Machine Learning Algorithms for Intrusion Detection in Legacy and Modern Systems

Van N. Nguyen, MSOL Graduate Student, University of California, Riverside

ENGR 296 Final Project – Summer 2020

# Abstract

The KDD Cup 1999 Data is a network traffic dataset has been used in numerous academic publications in the past two decades to demonstrate machine learning techniques for network intrusion detection. The dataset has been popular among the academic community because of the ease of use and prelabeled samples. However, today the dataset is considered outdated and generally not accepted for academic journals. In this paper, we will revisit the KDD Cup 1999 Data looking at the performance of machine learning classification algorithms against that data and then compare the same machine learning classification algorithms against a modern network traffic dataset that is similar in features and samples. We will examine whether the KDD Cup 1999 Data is relevant today and whether a more modern dataset is fundamentally different from a machine learning standpoint. The modern data we will be using is CSE-CIC-IDS2018 Data, a labeled dataset of network intrusion by the Communications Security Establishment (CSE) & the Canadian Institute for Cybersecurity (CIC). We will compare the performance of standard machine learning algorithms at classifying normal network traffic and network intrusion traffic against the legacy KDD Cup 1999 Data and the CSE-CIC-IDS2018 Data.

# Background

Network Intrusion Detection is an important use-case for data science and machine learning. Cybersecurity firms are increasingly using machine learning to thwart cyber threats for their clients. Machine learning is well suited for network intrusion detection because of the “cat and mouse nature” of the problem. As firms adapt their security to defend against threats, the cyber threats evolve and exploit gaps in the defense. Machine learning offers the ability to continuously learn and adapt to new threats. The global cyber security market size was estimated at $156.45 billion in 2019[[1]](#endnote-1). The global cost of cybercrime is estimated in the trillions of US dollars[[2]](#endnote-2). The opportunities to leverage this technology are enormous.

The KDD Cup 1999 Data is a data set used for The Third International Knowledge Discovery and Data Mining Tools Competition at KDD-99 The Fifth International Conference on Knowledge Discovery and Data Mining. The competition task was to build a network intrusion detector, a predictive model capable of distinguishing between ``bad'' connections, called intrusions or attacks, and ``good'' normal connections. The data includes a wide variety of intrusions simulated in a military network environment.

The advantage of the KDD Cup 1999 Data is that samples are labeled allowing for supervised machine learning techniques. The labels represent the type of network intrusion attack. Non-attack samples are also labeled as “normal”.

Labels:

back,buffer\_overflow,ftp\_write,guess\_passwd,imap,ipsweep,land,loadmodule,multihop,neptune,nmap,normal,perl,phf,pod,portsweep,rootkit,satan,smurf,spy,teardrop,warezclient,warezmaster.

Features:

duration: continuous.

protocol\_type: symbolic.

service: symbolic.

flag: symbolic.

src\_bytes: continuous.

dst\_bytes: continuous.

land: symbolic.

wrong\_fragment: continuous.

urgent: continuous.

hot: continuous.

num\_failed\_logins: continuous.

logged\_in: symbolic.

num\_compromised: continuous.

root\_shell: continuous.

su\_attempted: continuous.

num\_root: continuous.

num\_file\_creations: continuous.

num\_shells: continuous.

num\_access\_files: continuous.

num\_outbound\_cmds: continuous.

is\_host\_login: symbolic.

is\_guest\_login: symbolic.

count: continuous.

srv\_count: continuous.

serror\_rate: continuous.

srv\_serror\_rate: continuous.

rerror\_rate: continuous.

srv\_rerror\_rate: continuous.

same\_srv\_rate: continuous.

diff\_srv\_rate: continuous.

srv\_diff\_host\_rate: continuous.

dst\_host\_count: continuous.

dst\_host\_srv\_count: continuous.

dst\_host\_same\_srv\_rate: continuous.

dst\_host\_diff\_srv\_rate: continuous.

dst\_host\_same\_src\_port\_rate: continuous.

dst\_host\_srv\_diff\_host\_rate: continuous.

dst\_host\_serror\_rate: continuous.

dst\_host\_srv\_serror\_rate: continuous.

dst\_host\_rerror\_rate: continuous.

dst\_host\_srv\_rerror\_rate: continuous.

