Status Report: Anomaly Detection

Kushal Majmundar • 20.05.2019

Overview

Recent progress

- Trying anomaly detection methods like One Class SVM, PCA on proposed features

Problem to tackle

Determine the features that can help for anomaly detection

Method - Analyzing individual features

- Individual features were taken and were tested for outlier detection using One Class SVM
- Python library PyOd was used
- Every feature was appended into corresponding array. All the timestamps were taken in training set as in general case we will not know which timestamps are malicious
- If for a given user id, if the timestamp was mapped correctly to be malicious, it was counted in testing accuracy

Progress - Analyzing individual features

Low testing error

- Due to very less number of time stamps where the user behaves maliciously, the error on the testing error is very low
- Since the timestamps have been truncated it is very hard to determine

Non-zero testing error

- SUM SELL
- VOL SUM BUY
- CUM. SUM BUY

Link to results

Method - Analyzing features (2)

- A combination of features was now tried.
- We cannot combine any two features, the ones which have the same time stamps can only be taken together
- This leads to four clusters of time stamps
- All the possible combinations cannot be tested on as it is too time consuming
- Leave one feature out was followed to get the accuracies
- It is observed that if testing accuracy has some value, the training accuracy is very low for the malicious traders, hence low testing accuracy might not be much of a concern, as we have a measure who is behaving more maliciously.

Progress - Leave one out

Common observation: Features with non zero testing error have low training error, hence feature can be used

Low testing error again

- Due to very less number of time stamps where the user behaves maliciously

How many features to leave out

- All combinations of features cannot be tested upon
- Leave one out followed, common features can be later on taken

Imp features obtained

- Important features obtained are listed here
- The combinations are listed here
- The features obtained are the ones with low testing error

Link to results

Next steps

Try the model on these features

Try the CTO based model on accuracies obtained from this method

Try combinations of all these features

Try combinations of all these features to see which is the best combination that can be used instead of using them all

Problem being faced:

Only 19 malicious orders from 900 orders in total. Hence getting pointing out the malicious id is hard.