Anomaly Detection Algorithims

Nikhil Muthukrishnan

30 October 2018

Objective: Develope framework for experimenting with Anomaly Detection Algorithims

Contents:1 PI tag "SC3_FIC20267.pv" for 2500 datapoints and 1 algorithim "LOOP"

Sprint:PO_SPRINT_12_OCT_2018

Local Outlier Probabilities

Formal Definition:

LOOP computes a local density based on probabilistic set distance for observations, with a usergiven k-nearest neighbors. The density is compared to the density of the respective nearest neighbors,resulting in the local outlier probability. The values ranges from 0 to 1, with 1 being the greatest outlierness.

Below is the function:

```
outlier_score <- LOOP(dataset=a, k=10, lambda=3)
```

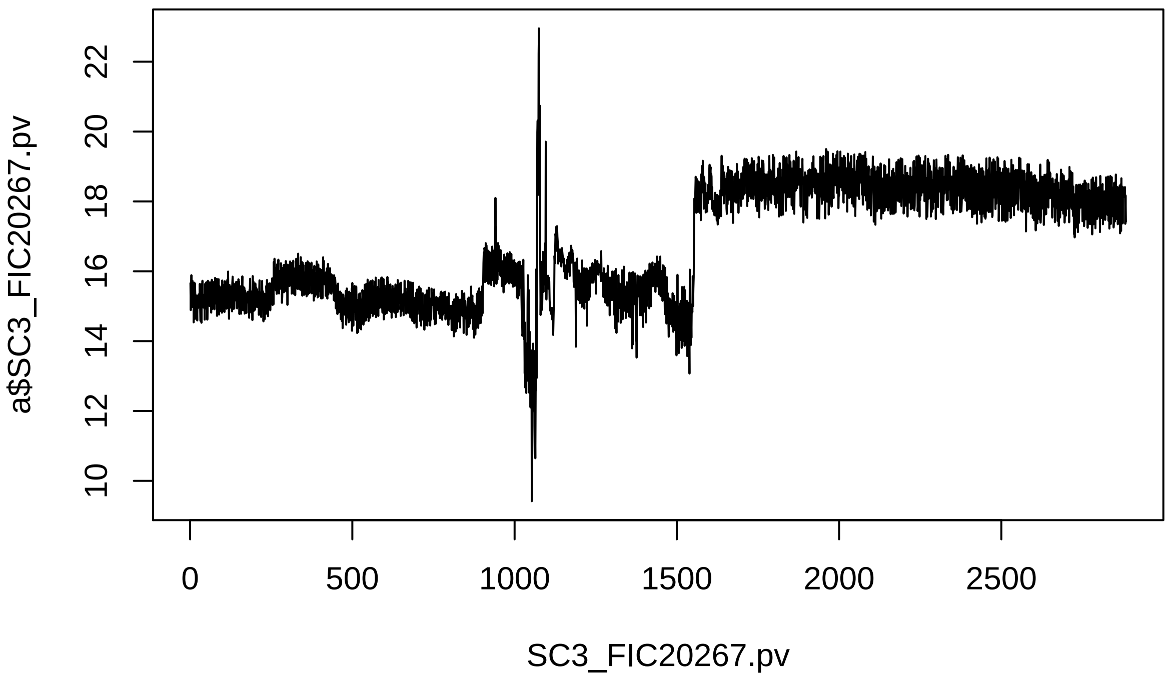
Below is the sample sensor reading "SC3_FIC20267.pv". As can be seen there are 2 obvious anomalies at ~1000 and ~1500 that should be detected by any anomaly detection algorithim.

The remainder of this document will apply the LOOP anomaly detection algorithim to this dataset with different values of "k" and "lambda".

k nearest neighbours represents the number of local data points that will be compared to the single data point to classify it as an outlier or not.