# Experiment Setup

## Feature Reduction Analysis

Before applying We will attempt to reduce the number of features using Random Forest Algorithm.

### KDD Cup 1999 Data Feature Reduction Analysis

The KDD Cup 1999 Data contains 41 features. Since the time complexity of kNN algorithm is O(n2)[[3]](#endnote-3),

Feature Importance Score

count 0.179769

srv\_count 0.124680

same\_srvrate 0.080571

src\_bytes 0.074948

protocol\_type 0.073758

dst\_host\_same\_src\_port\_rate 0.071520

diff\_srv\_rate 0.061298

flag 0.056767

service 0.055318

dst\_host\_same\_srv\_rate 0.034246

dst\_host\_diff\_srv\_rate 0.033881

dst\_bytes 0.033329

dst\_host\_srv\_count 0.019166

dst\_host\_serror\_rate 0.019150

logged\_in 0.015758

serror\_rate 0.013443

dst\_host\_count 0.010805

dst\_host\_srv\_serror\_rate 0.009024

dst\_host\_srv\_diff\_host\_ate 0.006068

dst\_host\_rerror\_rate 0.003736

hot 0.003478

wrong\_fragment 0.003440

num\_compromised 0.003362

dst\_host\_srv\_rerror\_rate 0.003152

rerror\_rate 0.002592

srv\_serror\_rate 0.002250

srv\_diff\_host\_rate 0.002100

duration 0.001076

srv\_rerror\_rate 0.000713

is\_guest\_login 0.000359

num\_failed\_logins 0.000091

root\_shell 0.000039

num\_ile\_creations 0.000028

num\_root 0.000027

land 0.000023

num\_access\_files 0.000013

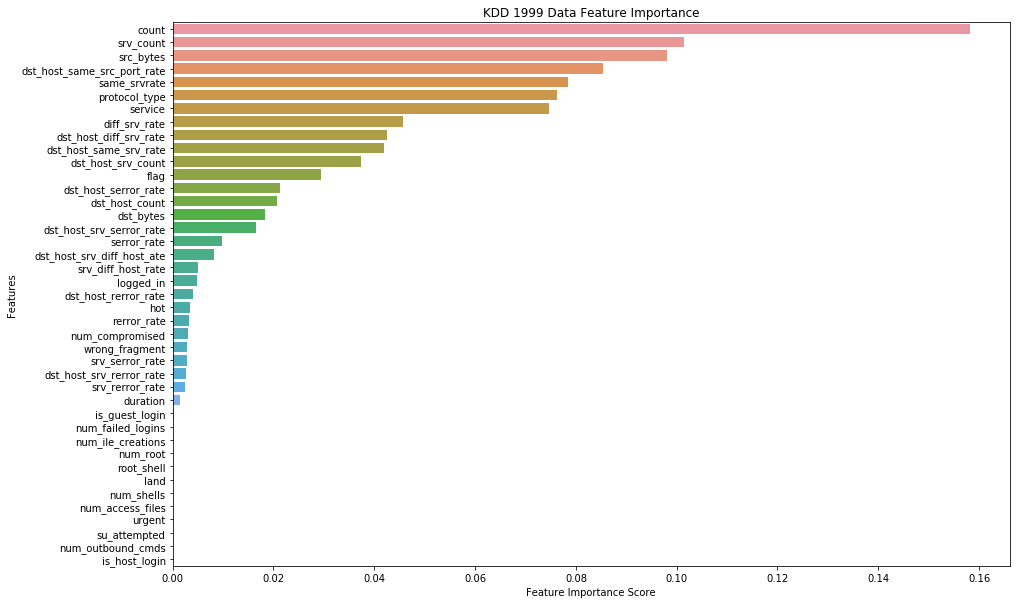
num\_shells 0.000011

urgent 0.000009

su\_attempted 0.000002

num\_outbound\_cmds 0.000000

is\_host\_login 0.000000



### CSE-CIC-IDS2018 Data Feature Reduction Analysis

The CSE-CIC-IDS2018 Data contains 80 features. We will remove the Timestamp feature because the time of the attack is not relevant to detecting future attacks. By leaving the Timestamp feature in the model is vulnerable to overfitting because within the training data the attacks may be time correlated. However, we know logically that the Timestamp feature is not relevant in detecting future attacks.

