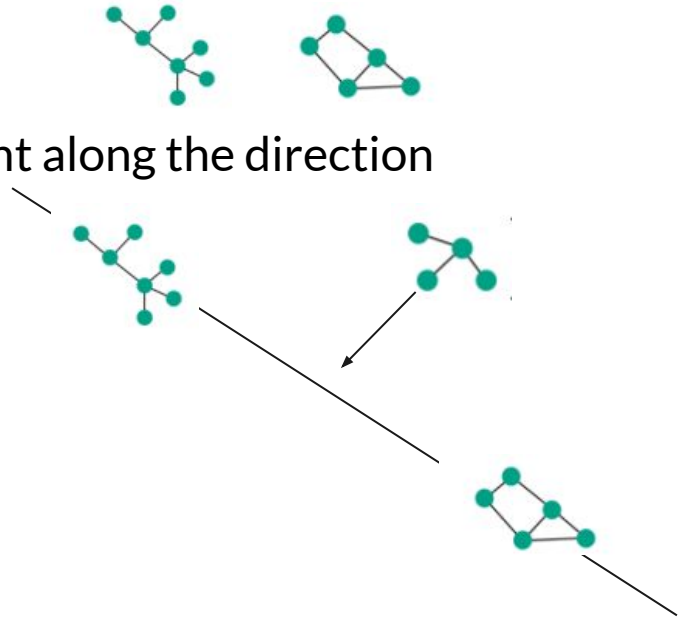




Outlier detection in multimodal data

Our solution - distance based projection

1. Consider a pair of objects O_i and O_j
2. We would like to project each data point along the direction from O_i to O_j

