# Proximity Isolation Forests 

Antonella Mensi ${ }^{1}$, Manuele Bicego ${ }^{1}$, David M.J. Tax ${ }^{2}$<br>${ }^{1}$ University of Verona, Italy (antonella.mensi@univr.it)<br>${ }^{2}$ 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.

No RF-based method for outlier detection exists!
-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 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).
$-\mathrm{O}-1 \mathrm{PS}{ }_{D}, \mathrm{O}-2 \mathrm{PS}_{D}, \mathrm{O}-2 \mathrm{PS}_{P}$ : Choice of the best pair based on an optimization function.


How to optimize the split choice?

1. Isolation of outliers $\longrightarrow$ decrease in variance.
2. No features $\longrightarrow$ no variance.
3. We measure the scatter $\longrightarrow$ sparseness of the distance values.


## Results

- 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 | $\mathbf{0 . 7 9 9}(0.799)$ |
| DelftGestures | 0.419 | 0.440 | 0.388 | 0.547 | 0.579 | 0.643 | $\mathbf{0 . 9 5 5}(0.976)$ |
| WoodyPlants | 0.451 | 0.390 | 0.383 | 0.659 | 0.639 | 0.714 | $\mathbf{0 . 9 1 0}(0.930)$ |
| Pendigits | 0.505 | 0.490 | 0.497 | 0.492 | 0.466 | 0.600 | $\mathbf{0 . 7 4 5}(0.755)$ |
| Zongker | 0.566 | 0.476 | 0.422 | 0.564 | 0.514 | 0.752 | $\mathbf{0 . 7 9 6}(0.811)$ |
| ChickenPieces | 0.462 | 0.462 | 0.425 | 0.456 | 0.444 | NaN | $\mathbf{0 . 8 2 5}(0.846)$ |
| Protein | 0.413 | 0.820 | 0.798 | 0.922 | 0.919 | 0.861 | $\mathbf{0 . 9 8 4}(0.985)$ |
| Flowcyto | 0.498 | 0.448 | 0.462 | 0.619 | 0.623 | 0.629 | $\mathbf{0 . 7 0 8}(0.737)$ |
| Average | 0.479 | 0.524 | 0.501 | 0.602 | 0.596 | 0.688 | $\mathbf{0 . 8 4 0}(0.855)$ |

