Isolation Forest for Anomaly Detection

Carol Stanton[1], Gilad Katz[2], and Dawn Song[2]
[1]Contra Costa College, [2]University of California, Berkeley

Abstract : Isolation Forest is a novel approach for anomaly detection proposed by Liu, Ting, and Zhou[1]. Using partitioning, binary search trees and data sampling, this method promises to be a fast and effective anomaly detector. We test the authors' claims regarding the performance of Isolation Forest on a variety of data sets. We also compare the Isolation Forest algorithm to other standard classifiers.

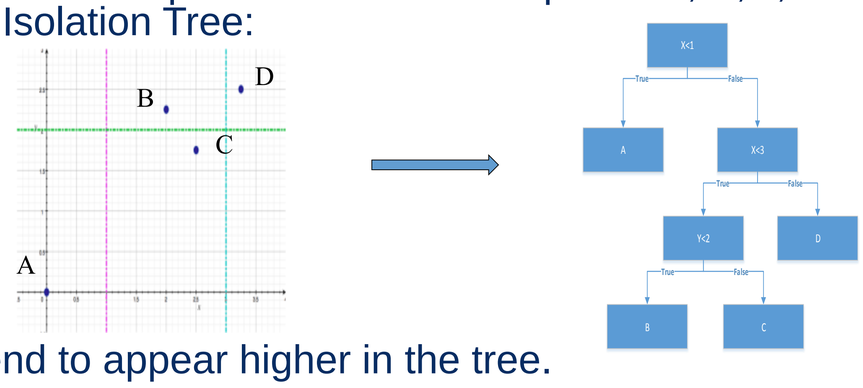
What Is Anomaly Detection?

- Anomalies in a data set are instances that are few in number and different from the majority of the instances.
- The ability to detect anomalous data is important in fields such as medicine, environmental science, and engineering. Anomaly detection is of primary importance in the area of cybersecurity.
- The drawbacks of most existing anomaly detection algorithms are that they are resource intensive, or have high computational complexity, or both.

How Isolation Forest Works

The Isolation Forest algorithm has linear time and space complexity. It relies on the simple idea that anomalous data instances can be isolated from normal data via recursive partitioning of the dataset.

Example : A recursive partition of the set of points A, B,C, and D produces an Isolation Tree:



Anomalies tend to appear higher in the tree.
An Isolation Forest is a collection of Isolation Trees.

- The algorithm uses subsamples of the data set to create an isolation forest.
- An anomaly score is computed for each data instance based on its average path length in the trees.
- Scores are normalized from 0 to 1; a score of 0 means the point is definitely normal, 1 represents a definite anomaly.
- Data instances whose anomaly scores are above a desired threshold are classified as anomalies.

Experimental Setup

Data : Algorithms were tested on over 30 datasets; "Real-world" data was obtained at [2]; synthetic data was generated in Java.
Dataset features:
- sizes : 35 to nearly 1 million
- dimensions : 2 to nearly 300
- percentage of anomalies : less than 1% to 49%

Environment : Algorithms were tested on the open-source machine learning platform Weka, run in their default configurations.

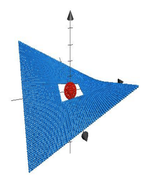
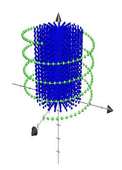
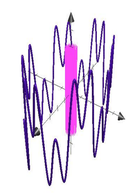
Results

- Summary of average performance results for Isolation Forest and the comparison algorithms:

Algorithm	number of datasets tested	average overall accuracy	average anomaly precision	average anomaly recall	average ROC area (AUC)
Isolation Forest	30	66%	40%	72%	0.77
Random Forest	30	91%	73%	69%	0.867
Naïve Bayes	30	89%	66%	69%	0.841
LibSVM	26	89%	50%	42%	0.703
LOF	20	41%	28%	59%	0.514

Note that Random Forest is best-performing overall, although Isolation Forest has a higher average anomaly recall in its default configuration.

- Isolation Forest – performance on geometric data sets

saddle	solenoid	carousel
5314 data points	754,528 data points	130,816 data points
19% anomalies	33% anomalies	36% anomalies
		
39% accuracy 0% recall	87% accuracy 100% recall	99% accuracy 98% recall

Conclusions

The Isolation Forest algorithm is a simple and elegant approach to anomaly detection. It is an efficient anomaly detector in terms of both time and space; the cost of this efficiency is reduced accuracy and precision as compared to Random Forest.

Future Work

Modify Isolation Forest algorithm to handle datasets with categorical attributes.

Acknowledgements

The E[3]S staff who ran the UCB Context Based RET Program.
Carol Stanton wishes to thank Drs. Gilad Katz and Dawn Song for their guidance.

References :

[1] Liu, F.T., Ting, K.M., and Zhou, Z.-H., Isolation-Based Anomaly Detection, ACM Transactions on Knowledge Discovery from Data, Vol. 6, No. 1, Article 3, Publication Date: March 2012.
[2] "UCI Machine Learning Repository: Data Sets." *UCI Machine Learning Repository: Data Sets*. Web.

Contact Information
cstanton@contracosta.edu ; (510) 215-3817

Support Information
This work was funded by National Science Foundation Award EEC-1405547.