Anomaly Detection Challenges - Challenge III

Hamza Tahir (03670002) and Muhammad Hamza Usmani (03669506)

Technical University of Munich

1 Introduction

This brief report serves as a purpose to present and explain the methodologies applied to tackle the third challenge in the Practical: Anomaly Detection Challenges. Section 2 discusses the challenge task and the data set for the machine learning/anomaly detection task. Section 3 explains the approaches adopted for the task.

2 The Challenge

Machine Learning Task The machine learning task for this challenge is to determine if a network trace is a normal transaction or an attack.

Table 1. Decision Classes

Class	Representation
Normal	1
Attack	0

The Data set The dataset consists of network traces. There are 56,041 records in the training set with 43 network related features. The training dataset consists of 56,000 normal records and 41 attack rows, the attack rows consist of 9 different attack types, the machine learning task however for this challenge is to classify a record as attack or normal.

3 Methodology

This section explains the data analysis and the machine learning process to build the model for classifying the given samples as attack or normal record.

3.1 Machine Learning Based NIDS

One of the most common reasons that machine Learning based network intrusion detection techniques are challenging is because of it is an outlier detection technique [1], this challenge is even more difficult because of very few attack records in the training set, that have 9 different attack types, making it difficult to develop a profile of the attack behavior.

3.2 Data Visualization

To select classification techniques and to better analyze the data, a technique that the authors used is data visualization. Since the original data set is in high dimensional space, techniques were used to reduce the dimensions to visualize the attack and normal records. t-distributed Stochastic Neighbor Embedding (TSNE) was used to reduce the dimension of the data to two basic features, and subsequently sample points were plotted. TSNE converts similarities between data points to joint probabilities and it tries to minimize the Kullback-Leibler divergence between the joint probabilities. The visualized attacks and selected normal data are shown in Figure 1 and 2.

As is evident from the projected, reduced space, the data is not easily separable (at least in this approximation).

3.3 Feature Scaling

The given features and the extended behavioral features have different scales, this required to normalize all features on one scale, so that the different ranges and scales of features do not contribute to relative weights of those features. To normalize, the following method was used:

$$X^{'} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

3.4 Re-sampling

The training data set is higher number of normal network traffic in comparison to attack records, thus the data was re-sampled to build a robust classifier. The re-sampled data set had 40 normal records that were randomly chosen from the data set, and 41 fake records from the training data set.

3.5 Machine Learning Techniques

Following supervised and unsupervised machine learning techniques were used to classify given samples as "Normal" (represented as 0) or "Attack"(represented as 1):

1. Random Forest (Supervised)
2. KNN Classifier (Unsupervised)
3. Logistic Regression (Supervised)
4. KNN Regressor (Semi-Supervised)

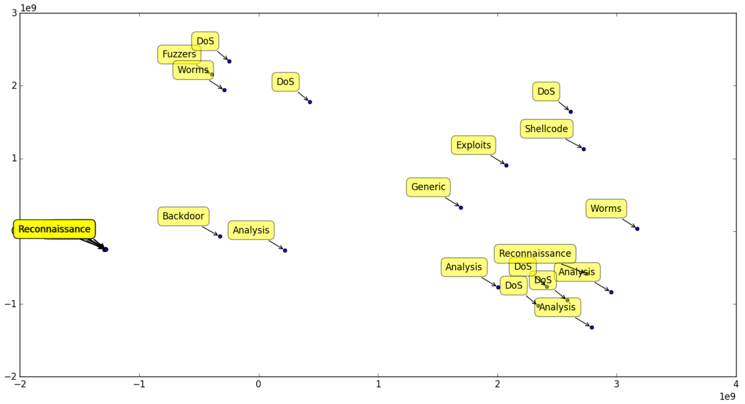

Fig. 1. Anomalies in reduced space

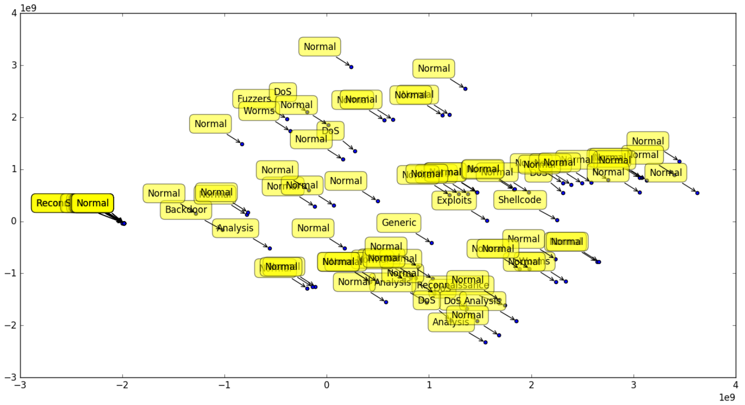

Fig. 2. Anomalies and Selected Data in reduced space

3.6 Generalizing - Developing a Robust Classifier

To build a robust classifier, that is relatively general and is not restricted only to the given training set following techniques were used:

Three-fold Cross Validation: Three-fold cross validation was used to overcome the problem of over-fitting, and to build a model to that will generalize to an independent dataset, Hawkins et. al. (see [2]).

4 Results

Results of the challenge are summarized in this section. We present results based upon training accuracies only. The results presented are achieved after feature scaling and random re-sampling and three-fold cross validation.

Table 2. Best Average Cross-Validation Accuracies

Technique	Training Accuracy	Testing Accuracy
Random Forest	91.347	82.063
KNN Classifier	87.224	77.319
KNN Regressor	88.892	74.669
Logistic Regression	92.119	71.958
Average	**97.948**	**76.503**

As you can see, there is quite a big gap between training and test accuracies. This can be explained due to the fact that we used a very small sample for our classifier, due to the under sampling and the fact that there are only about 50 anomalous entries. This would make our k-fold classification approach tend to overfit.

As we examined the data, we tried to project the anomalies to 2D space to visualize it. We can see the result in Fig. 1, where we only plot anomalies and Fig. 2 where we also plot it with some normal points. As it can be seen, there is no obvious separation in this space between anomalous and normal entries. However, we did observe some anomalous entries were groped together in terms of their attack categories. Therefore, we decided to use an semi-supervised clustering approach to try to cluster the anomalies together, and then extrapolate the testing set from that. We used a KNN regressors approach.

However, this approach did not fair any better than the supervised Random Forest approach which scored the best out of all classifiers.

References

1. P. Garca-Teodoroa, J. Daz-Verdejoa, G. Macia-Fernandeza, E. Vazquezb. Anomaly-based network intrusion detection: Techniques, systems and challenges aDepartment

of Signal Theory, Telematics and Communications Computer Science and Telecommunications Faculty, University of Granada, Granada, Spain (2008)

2. Hawkins D. , Basak S. , and Denise M. Assessing Model Fit by Cross-Validation J. Chem. Inf. Comput. Sci., 2003, 43 (2), pp 579586 (2003)