We perform a Random Forest

Feature Importance Score

Dst Port 0.100080

Fwd Pkts/s 0.056449

Flow IAT Max 0.056117

Flow Duration 0.053613

Flow IAT Min 0.053206

Flow IAT Mean 0.052637

Flow Pkts/s 0.051192

Bwd Pkts/s 0.051181

Flow Byts/s 0.041710

Init Fwd Win Byts 0.037802

Fwd IAT Max 0.024511

Fwd IAT Min 0.023884

Fwd Seg Size Min 0.023633

Fwd IAT Mean 0.023305

Flow IAT Std 0.021666

Fwd IAT Tot 0.020397

Fwd IAT Std 0.017259

Bwd IAT Min 0.013001

Bwd IAT Mean 0.012715

Bwd IAT Max 0.012361

Bwd IAT Std 0.012061

Pkt Len Mean 0.012033

Bwd IAT Tot 0.011217

Init Bwd Win Byts 0.010274

Pkt Size Avg 0.009991

Fwd Pkt Len Max 0.008927

Fwd Pkt Len Mean 0.008841

Fwd Header Len 0.008508

Pkt Len Max 0.008263

Pkt Len Var 0.007909

Fwd Seg Size Avg 0.007696

Subflow Fwd Byts 0.007648

Pkt Len Std 0.007605

TotLen Fwd Pkts 0.007345

PSH Flag Cnt 0.007330

Bwd Seg Size Avg 0.006940

Bwd Pkt Len Mean 0.006765

TotLen Bwd Pkts 0.006421

Subflow Bwd Byts 0.006414

Active Min 0.005612

Active Max 0.005263

Subflow Fwd Pkts 0.005230

Active Mean 0.005161

Tot Fwd Pkts 0.005098

Bwd Pkt Len Max 0.004822

Idle Min 0.004647

Idle Std 0.004320

Bwd Header Len 0.004129

Bwd Pkt Len Std 0.003915

Idle Mean 0.003903

Fwd Pkt Len Std 0.003697

Bwd Pkt Len Min 0.003425

Active Std 0.003294

