Anomaly Detection with StrOUD Algorithm

CS584 Homework 4

Edvin Beqari and Doug Williamson

GMU November 22, 2020

This project focused on anomaly detection using the Strangeness-Based Outliner Detection (StrOUD) methodology using Local Outlier Factor (LOF) to measure strangeness. The LOF works similarly to the Nearest Neighbor algorithm in that it first identifies the k nearest data points and calculates the reachability distance to each of those neighbors. Then it calculates the reachability distance of each of those neighbors to their respective k-neighbors to provide a density function. The StrOUD methodology then compares the density of points near our new data point to the density of points around its neighbors to determine if the target point is, or is not, an outlier. This project uses the StrOUD methology to analyze centrifuge data for anomalies.

Team Ranking

Our resulting application was submitted under the user name of "*beqarie*." We submitted two different versions, one using Fast Fourier Transform (FFT) and one without FFT. The results were as follows:

- *Submission score with FFT*: 0.96 (*Submission Date*: Nov. 23, 2020, 9:31 AM)
- *Submission score w/o FFT*: 1.00 (*Submission Date*: Nov. 23, 2020, 9:35 AM)

Approach

For our approach, we developed the StrOUD classifier and LOF calculator as required. To train and test our model, we were able to use all of the provided non-empty data files. We calculated the AUC values from ROC plots as our measure of effectiveness.

We also added two additional elements to our solution. First, we performed a full cross-validation to select the best k value for our LOF calculations. Our cross validation process and results are described in more detail in the next section.

Second, we performed out analysis using two different approaches to evaluate the effects on performance and accuracy. For our first run, we pre-processed the input data using a Fast Fourier Transform (FFT). For the second run, we utilized the raw data without any pre-processing. The results of both are shown throughout this report.

Validation Methodology

We implemented the Stratified K-fold cross validation to determine the best k value for our LOF calculations. We ran the algorithm for values of k from 2 to 400 using all the training data. The results of our cross validation are as follows:

- *With FFT best k*: 32 (*auc-roc score:* 0.94)
- *W/O FFT best k*: 30 (*auc-roc score:* 0.96)

Displaying all 399 data points would obscure the details, so Figures 1, and 2, show just the portion of the results which include our optimal values.

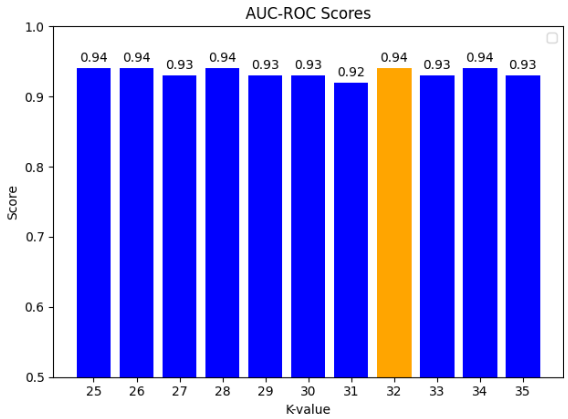

***Figure 1**: Cross-validation results with FFT Magnitude*

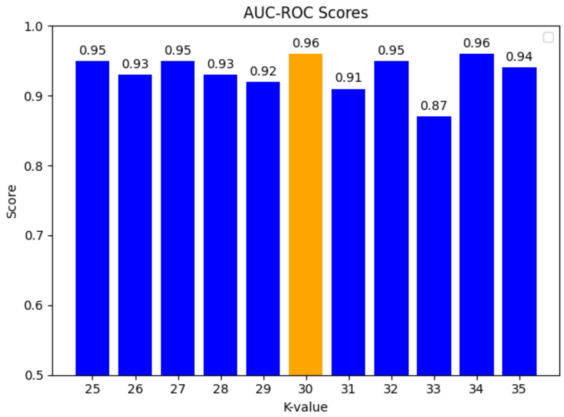

***Figure 2**: Cross-validation results without FFT Magnitude*

Once the optimal k values were determined for each of our LOF calculations, we generated the ROC curves for each and calculated the AUC metrics. The results are shown in Figures 3, and 4.

