Outlier detection on network flow analysis

Quang-Vinh Dang

Noname manuscript No.
(will be inserted by the editor)

Outlier detection on network flow analysis

Quang-Vinh Dang

Received: date / Accepted: date

Abstract It is important to be able to detect and classify malicious network traffic flows such as DDoS attacks from benign flows. Normally the task is performed by using supervised classification algorithms. In this paper we analyze the usage of outlier detection algorithms for the network traffic classification problem.

Keywords outlier detection · classification · ddos detection

1 Introduction

A denial-of-service (DoS) attack is characterized "by an explicit attempt by attackers to prevent the legitimate use of a service" [Mirkovic and Reiher, 2004]. If the attackers coordinate the DDos traffic from multiple sources to perform the attack, it will be Distributed Denial-of-Service (DDos) [Koay et al, 2018].

Multiple studies have analyzed the detection and prevention strategies of DDoS attacks by using classification algorithms [Douligeris and Mitrokotsa, 2004; Fouladi et al, 2016; Koay et al, 2018; Alsirhani et al, 2018]. While these methods achieved a lot of success, they suffer from imbalanced dataset problem [Krawczyk, 2016] and lack of detecting unfamiliar flows. For instance, these techniques mail fail to detect a new DDoS attack technique that they did not see during the training period. Furthermore, supervised classification algorithms usually exhaust of data.

Outlier detection algorithms [Aggarwal, 2017] try to distinguish outlier points from normal traffic data. Hence, the techniques might be performed in unsupervised manner [Campos et al, 2016]. Furthermore, outlier detection

F. Author
TMC Data Science, Eindhoven, the Netherlands
E-mail: vinh.dang@tmc.nl

algorithms can deal well with extremed imbalanced dataset, such as 1:1000 ratio [Krawczyk, 2016].

In this paper we evaluate the performance of outlier detection algorithms on detecting DDos traffic. Our work is related to other evaluation studies, such as of [Hodge and Austin, 2018]. However, the authors of [Hodge and Austin, 2018] do not analyze the performance in case of imbalanced datasets.

2 Evaluation

2.1 Algorithms

We evaluated the following algorithms:

- **CBLOF** (Clustering-Based Local Outlier Factor) [He et al, 2003].
- **HBOS** (Histogram-Based Outlier Score) [Goldstein and Dengel, 2012].
- **IForest** Isolation Forest [Liu et al, 2008].
- **k-NN** (k Nearest Neighbors)[Ramaswamy et al, 2000].
- **MCD** (Minimum Covariance Determinant) [Rousseeuw and van Driessen, 1999].
- **OCSVM** (One-Class SVM) [Ma and Perkins, 2003].
- **PCA** (Principal Component Analysis) [Shyu et al, 2003].

2.2 Dataset

We used the dataset provided by [Ghorbani and Lashkari, 2018] that contains 464,976 samples that are assigned the labels "Attack" or "Benign". We consider these labels as ground truth. A set of 76 features is provided. Some of the features are:

- Flow.Duration
- Tot.Fwd.Pkts (Total Forward Packets)
- Tot.Bwd.Pkts (Total Backward Packets)
- Fwd.Pkt.Len.Max
- Fwd.Pkt.Len.Min

The "Attack" traffic contains 3.76% of the whole dataset (17,462 samples). In order to evaluate the performance of the outlier detection algorithms, we create different datasets with different "Attack" Ratio, range from 0.01 to 0.99.

To change the class ratio, we applied the following function to select a subset of the dataset:

```
def make_dataset(benign_ratio = 0.9,
        df_attack = df_attack,
        df_benign = df_benign):
    N_attack = len (df_attack.index)
    N_benign = len (df_benign.index)
```

```
n_attack = N_benign /
    (benign_ratio / (1-benign_ratio))
sample_df_attack = df_attack.
    sample(int(n_attack),
    replace=False)
res = pd.concat([df_benign, sample_df_attack])
res = shuffle (res)

res = res[[col for col in res
    if not len(set(df[col]))==1]]

#divide train/test
train = res.sample(frac = 0.7, random_state=200)
test = res.drop(train.index)
return (train,test)
```

3 Results

In this section we present the performance in term of *AUC* and *Accuracy* scores of each algorithms. Overall, while the *benign_ratio* increases the *Accuracy* scores increase for all algorithms, that can be explained by the imbalanced of the dataset. Another observation is that the performance of the algorithms are very similar between training and testing set due to the fact that these algorithms are all unsupervised. In term of *AUC* score we could see that IForest and PCA algorithms achieved the best scores while the *benign_ratio* increases.