ACK Flag Cnt 0.003017

Fwd Pkt Len Min 0.002982

Subflow Bwd Pkts 0.002981

Idle Max 0.002948

Tot Bwd Pkts 0.002925

Pkt Len Min 0.002917

URG Flag Cnt 0.001940

Down/Up Ratio 0.001858

Fwd Act Data Pkts 0.001264

RST Flag Cnt 0.001100

ECE Flag Cnt 0.001051

SYN Flag Cnt 0.000833

Fwd URG Flags 0.000653

Fwd PSH Flags 0.000633

CWE Flag Count 0.000632

Protocol 0.000541

FIN Flag Cnt 0.000307

Fwd Pkts/b Avg 0.000000

Fwd Blk Rate Avg 0.000000

Fwd Byts/b Avg 0.000000

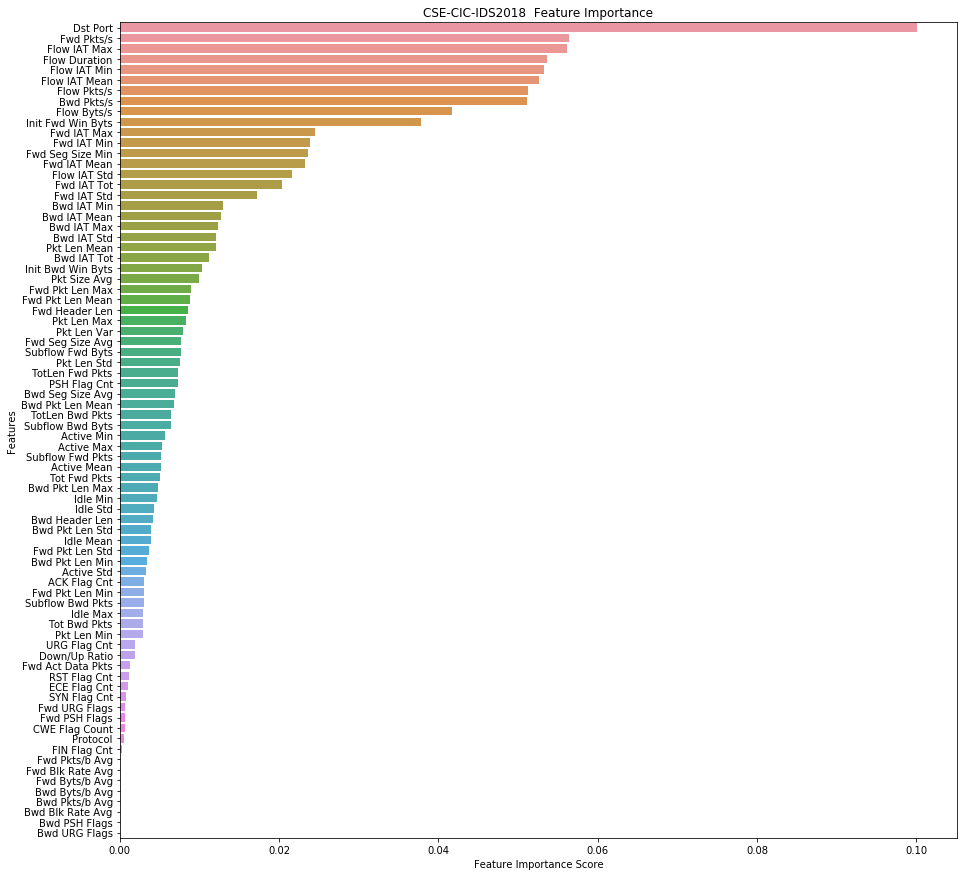
Bwd Byts/b Avg 0.000000

Bwd Pkts/b Avg 0.000000

Bwd Blk Rate Avg 0.000000

Bwd PSH Flags 0.000000

Bwd URG Flags 0.000000



## Model Selection

“Logistic regression is not able to handle a large number of categorical features/variables. It is vulnerable to overfitting. Also, can't solve the non-linear problem with the logistic regression that is why it requires a transformation of non-linear features. Logistic regression will not perform well with independent variables that are not correlated to the target variable and are very similar or correlated to each other.”[[4]](#endnote-4)

# Results

## Decision Tree Model

### KDD Cup 1999 Data

The following are the results of testing a Decision Tree Model for the KDD Cup 1999 Data with 80% training data and 20% test data.

