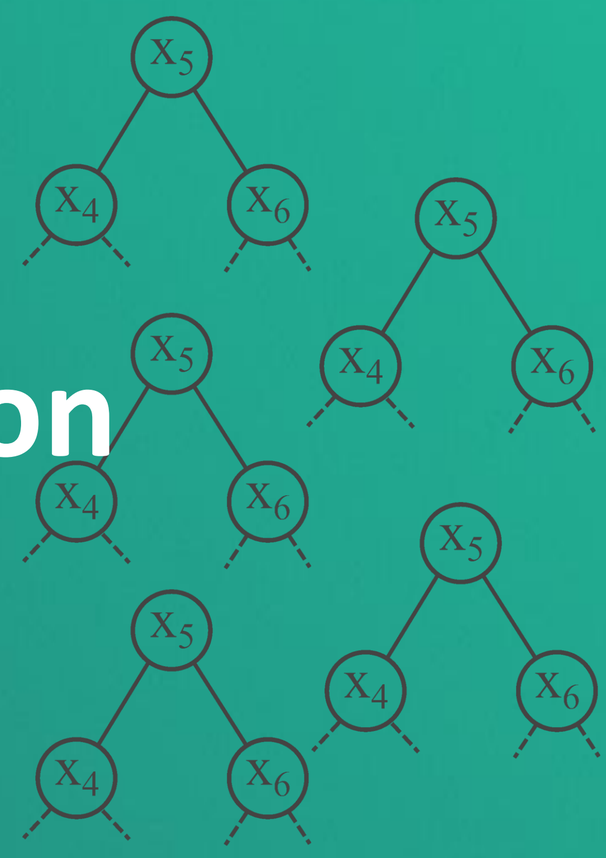
Equiping **Isolation Forest** with **multi-modal similarity projection** improves outlier detection



RSIF: Random Similarity Isolation Forest

CHWILCZYNSKI Sebastian, BRZEZINSKI Dariusz

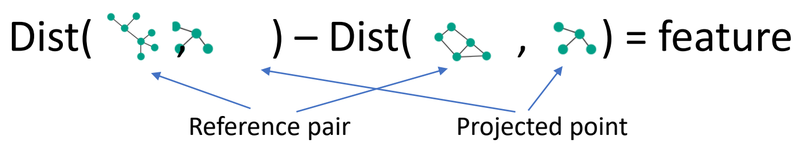
MOTIVATION

- Multimodal data - objects are described by many feature types (graph, numerical, set)

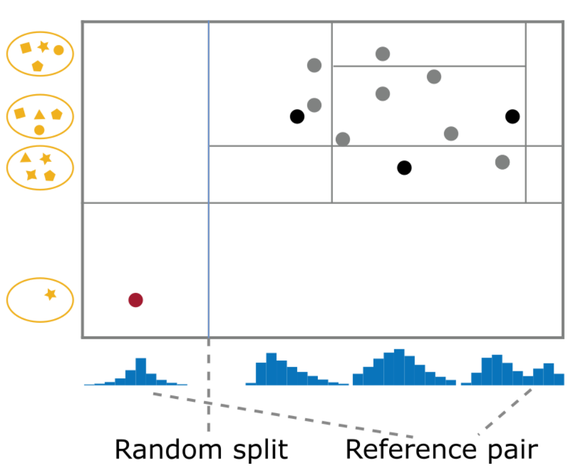
- Detecting outliers in multimodal data requires either omitting some data, creating separate models, or simplifying the data

APPROACH

- By utilizing distance based projections, we can handle features of arbitrary data types while retaining each feature's characteristic



- RSIF creates random splits in the projection space of each feature. Objects, which are separated with just few splits are treated as outliers.



RESULTS

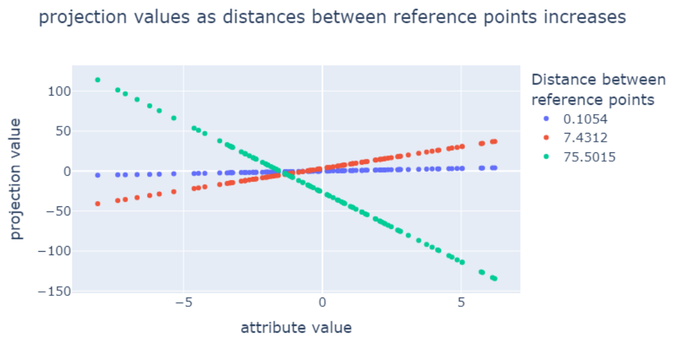
LOF (Local Outlier Factor), SF (Similarity Forest) and RSIF results were obtained by finding optimal distance functions.

