Stella White
Pattern Recognition
Fall 2024

Machine Learning Algorithms for Anomaly Detection

Summary:

For my project, I will compare four different ML algorithms (Support Vector Machines, Random Forest, k-nearest Neighbour, and K-means Clustering) on their accuracy in detecting data anomalies. I will compare the results to a paper called "Evaluation of Machine Learning-based Anomaly Detection Algorithms on an Industrial Modbus/TCP Data Set." This paper also compares these ML algorithms. I will use a different dataset and add more result metrics. These metrics are used in "Utilizing Machine Learning Approaches for Anomaly Detection in Industrial Control Systems." I will use two datasets, one with anomalies and one without anomalies.

Approach:

The four ML algorithms (Support Vector Machines, Random Forest, k-nearest Neighbour, and K-means Clustering) will be used to detect if there is data anomaly and will be evaluated on accuracy, precision, recall,f1-score, AUC-ROC, AUC-PR, false positives, and false negatives. These metrics are commonly used for binary classification. In this project, a positive will indicate an anomaly. Support Vector machines will divide the set by a linear kernel. If the set cannot be divided by the kernel, the set will be mapped into a higher dimensional feature space. In Random Forest, 2000 decision trees will be used. K-means clustering will use 2 clusters.

Data:

I will be using two public datasets from Numenta Anomaly Benchmark as found here: numenta/NAB: The Numenta Anomaly Benchmark. I plan on using the artificially-generated datasets of datastreams with and without anomalies. For the dataset without anomalies, I will be using art_daily_no_noise.csv, art_daily_perfect_square_wave.csv, art_daily_small_noise.csv, art_flatline.csv, and art_noisy.csv. For the dataset with anomalies, I will be using art_daily_flatmiddle.csv, art_daily_jumpsdown.csv, art_daily_jumpsup.csv, art_daily_nojump.csv, art_increase_spike_density.csv, and art_load_balancer_spikes.csv. I plan on using 80% of each data stream dataset for training and the other 20% for testing.

Expected Results:

I expected Support Vector Machines and Random Forest to perform the best. These performed best in the "Evaluation of Machine Learning-base Anomaly Detection Algorithms on an Industrial Modbus/TCP Data Set" paper and are commonly used, such as in the "Utilizing Machine Learning Approaches for Anomaly Detection in Industrial Control Systems." The AUC-ROC score will measure the tradeoff between false positives and falses negatives, while the AUC-PR score will measure the tradeoff between precision and recall.