4 Conclusions

In this paper we evaluate the performance of unsupervised outlier detection algorithms in detecting DDoS attacks. We showed that the outlier detection algorithms perform, particularly Isolation Forest or PCA-based algorithms, perform best if the proportion of outlier instances is small. This contrasts with popular classification algorithms. In the future we will focus on analyzing Isolation Forest and PCA algorithms in other scenarios.

References

Aggarwal CC (2017) Outlier Analysis, 2nd edn. Springer

Alsirhani A, Sampalli S, Bodorik P (2018) Ddos attack detection system: Utilizing classification algorithms with apache spark. In: NTMS, IEEE, pp 1–7

Campos GO, Zimek A, Sander J, Campello RJGB, Micenková B, Schubert E, Assent I, Houle ME (2016) On the evaluation of unsupervised outlier

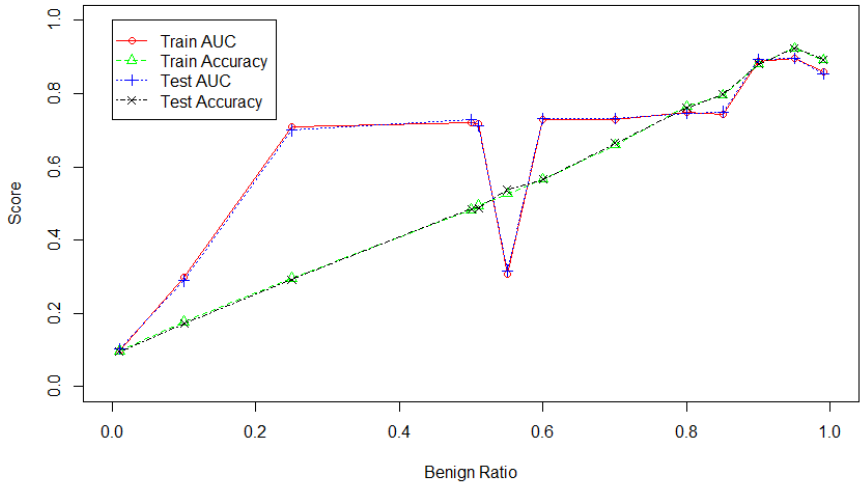

Fig. 1: CBLOF

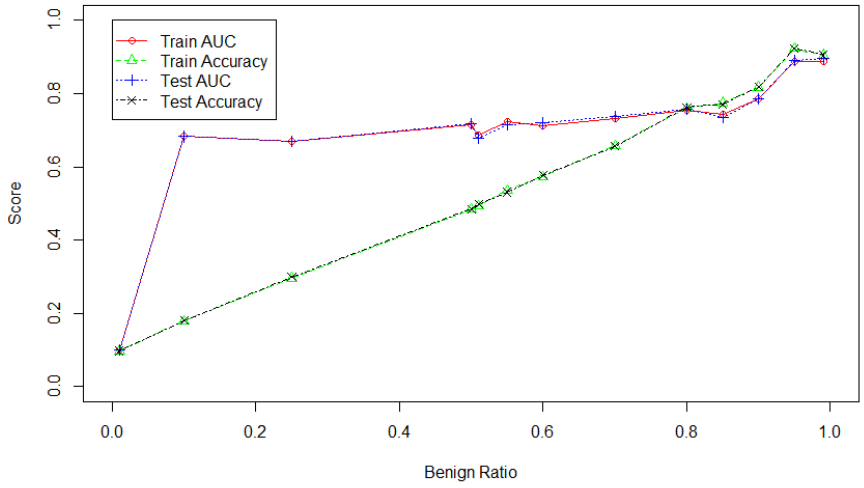

Fig. 2: HBOS

detection: measures, datasets, and an empirical study. Data Min Knowl Discov 30(4):891–927