Dataset	Type	AUC					
		iForest	LOF	HBOS	ECOD	**SF**	**RSIF**
glass	numeric	0.72	0.74	0.76	0.60	0.77	**0.80**
letter		0.61	**0.92**	0.59	0.55	0.74	0.77
musk		1.00	0.60	1.00	0.96	**1.00**	1.00
annthyroid		0.80	0.71	0.61	0.78	0.70	**0.84**
satimage		0.99	0.34	0.97	0.96	0.98	**0.99**
thyroid		0.98	0.48	0.95	0.98	0.98	**0.98**
vowels		0.69	**0.93**	0.66	0.59	0.63	0.91
waveform		0.73	0.74	0.69	0.59	**0.80**	0.76
wbc		**1.00**	0.92	0.99	1.00	0.99	**1.00**
wdbc		0.99	0.98	0.99	0.97	**1.00**	0.99
wilt		0.46	**0.69**	0.41	0.39	0.35	0.53
aid	categorical	0.65	0.58	0.66	**0.66**	0.63	0.62
apascal		0.49	0.55	0.66	0.66	**0.67**	0.64
cmc		0.57	0.51	**0.59**	0.59	0.52	0.57
reuters		0.98	0.95	0.99	**0.99**	0.98	0.98
solarflare		0.80	0.55	0.84	0.84	**0.84**	0.83
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
earthquakes	time series	0.61	0.57	0.49	0.56	0.51	**0.64**
aibo		0.50	**0.63**	0.50	0.46	0.59	0.53
ECGFiveDays		0.80	**0.91**	0.75	0.67	0.69	0.79
MiddlePhalanx		0.68	**0.75**	0.62	0.53	0.65	0.62
amazon	text	0.52	**0.55**	0.51	0.52	0.50	0.50
imdb		0.47	**0.52**	0.47	0.47	0.47	0.50
yelp		0.54	**0.59**	0.55	0.56	0.51	0.54
cifar	image	0.73	**0.73**	0.68	0.71	0.69	0.71
fashionmnist		**0.84**	0.74	0.76	0.83	0.82	0.83
svhn		0.56	**0.66**	0.48	0.54	0.57	0.55
item	sequences	0.83	0.83	**0.84**	0.84	0.67	0.77
length		0.85	0.87	0.92	**0.92**	0.86	0.84
order		0.53	0.53	0.55	**0.59**	0.46	0.56
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

DISCUSSION

- RSIF is highly flexible and can work both as IF (Isolation Forest) and SF when proper projections are selected.
- Distance based outlier detection methods can work exceptionally well after finding optimal distance functions.
- 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.

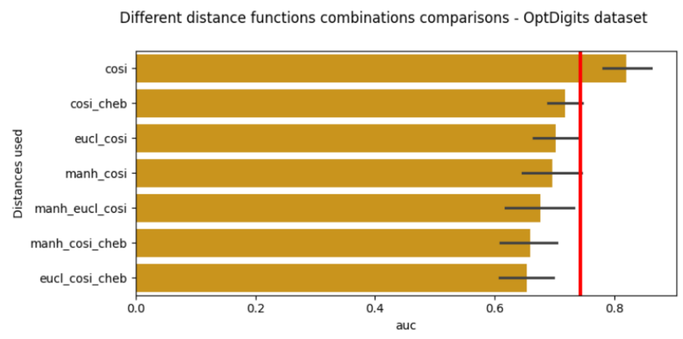
Why reference pair selection strategy is important?



Future work:

- Evaluation of other strategies. Currently only furthest points within currently considered node was checked.

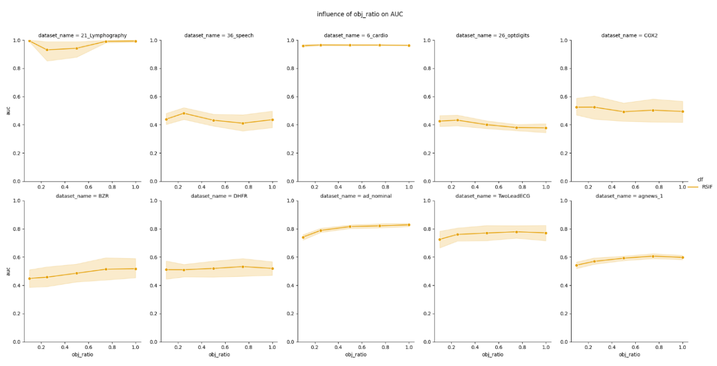
Why distance selection is important?



Future work:

- Finding best distances in unsupervised fashion. Currently it is done via nested Cross Validation.
- Exploration of more distance functions

How to overcome costly distance calculations?



Future work:

- Find a way to select as little and as good samples for distance calculation as possible

POZNAN UNIVERSITY OF TECHNOLOGY

Group of Horribly Optimistic STatisticians

Take a picture to visit GitHub repository