Wydział Elektryczny
POLITECHNIKA WARSZAWSKA

The influence of the dataset bias on the accuracy of the NIDS systems

Franciszek Pelc

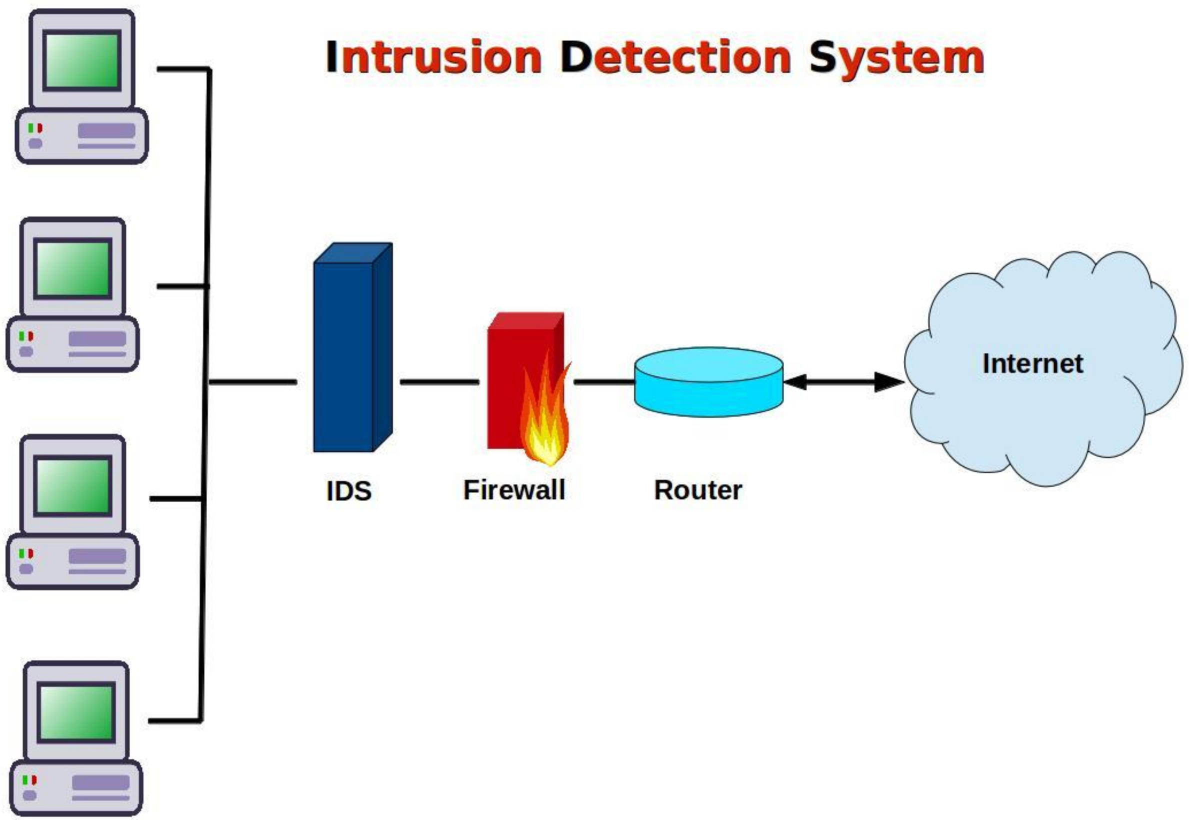
supervisor: dr hab. inż. Marcin Iwanowski

CITEE conference presentation

Politechnika Warszawska

Motivation

The goal of our research was to measure the impact of training datasets on accuracy of machine learning Network Intrusion Detection System (NIDS) models.