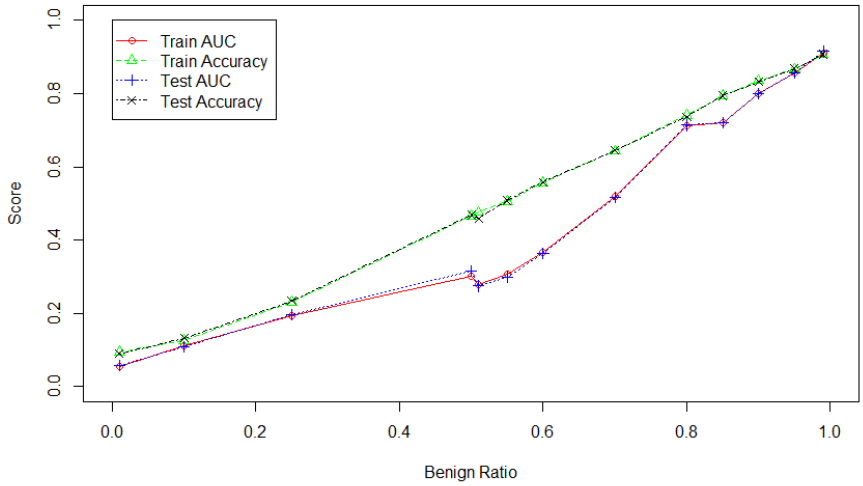

Fig. 3: IForest

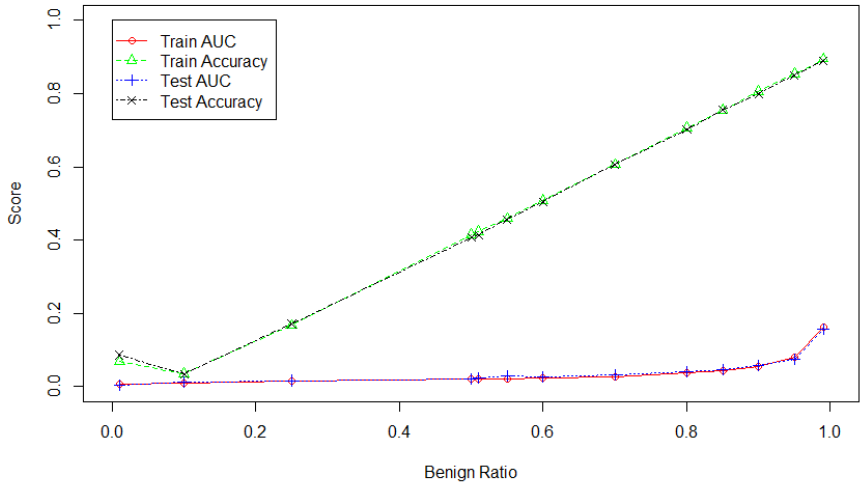

Fig. 4: kNN

Douligeris C, Mitrokotsa A (2004) Ddos attacks and defense mechanisms: classification and state-of-the-art. Computer Networks 44(5):643–666

