The Journal of Open Source Software

eif: Extended Isolation Forest

Sahand Hariri[1] **and Matias Carrasco Kind**[2]

1 Mechanical Science and Engineering, University of Illinois at Urbana-Champaign, USA **2** National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign. 1205 W Clark St, Urbana, IL USA 61801

DOI: 00.00000/joss.00000

Software

- Review
- Repository
- Archive

Submitted: 00 January 0000
Published: 00 January 0000

Summary

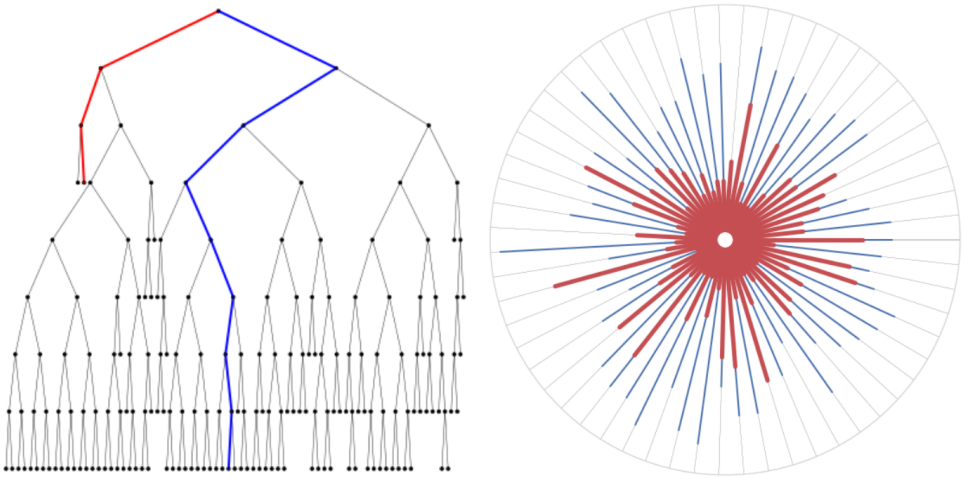
We present an extension to the model-free anomaly detection algorithm, Isolation Forest. This extension, named Extended Isolation Forest (EIF), improves the consistency and reliability of the anomaly score produced by standard methods for a given data point. We show that the standard Isolation Forest produces inconsistent scores using score maps, and that these score maps suffer from an artifact produced as a result of how the criteria for branching operation of the binary tree is selected. We propose two different approaches for improving the reliability of anomaly detection. First we propose methods for transforming the data before the creation of each tree in the forest. Second, which is the preferred method of this paper, is to allow the slicing of the data to use hyperplanes with random slopes. This approach results in improved score maps. We show that the consistency and reliability of the algorithm is much improved using this extension by looking at the variance of scores of data points distributed along constant score lines. We find no appreciable difference in the rate of convergence nor in computational time between the standard Isolation Forest and EIF which highlights its potential as anomaly detection algorithm

`eif`

Acknowledgements

MCK is supported by the National Science Foundation under Grant NSF AST 07-15036 and NSF AST 08-13543

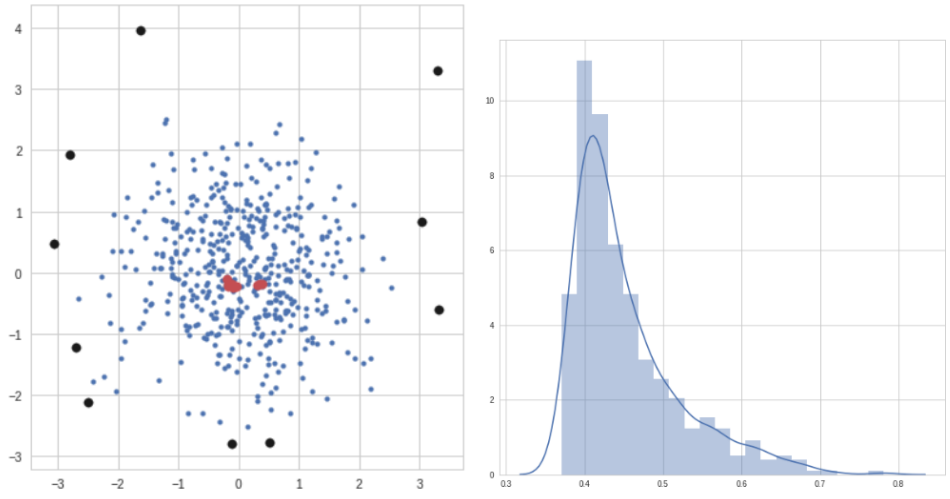
References



(a) Representation of a single tree in a forest.

(b) Representation of a full forest where each radial line corresponds to a tree.

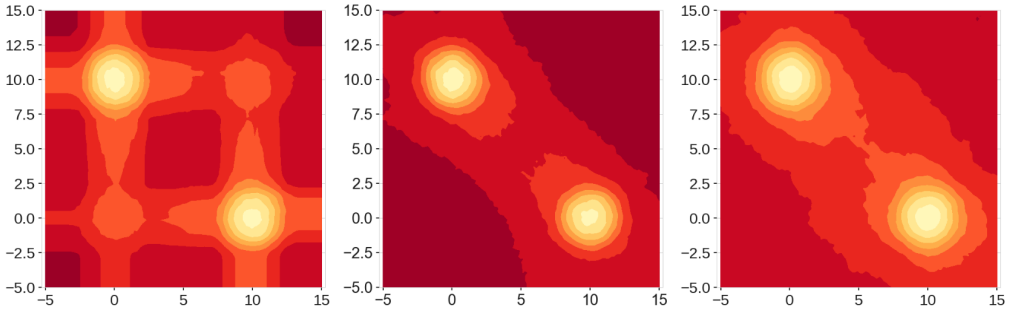
Figure 1: a) Shows an example tree formed from the example data while b) shows the forest generated where each tree is represented by a radial line from the center to the outer circle. Anomalous points (shown in red) are isolated very quickly,which means they reach shallower depths than nominal points (shown in blue).



(a) Single 2D blob with anomalies (black) and nominal points (red).

(b) Anomaly score Distribution

Figure 2: a) Shows the dataset used, some sample anomalous data points discovered using the algorithm are highlighted in black. We also highlight some nominal points in red. In b), we have the distribution of anomaly scores obtained by the algorithm.



(a) Standard IF (b) Rotated IF (c) Extended IF

Figure 3: Comparison of the standard Isolation Forest with rotated Isolation Forest, and Extended Isolation Forest for the case of two blobs.