precision recall f1-score support

back. 1.00 1.00 1.00 467

buffer\_overflow. 0.67 0.67 0.67 9

ftp\_write. 1.00 0.50 0.67 4

guess\_passwd. 1.00 0.83 0.91 12

imap. 0.75 0.75 0.75 4

ipsweep. 1.00 1.00 1.00 2443

land. 1.00 0.75 0.86 4

loadmodule. 0.33 0.50 0.40 2

multihop. 0.00 0.00 0.00 1

neptune. 1.00 1.00 1.00 214905

nmap. 0.99 0.99 0.99 452

normal. 1.00 1.00 1.00 194337

perl. 1.00 1.00 1.00 2

phf. 1.00 1.00 1.00 2

pod. 0.97 1.00 0.98 56

portsweep. 1.00 1.00 1.00 2015

rootkit. 0.00 0.00 0.00 0

satan. 1.00 1.00 1.00 3104

smurf. 1.00 1.00 1.00 561466

teardrop. 1.00 1.00 1.00 204

warezclient. 0.96 0.98 0.97 195

warezmaster. 0.75 1.00 0.86 3

accuracy 1.00 979687

macro avg 0.84 0.82 0.82 979687

weighted avg 1.00 1.00 1.00 979687

### CSE-CIC-IDS2018 Data

The following are the results of testing a Decision Tree Model for the CSE-CIC-IDS2018 Data with 80% training data and 20% test data.

precision recall f1-score support

Benign 0.79 0.81 0.80 47240

Infilteration 0.48 0.44 0.46 18397

accuracy 0.71 65637

macro avg 0.63 0.63 0.63 65637

weighted avg 0.70 0.71 0.70 65637

The following are the results of testing a Decision Tree Model for the CSE-CIC-IDS2018 Data with 80% training data and 20% test data.

precision recall f1-score support

Benign 1.00 1.00 1.00 197791

DoS attacks-GoldenEye 1.00 1.00 1.00 8096

DoS attacks-Slowloris 1.00 1.00 1.00 2223

accuracy 1.00 208110

macro avg 1.00 1.00 1.00 208110

weighted avg 1.00 1.00 1.00 208110

Combined Data

precision recall f1-score support

Benign 0.95 0.96 0.96 244569

DoS attacks-GoldenEye 1.00 1.00 1.00 8439

DoS attacks-Slowloris 1.00 1.00 1.00 2200

Infilteration 0.42 0.37 0.39 18538

accuracy 0.92 273746

macro avg 0.84 0.83 0.84 273746

weighted avg 0.92 0.92 0.92 273746

## K Nearest Neighbor Model

### KDD Cup 1999 Data

The following are the results of testing a k-Nearest Neighbor Model for the the KDD Cup 1999 Data with 80% training data and 20% test data.

precision recall f1-score support

back. 1.00 0.99 0.99 438

buffer\_overflow. 0.33 0.33 0.33 3

ftp\_write. 0.00 0.00 0.00 1

guess\_passwd. 1.00 1.00 1.00 8

imap. 0.00 0.00 0.00 1

ipsweep. 0.99 0.99 0.99 2534

land. 0.50 0.33 0.40 3

loadmodule. 0.00 0.00 0.00 3

neptune. 1.00 1.00 1.00 215040

nmap. 0.99 0.98 0.99 476

normal. 1.00 1.00 1.00 194639

perl. 0.50 1.00 0.67 1

pod. 1.00 1.00 1.00 56

portsweep. 1.00 1.00 1.00 2069

rootkit. 0.00 0.00 0.00 3

satan. 1.00 0.99 1.00 3188

smurf. 1.00 1.00 1.00 560820

teardrop. 1.00 1.00 1.00 198

warezclient. 0.92 0.94 0.93 203

warezmaster. 1.00 0.67 0.80 3

accuracy 1.00 979687

macro avg 0.71 0.71 0.70 979687

weighted avg 1.00 1.00 1.00 979687

### CSE-CIC-IDS2018 Data

The following are the results of testing a k-Nearest Neighbor Model for the CSE-CIC-IDS2018 Data with 80% training data and 20% test data.

precision recall f1-score support

Benign 0.79 0.90 0.84 47134

Infilteration 0.60 0.38 0.46 18503

accuracy 0.75 65637

macro avg 0.69 0.64 0.65 65637

weighted avg 0.73 0.75 0.73 65637

The following are the results of testing a k-Nearest Neighbor Model for the CSE-CIC-IDS2018 Data with 80% training data and 20% test data.