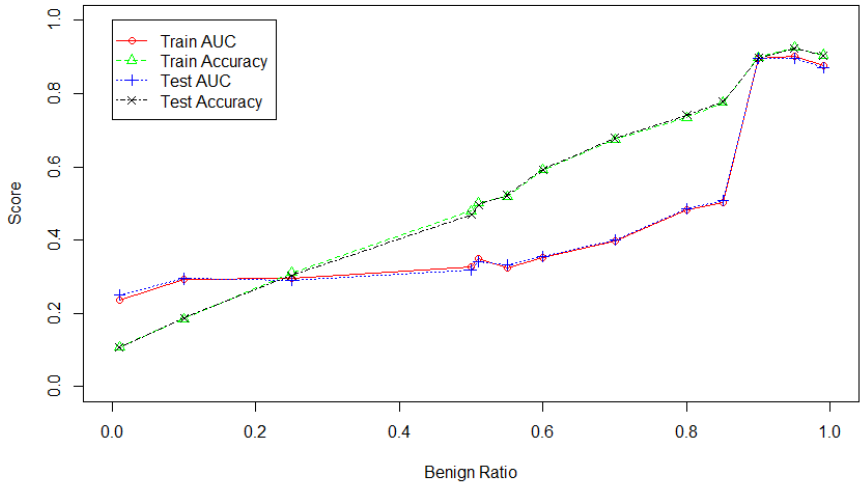

Fig. 5: MCD

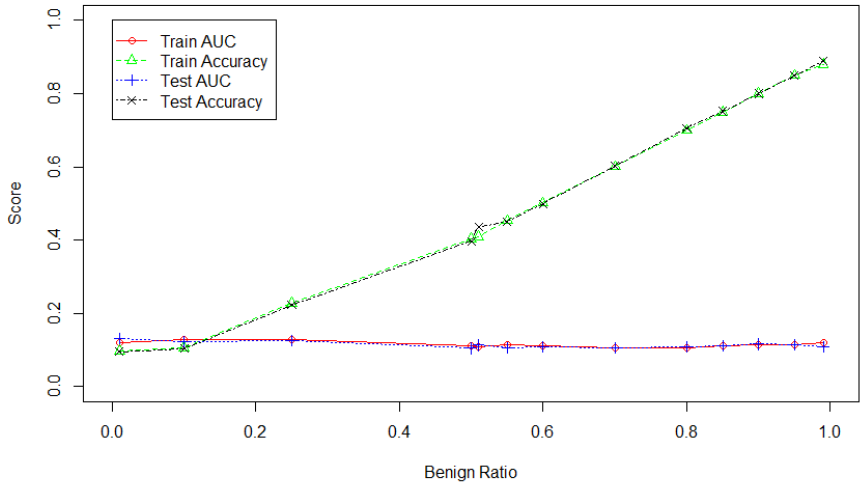

Fig. 6: OCSVM

Fouladi RF, Kayatas CE, Anarim E (2016) Frequency based ddos attack detection approach using naive bayes classification. In: TSP, IEEE, pp 104–107

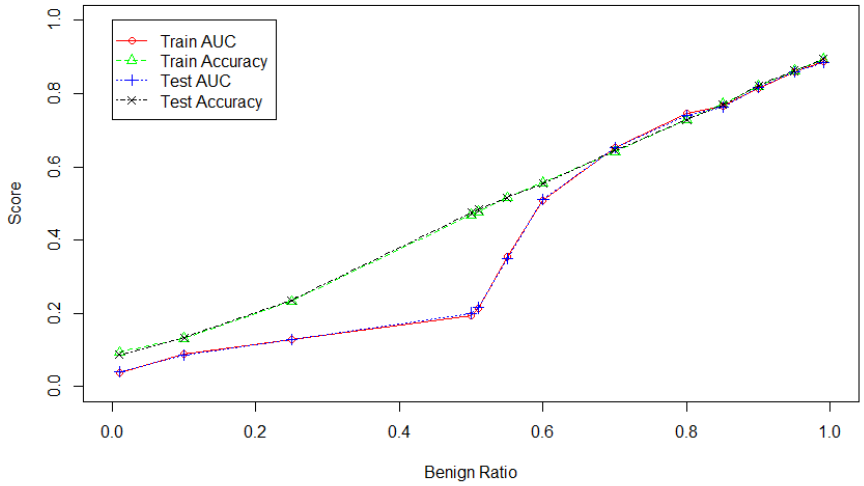

Fig. 7: PCA

Ghorbani A, Lashkari AH (2018) Cdmc2018 dataset: Ddos attacks detection for enterprise network security

Goldstein M, Dengel A (2012) Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. KI-2012: Poster and Demo Track pp 59–63

He Z, Xu X, Deng S (2003) Discovering cluster-based local outliers. Pattern Recognition Letters 24(9-10):1641–1650

Hodge VJ, Austin J (2018) An evaluation of classification and outlier detection algorithms. CoRR abs/1805.00811, URL `http://arxiv.org/abs/1805.00811`, `1805.00811`

Koay A, Chen A, Welch I, Seah WKG (2018) A new multi classifier system using entropy-based features in ddos attack detection. In: ICOIN, IEEE, pp 162–167

Krawczyk B (2016) Learning from imbalanced data: open challenges and future directions. Progress in AI 5(4):221–232

Liu FT, Ting KM, Zhou Z (2008) Isolation forest. In: ICDM, IEEE Computer Society, pp 413–422

Ma J, Perkins S (2003) Time-series novelty detection using one-class support vector machines. In: Neural Networks, 2003. Proceedings of the International Joint Conference on, IEEE, vol 3, pp 1741–1745

Mirkovic J, Reiher PL (2004) A taxonomy of ddos attack and ddos defense mechanisms. Computer Communication Review 34(2):39–53

Ramaswamy S, Rastogi R, Shim K (2000) Efficient algorithms for mining outliers from large data sets. In: SIGMOD Conference, ACM, pp 427–438

Rousseeuw PJ, van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41(3):212–223

Shyu ML, Chen SC, Sarinnapakorn K, Chang L (2003) A novel anomaly detection scheme based on principal component classifier. Tech. rep., MIAMI UNIV CORAL GABLES FL DEPT OF ELECTRICAL AND COMPUTER ENGINEERING