Proximity Isolation Forests

Antonella Mensi¹, Manuele Bicego¹, David M.J. Tax²

¹ University of Verona, Italy (antonella.mensi@univr.it) 2 TU Delft, The Netherlands

Motivation

- Isolation Forests: successful method for outlier detection based on Random Forests.
- Isolation Forests+extensions work only with vectorial data.
- Several outlier detection problems deal with non-vectorial data such as: sequences, images, etc.
- There exist many distance measures for non-vectorial data.
- We can work directly with non-vectorial data by employing pairwise distances.

Proposal: Proximity Isolation Forest: RF-based methodology for outlier detection. It works with all types of

No RF-based method for outlier detection exists!

data for which a distance measure is defined.

Proximity Isolation Forests (PIF)

Proximity Isolation Tree (PIT): recursively built on a distance matrix \mathbf{D} containing pairwise distances.

• Two ways to traverse a node n in a PIT:



• Five ways to split a node n in a PIT:

-**R-1P**, **R-2P**: random selection of a pair of prototype and threshold (or prototypes).

 $-\mathbf{O}-\mathbf{1PS}_D$, $\mathbf{O}-\mathbf{2PS}_D$, $\mathbf{O}-\mathbf{2PS}_P$: Choice of the best pair based on an *optimization* function.

How to optimize the split choice?

1. Isolation of outliers _____ decrease in variance.

2. No features **mathemathe in the set of the**

3. We measure the scatter **sparseness** of the distance values.





•8 datasets. 10 repetitions per experiment. Accuracy measure: AUC.

• Comparison with 6 distance and densitybased methods.

PIF: Guideline-based and *best* parametrizations.

(More details and results in the paper.)

Dataset	NNd	KNNd	KNNd-Av	LOF	LOF-Range	K-Centers	PIF
DelftPedestrians	0.524	0.567	0.534	0.553	0.579	0.629	0.799 (0.799)
DelftGestures	0.419	0.440	0.388	0.547	0.579	0.643	0.955 (0.976)
WoodyPlants	0.451	0.390	0.383	0.659	0.639	0.714	0.910 (0.930)
Pendigits	0.505	0.490	0.497	0.492	0.466	0.600	0.745 (0.755)
Zongker	0.566	0.476	0.422	0.564	0.514	0.752	0.796 (0.811)
ChickenPieces	0.462	0.462	0.425	0.456	0.444	NaN	0.825 (0.846)
Protein	0.413	0.820	0.798	0.922	0.919	0.861	0.984 (0.985)
Flowcyto	0.498	0.448	0.462	0.619	0.623	0.629	0.708 (0.737)
Average	0.479	0.524	0.501	0.602	0.596	0.688	0.840 (0.855)