computersecuritypgp.blogspot.com

Datasets

- UNSW-NB15
- BoT-IoT
- ToN-IoT
- CIC-CSE-IDS2018

Klasa	BoT %	IDS2018 %	NB15 %	ToN %
Analysis	0	0	0.0962	0
Backdoor	0	0	0.0907	0.0992
Benign	0.3576	88.0482	96.0233	36.0053
Bot	0	0.7574	0	0
Brute Force	0	0.6562	0	0
DDoS	48.5438	7.3584	0	11.9608
DoS	44.1516	2.5617	0.2424	4.2065
Exploits	0	0	1.32	0
Fuzzers	0	0	0.9334	0
Generic	0	0	0.6928	0
Infilteration	0	0.6158	0	0
Reconnaissance	6.9406	0	0.5346	0
Shellcode	0	0	0.0596	0
Theft	0.0063	0	0	0
Worms	0	0	0.0069	0
injection	0	0.0023	0	4.0404
mitm	0	0	0	0.0456
password	0	0	0	6.8081
ransomware	0	0	0	0.0202
scanning	0	0	0	22.3218
xss	0	0	0	14.4919

Percentage of datasets consisting of particular class.

Features

- Originaly there have been 43 derived from flow and other data.
- Some features have been discarded
- Ultimately 32 features were considered for each record.

Feature
L7_PROTO
PROTOCOL
TCP_FLAGS
SERVER_TCP_FLAGS
ICMP_TYPE
MIN_TTL
DNS_QUERY_TYPE
DNS_TTL_ANSWER
FTP_COMMAND_RET_CODE
IN_BYTES
IN_PKTS
OUT_BYTES
OUT_PKTS
FLOW_DURATION_MILLISECONDS
DURATION_IN
DURATION_OUT
LONGEST_FLOW_PKT
SHORTEST_FLOW_PKT
MIN_IP_PKT_LEN
SRC_TO_DST_SECOND_BYTES
DST_TO_SRC_SECOND_BYTES
RETRANSMITTED_IN_BYTES
RETRANSMITTED_IN_PKTS
RETRANSMITTED_OUT_BYTES
SRC_TO_DST_AVG_THROUGHPUT
DST_TO_SRC_AVG_THROUGHPUT
NUM_PKTS_UP_TO_128_BYTES
NUM_PKTS_128_TO_256_BYTES
NUM_PKTS_256_TO_512_BYTES
NUM_PKTS_512_TO_1024_BYTES
TCP_WIN_MAX_IN
TCP_WIN_MAX_OUT

Classifiers

Dataset / Hyperparameter	criterion
Default	gini
BoT	log_loss
IDS2018	entropy
NB15	entropy
ToN	log_loss

TABLE III
OPTIMIZED PARAMETERS FOR DECISION TREE CLASSIFIER

- Decision Tree Classifier
- Random Forest Classifier
- Extra Trees Classifier

Test have been conducted both on default

as well as optimized hiperparameters

Dataset / Hyperparameter	criterion
Default	gini
BoT	entropy
IDS2018	log_loss
NB15	log_loss
ToN	entropy

TABLE IV
OPTIMIZED PARAMETERS FOR RANDOM FOREST CLASSIFIER

Dataset / Hyperparameter	criterion	max_features	bootstrap
Default	gini	sqrt	False
BoT	entropy	None	True
IDS2018	log_loss	log2	True
NB15	log_loss	None	True
ToN	entropy	None	True

TABLE V
OPTIMIZED PARAMETERS FOR EXTRA TREES CLASSIFIER

Experiments

1. Training each classifier on every dataset
2. Tests for each model on every dataset
3. Accuracy measured separately for each class
4. Accuracy = (TP + TN)/(All records)

Benign
DDoS
DoS
Reconnaissance
Backdoor
All classes

Results

Trained on dataset BoT

Accuracy for class [%] \ tested on dataset	**IDS2018**	**IDS2018 optimized**	**NB15**	**NB15 optimized**	**ToN**	**ToN optimized**	**BoT**	**BoT optimized**
Benign	55.4610	57.2694	28.6710	24.7427	55.5062	58.145	_99.9855_	_99.984_
DDoS	92.5474	92.64	_99.9802*_	_99.998*_	87.9844	87.9911	99.874	99.87
DoS	98.4526	97.5279	_98.7986_	_99.179_	90.6257	89.4078	99.6704	99.6709
Reconnaissance	_52.5663*_	_49.7831*_	**28.624**	**24.2688**	_14.9694*_	_20.5188*_	99.774	99.7825
Backdoor	_100.0*_	_100.0*_	_99.9093_	_99.9093_	_99.9008_	_99.9008_	_100.0*_	_100.0*_
All classes	41.182	46.7863	26.7328	22.3717	3.3359	3.8008	99.6505	99.6525