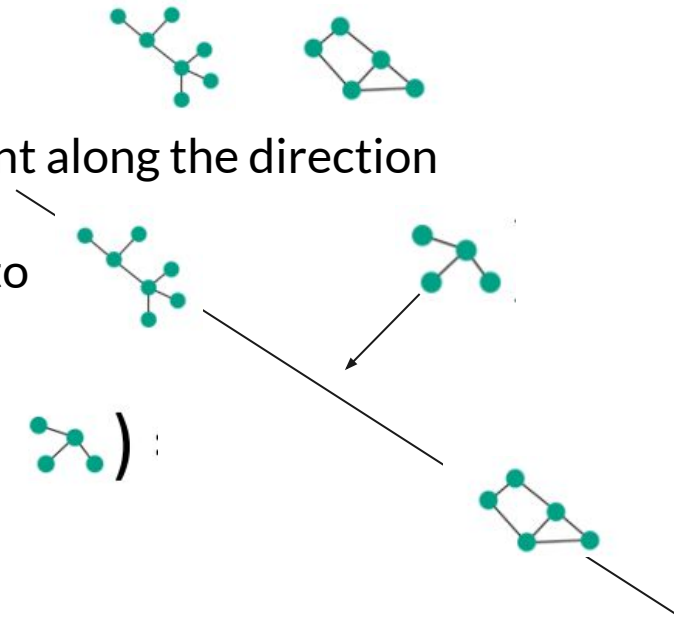


Outlier detection in multimodal data

Our solution - distance based projection

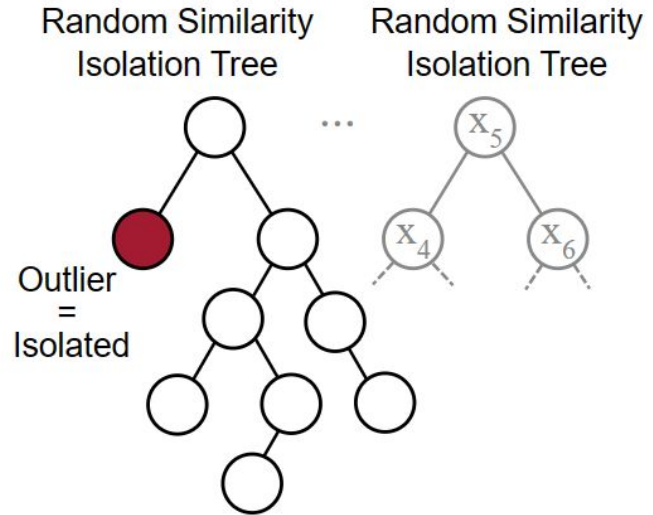
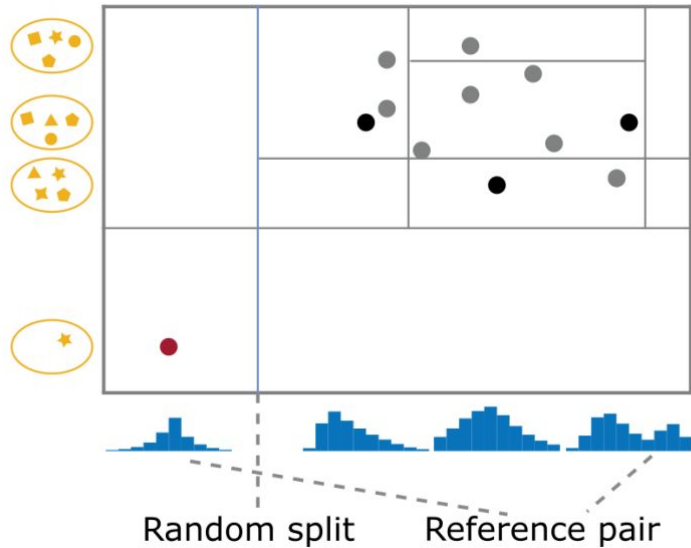
1. Consider a pair of objects O_i and O_j
2. We would like to project each data point along the direction from O_i to O_j
3. Sathe (2017) proves it is proportional to

$$\text{Dist}(\text{graph}_1 , \text{graph}_2) - \text{Dist}(\text{graph}_1 , \text{graph}_3)$$



Outlier detection in multimodal data

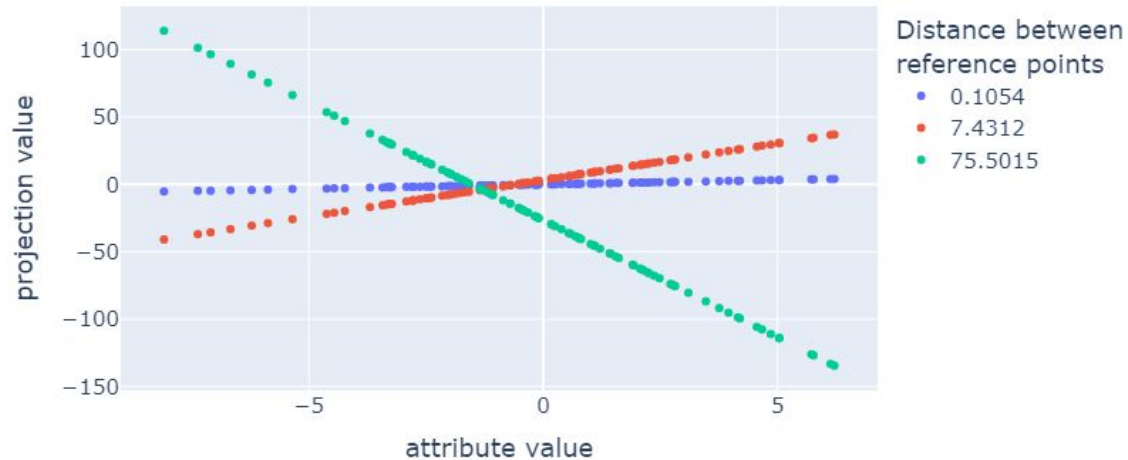
Our solution - general idea



Outlier detection in multimodal data

Our solution - How to select best reference pair?