precision recall f1-score support

Benign 1.00 1.00 1.00 197559

DoS attacks-GoldenEye 1.00 1.00 1.00 8389

DoS attacks-Slowloris 1.00 1.00 1.00 2162

accuracy 1.00 208110

macro avg 1.00 1.00 1.00 208110

weighted avg 1.00 1.00 1.00 208110

Combined Data

precision recall f1-score support

Benign 0.95 0.99 0.97 244620

DoS attacks-GoldenEye 1.00 1.00 1.00 8302

DoS attacks-Slowloris 1.00 1.00 1.00 2189

Infilteration 0.68 0.32 0.44 18635

accuracy 0.94 273746

macro avg 0.91 0.83 0.85 273746

weighted avg 0.93 0.94 0.93 273746

# Discussion

# Summary

# References

## Academic Papers

1. Combining Feature Selection and Local Modelling in the KDD Cup 99 Dataset. Iago Porto-Diaz, David Martiınez-Rego, Amparo Alonso-Betanzos, and Oscar Fontenla-Romero. Department of Computer Science, University of A Coruna, Spain
2. Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99. H. G. Kayacik, A. N. Zincir-Heywood, M. Heywood. Published 2005.
3. Intrusion Detection System Classification Using Different Machine Learning Algorithms on KDD-99 and NSL-KDD Datasets - A Review Paper. R. Rama Devi, Munther Abualkibash. Published 2019. Computer Science International Journal of Computer Science and Information Technology.
4. Problems of KDD Cup 99 Dataset Existed and Data Preprocessing. Yan Wang\*, Kun Yang, Xiang Jing, Huang Long Jin. Published 2014. Applied Mechanics and Materials (Volume 667)
5. Mao, J. K., & Zhan, F. (2015). Study on Intrusion Detection System Based on Data Mining. Applied Mechanics and Materials, 713–715, 2499–2502. https://doi.org/10.4028/www.scientific.net/amm.713-715.2499
6. Luo, Y. J., Meng, J., & Zhao, H. Y. (2014). Research on Intrusion Detection Technology Based on Ad Hoc Network. Applied Mechanics and Materials, 687–691, 2622–2625. https://doi.org/10.4028/www.scientific.net/amm.687-691.2622
7. Satish Kumar, Sunanda, Sakshi Arora. A Statistical Analysis on KDD Cup’99 Dataset for the Network Intrusion Detection System. Part of the Lecture Notes in Networks and Systems book series (LNNS, volume 125).

## Non-Academic Articles

1. 3 New Techniques for Data-Dimensionality Reduction in Machine Learning. <https://thenewstack.io/3-new-techniques-for-data-dimensionality-reduction-in-machine-learning/>
2. Building an Intrusion Detection System using KDD Cup’99 Dataset. <https://medium.com/analytics-vidhya/building-an-intrusion-detection-model-using-kdd-cup99-dataset-fb4cba4189ed>

## Data Sets Used

A Realistic Cyber Defense Dataset (CSE-CIC-IDS2018) <https://registry.opendata.aws/cse-cic-ids2018/>

KDD Cup 1999 Data The UCI KDD Archive, Information and Computer Science, University of California, Irvine <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>

## Code Repository

<https://github.com/nnvan/CS296>

## Endnotes

1. <https://www.grandviewresearch.com/industry-analysis/cyber-security-market> [↑](#endnote-ref-1)
2. <https://cybersecurityventures.com/hackerpocalypse-cybercrime-report-2016> [↑](#endnote-ref-2)
3. The Analysis and Optimization of KNN Algorithm Space-Time Efficiency for Chinese Text Categorization Ying Cai and Xiaofei Wang Dept. of Computer Science and Technology, Beijing Information Science & Technology University Beijing, 100101, P. R. China [↑](#endnote-ref-3)
4. <https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python> [↑](#endnote-ref-4)