Trained on dataset IDS2018

Accuracy for class [%] \ tested on dataset	**BoT**	**BoT optimized**	**NB15**	**NB15 optimized**	**ToN**	**ToN optimized**	**IDS2018**	**IDS2018 optimized**
Benign	_51.807_	_7.0542_	90.5869	93.7148	44.2014	41.8339	55.423	99.5255
DDoS	**51.4062**	**51.4368**	_99.9993*_	_99.9990*_	87.8835	88.0384	92.5904	99.9755
DoS	**55.8207**	**57.2276**	_97.7408_	_99.7512_	76.7222	83.2694	98.4299	99.9985
Reconnaissance	93.0594	93.0594	_99.4654_	_99.4654_	_100.0*_	_100.0*_	_52.535*_	_100.0*_
Backdoor	_100.0*_	_100.0*_	_99.9093_	_99.9093_	_99.9008_	_99.9008_	_100.0*_	_100.0*_
All classes	0.2438	1.6812	94.0112	93.1099	26.2241	25.7695	41.2095	99.5015

Trained on dataset NB15

Accuracy for class [%] \ tested on dataset	**BoT**	**BoT optimized**	**IDS2018**	**IDS2018 optimized**	**ToN**	**ToN optimized**	**NB15**	**NB15 optimized**
Benign	_41.6776_	_2.75089_	61.4127	52.1712	48.3047	53.6949	28.773	99.6695
DDoS	**51.4561**	**51.4561**	92.6416	92.6416	88.0391	88.0391	_99.9805*_	_100.0*_
DoS	**55.6830**	**55.8262**	66.0998	90.2995	91.1191	86.9492	_98.812_	_99.655_
Reconnaissance	84.8808	93.0594	_99.5719*_	_99.9984*_	_99.9956*_	_99.9944*_	_28.7_	_99.818_
Backdoor	_99.9994*_	_99.9986*_	_99.991*_	_99.9864*_	_99.8944_	_99.8803_	_99.911_	_99.842_
All classes	0.3282	0.3308	45.1172	41.5032	33.5194	33.2048	26.8510	98.8225

Trained on dataset ToN

Accuracy for class [%] \ tested on dataset	**BoT**	**BoT optimized**	**IDS2018**	**IDS2018 optimized**	**NB15**	**NB15 optimized**	**ToN**	**ToN optimized**
Benign	_53.3776_	_17.221_	58.9632	77.2498	58.062	75.1247	55.5845	99.2205
DDoS	**51.413**	**51.4047**	94.5719	98.2498	_95.7979*_	_88.7531*_	87.89	99.681
DoS	**55.8401**	**55.846**	95.5346	97.2593	_99.1569_	_99.7511_	90.728	99.181
Reconnaissance	93.0594	93.0594	_99.4654*_	_99.4654*_	_99.4654_	_99.4654_	_14.915*_	_100.0*_
Backdoor	_99.9975*_	_99.9995*_	_99.8862*_	_99.9058*_	_99.909_	_99.9093_	_99.9035_	_99.9995_
All classes	0.3215	0.3236	57.5044	76.4519	69.878	74.4936	3.3655	97.3485

Results for Decision Tree Classifier

Results

1. Results in tests may vary greatly between testing on dataset model was trained on and testing on other datasets.
2. Optimizing classifiers improves performance on datasets training was perfomed but may decrease for others.
3. The greatest variance in results was found in Benign, DDoS and DoS classes

Conclusions

1. There is a need to carefully select the training data for IDS models
2. The set-up specific properties as well as great diversity of traffic present obstacle in training effective models.
3. For IDS model there is a risk of selection bias, sampling bias, exlcusion bias as well as reporting bias.

Wydział Elektryczny
POLITECHNIKA WARSZAWSKA

Thank you for your attention

Franciszek Pelc

Politechnika Warszawska