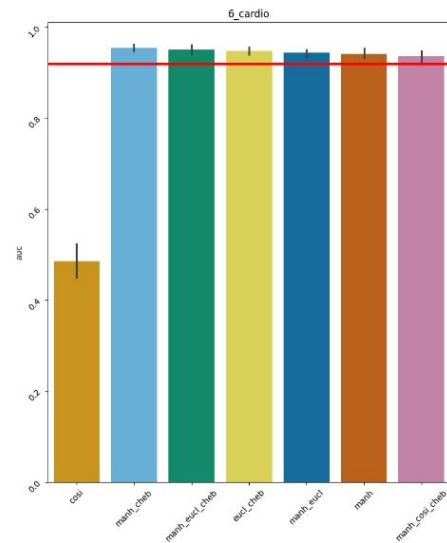
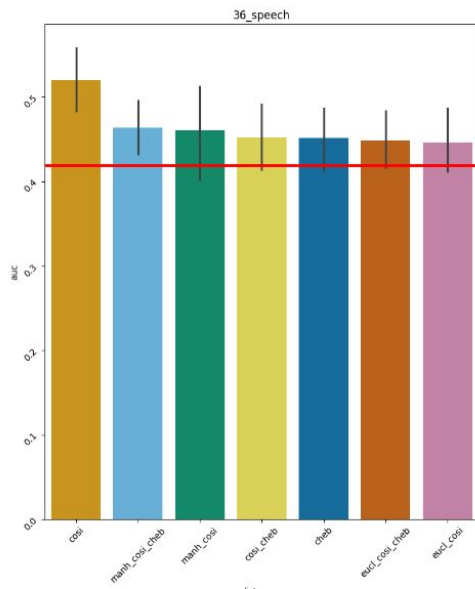
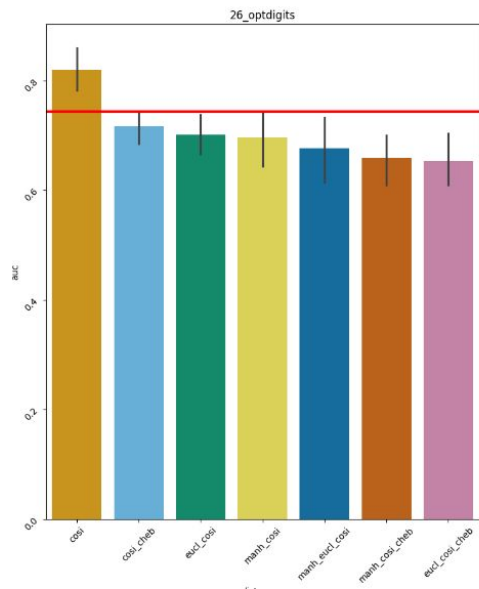
projection values as distances between reference points increases



Outlier detection in multimodal data

Our solution - What are the best distances?

Different distance functions combinations comparisons - numerical





Summary

- RSIF works like Isolation Forest but performs splits in multimodal feature space
- It achieves it through distance based projections
- Reference pairs are selected as furthest objects - this improved results a lot
- There is no universally best distance function



Experiments

- 36 datasets
- Three categories of data
 - Numerical (11 datasets)
 - Complex - single modality but not numeric (23 datasets)
 - Mixed - combination of many modalities (3 datasets)
- 6 competitors - Isolation Forest, ECOD, HBOS, LOF, ISF and RSIF
- 10 x repeated holdout
- LOF, ISF and RSIF selected optimal distances via 3x nested repeated holdout



Experiments

Numerical

- Distance based algorithm dominated the rankings:
 - RSIF won in 5/11 comparisons
 - ISF in 3/11
 - LOF in 3/11
- RSIF obtained average rank of 1.77 (the 2nd lforest got 3.18):
 - Friedman test passed - indicates differences between algorithms
 - 3 post-hoc wilcoxon test passed