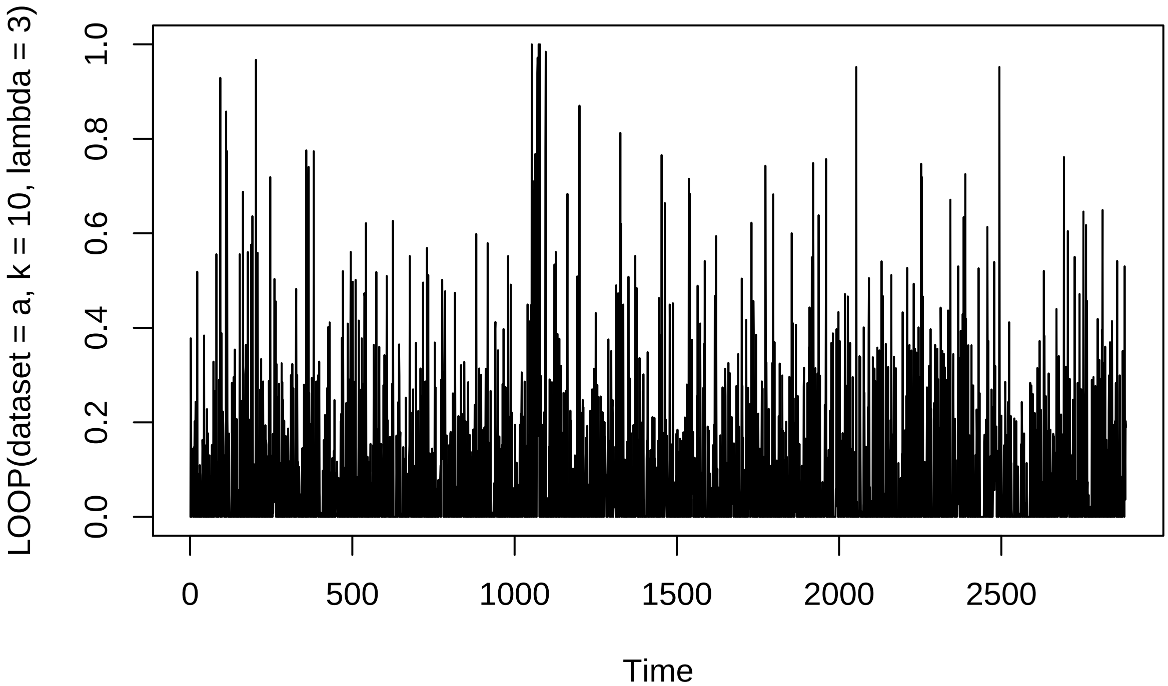
Lambda represents Multiplication factor for standard deviation. The greater lambda, the smoother results.Default is 3 as used in original papers experiments

The graph below is a simple timeseries plot of a sensor which will be compared to the outlier scores.



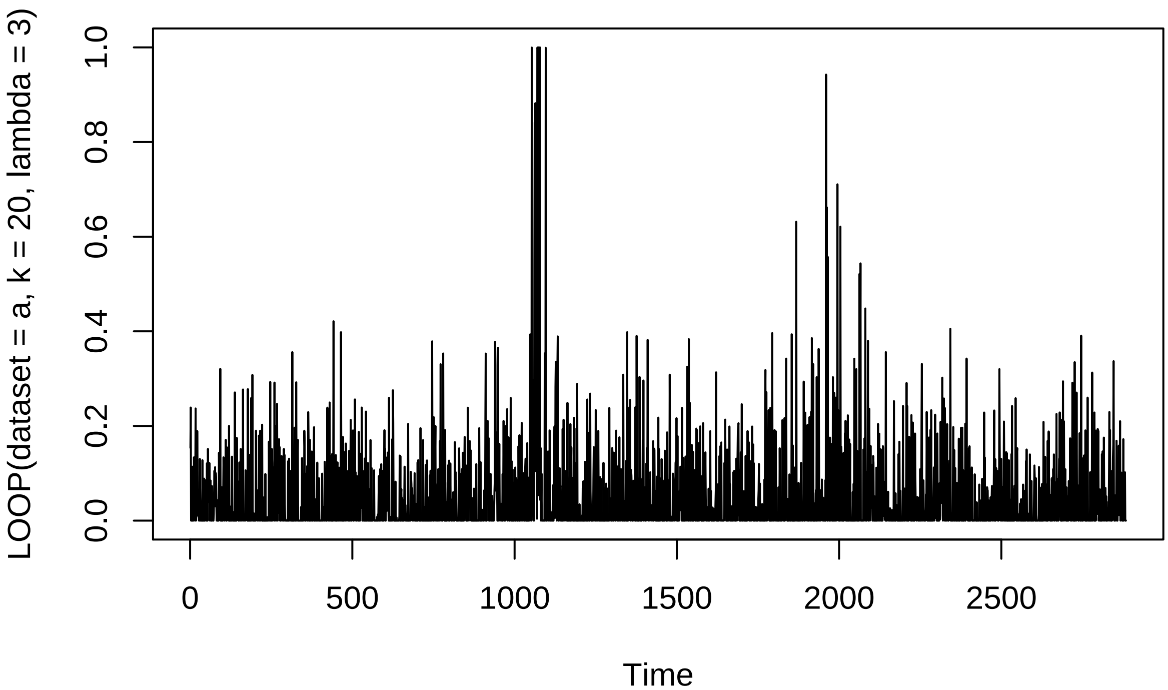
The graph below is the output of LOOP with k=10 & lambda=3. In this case the data points taken into consideration for datapoint number ~1000 is between datapoint ~995 to ~1005.As can be seen the algorithim has caputured the 2 datapoints viz. 1000 and ~1500 however it has also given a high outlier score for data points ~2000 and ~3000 which we consider as false positives.

```
ts.plot(LOOP(dataset=a, k=10, lambda=3))
```