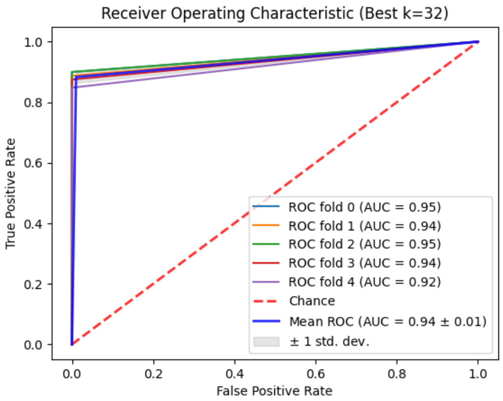

***Figure 3**: ROC curve for best k with FFT*

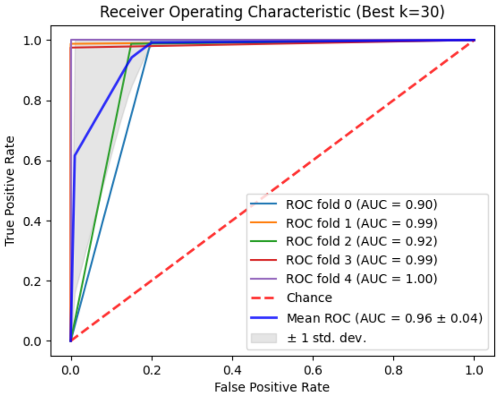

***Figure 4**: ROC curve for best k without FFT*

We also wanted to plot the StrOUD results to see if we could visualize the results. The original data values included roughly 20,000 dimensions each, so plotting this was not feasible. To compensate, we performed Principal Component Analysis (PCA) on the raw data to reduce it to two dimensions so that it could be plotted. The resulting plots appear in Figures 5, and 6. The dots represent the data points and the red circles mark the ones designated as outliers. We were pleasantly surprised to see that there was a detectable difference, with the FFT plot clearly showing more “outliers” like in the region in the upper right marked with the arrow. While a two-dimensional plot of points with 20,000 dimensions remains notional at best, the plots do show analysis using the pre-processed FFT data is predicting more outliers than the analysis on the raw data, which is consistent with our other measures.

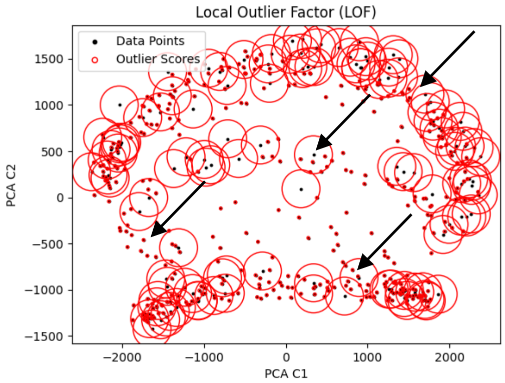

***Figure 5**: PCA dimensionality reduction on signal points with FFT*

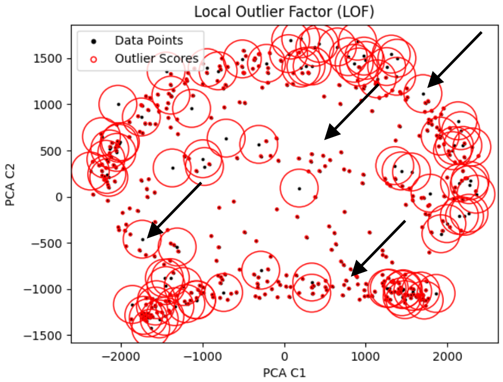

***Figure 6:** PCA dimensionality reduction on signal points without FFT*

Conclusions

Development of the analysis was more complex than originally anticipated with a lot of details required to make it work. However, the final result showed that it was possible to utilize the StrOUD process to detect the anomalies in data with high dimensionality. By examining both the raw data and data pre-processed with FFT, we also demonstrated the desirability of trying different approaches in order to tune the process for maximum accuracy.