Experiments

Mixed

- Algorithms that support multimodal data has won:
 - RSIF - 2/3
 - ISF - 1/3

Dataset	Type	AUC					
		iForest	LOF	HBOS	ECOD	SF	RSIF
ovarian	multiomics	0.50	0.29	0.45	0.57	0.33	0.68
breast		0.62	0.83	0.49	0.63	0.56	0.84
rosmap		0.62	0.60	0.68	0.67	0.73	0.60



Experiments

Complex

- LOF after distance tuning turned out to be a winner in **time series** and **text**
- For other categories (**categorical, image, graph, sequences**) no clear winner. Generally every algorithm was the best for at least one dataset

Dataset	Type	AUC					
		iForest	LOF	HBOS	ECOD	SF	RSIF
nci1	graph	0.48	0.56	0.46	0.49	0.50	0.51
aids		0.92	0.83	0.96	0.92	0.99	0.99
enzymes		0.76	0.61	0.68	0.72	0.66	0.59
proteins		0.54	0.58	0.35	0.67	0.68	0.66



Conclusions

- Distance based methods work exceptionally well after finding optimal distances.



Conclusions

- Distance based methods work exceptionally well after finding optimal distances.
- Good results in numerical and mixed realm shows RSIF potential.



Conclusions

- Distance based methods work exceptionally well after finding optimal distances.
- Good results in numerical and mixed realm shows RSIF potential.
- It seems there is neither best nor worst algorithm.



Conclusions

- Distance based methods work exceptionally well after finding optimal distances.
- Good results in numerical and mixed realm shows RSIF potential.
- It seems there is neither best nor worst algorithm.
- Outliers can exhibit themselves in different ways. Different nature of outlier, different measure needed



Conclusions

- Distance based methods work exceptionally well after finding optimal distances.
- Good results in numerical and mixed realm shows RSIF potential.
- It seems there is neither best nor worst algorithm.
- Outliers can exhibit themselves in different ways. Different nature of outlier, different measure needed
- Being unsupervised, most outlier detection methods heavily depend on the data representation quality



Conclusions

- Distance based methods work exceptionally well after finding optimal distances.
- Good results in numerical and mixed realm shows RSIF potential.
- It seems there is neither best nor worst algorithm.
- Outliers can exhibit themselves in different ways. Different nature of outlier, different measure needed
- Being unsupervised, most outlier detection methods heavily depend on the data representation quality
- RSIF can easily imitate ISF and IF by setting it's projections properly.



Future work

- Evaluation of other strategies for selecting reference pairs for projections



Future work

- Evaluation of other strategies for selecting reference pairs for projections
- Finding best distances in unsupervised fashion.



Future work

- Evaluation of other strategies for selecting reference pairs for projections
- Finding best distances in unsupervised fashion.
- Exploration of more distance functions to improve work on complex data.



Future work

- Evaluation of other strategies for selecting reference pairs for projections
- Finding best distances in unsupervised fashion.
- Exploration of more distance functions to improve work on complex data.
- Select as little and as good samples for distance calculation to reduce computational complexity.



Future work

- Evaluation of other strategies for selecting reference pairs for projections
- Finding best distances in unsupervised fashion.
- Exploration of more distance functions to improve work on complex data.
- Select as little and as good samples for distance calculation to reduce computational complexity.
- Case study



You can try RSIF yourself





Conclusions

- Distance based methods work exceptionally well after finding optimal distances.
- Good results in numerical and mixed regime shows RSIF potential.
- It seems there is no best nor worst algorithm
- Outliers can exhibit themselves in different ways. Different nature of outlier, different measure needed
- Being unsupervised, most outlier detection methods heavily depend on the data representation quality
- RSIF can easily imitate ISF and IF by setting it's projections properly.

Future work

- Evaluation of other strategies for selecting reference pairs for projections
- Finding best distances in unsupervised fashion.
- Exploration of more distance functions to improve work on complex data.
- Select as little and as good samples for distance calculation to reduce computational complexity.
- Case study