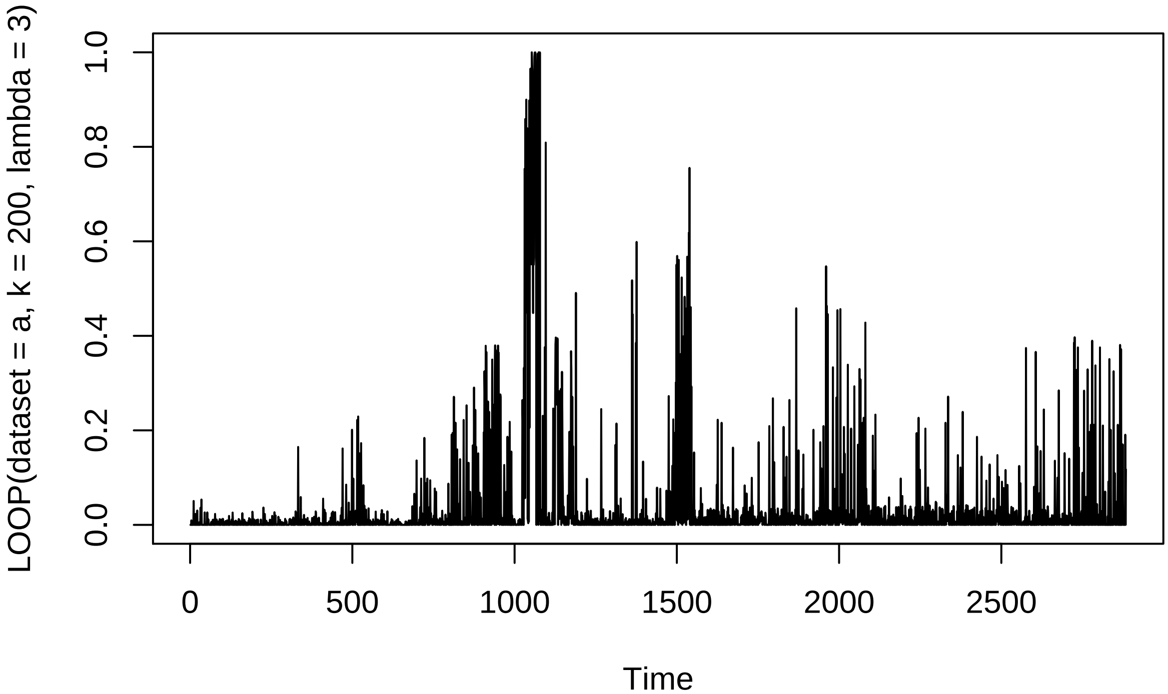
Here k has been set 20,lambda unchanged at 3 and it can be seen that ~1000 has been correctly classified as most outlying again but the next most outlying datapoints with outlier score of 0.8 has been awarded to data point ~2000 however we want the second highest outlier score to fall on datapoint ~1500, meaning our parameters are still not optimum.

```
ts.plot(LOOP(dataset=a, k=20, lambda=3))
```



The below graph is slightly better at detecting anomalies as the data point with highest outlier score is ~1000 as expected and the second most outlying data point is ~2000 and not our expected outlier viz ~1500. Hence we have 1 True Positive(~1000) but also 1 False Negative(~1500) and 1 True Negative (~2000). The parameters will be adjusted again for hopefully better results.

```
ts.plot(LOOP(dataset=a, k=200, lambda=3))
```



To speeden the research k has been taken at 200 which is the minimum value to classify datapoint 1500 as the second most outlying datapoint. The ideal k value has been achieved for this dataset and the 2 expected anomalies following this further experiments will be done on other sensor readings of a single batch.