

## **Feature Selection in the Corrected KDD -dataset**

ZARGARI, Shahrzad

Available from Sheffield Hallam University Research Archive (SHURA) at:

<http://shura.shu.ac.uk/17048/>

---

This document is the author deposited version. You are advised to consult the publisher's version if you wish to cite from it.

### **Published version**

ZARGARI, Shahrzad (2017). Feature Selection in the Corrected KDD -dataset. In: International Conference on Big Data in Cyber Security 2017, Cyber Academy, Edinburgh, 10 May 2017. (Unpublished)

---

### **Repository use policy**

Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. Users may download and/or print one copy of any article(s) in SHURA to facilitate their private study or for non-commercial research. You may not engage in further distribution of the material or use it for any profit-making activities or any commercial gain.

## Feature Selection in the Corrected KDD-dataset

Shahrzad Zargari

Computing Department, Sheffield Hallam University



# Contents



Introduction

Objective

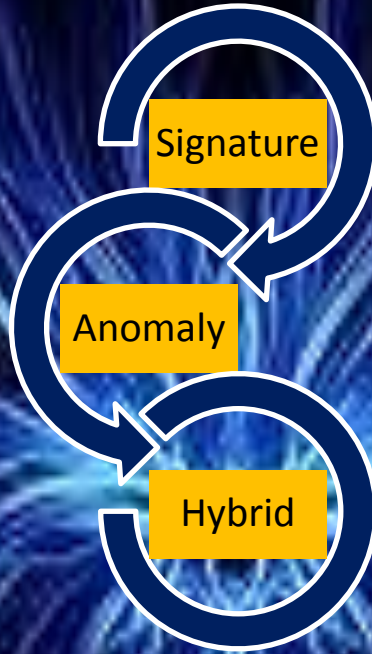
Methodology

Experimental Work

Conclusions

# Introduction

Intrusion detection systems



Anomaly intrusion detection deals with detecting of unknown attacks in the network traffic, therefore, they are difficult to identify without human intervention. IT administrators struggle to keep up with Intrusion Detection System (IDS) alerts, and often manually examine system logs to discover potential attacks.

Final Goal



Automation of Intrusion detection by using data mining and statistical techniques

# Objective

To propose a subset of features that can produce high intrusion detection rates while keeping the false positives at a minimum level. Therefore this will tackle the curse of dimensionality (e.g. reducing the computational complexity, time and power consumption)

# Challenges

Challenges

It is difficult to find published data for analysis

It is difficult to determine the normal traffic

the concept of normal traffic varies within different network

The KDD CUP 1999<sup>1</sup> is the first published dataset to be used in intrusion detection which has been used widely by researchers despite of the reported criticisms (McHugh, 2000) due to the lack of data

1) <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>



# The KDD-CUP 1999 datasets

The KDD CUP 1999 dataset is a version of the dataset produced by the DARPA (1998) Intrusion Detection Evaluation Program which included nine weeks of raw TCP dump data for a local-area network (LAN) simulating a typical U.S. Air Force LAN. The LAN was operated as if it were a true Air Force environment, but peppered it with multiple attacks.

The KDD-CUP  
1999

The full data: `Kddcup.data.gz`

A 10% subset: `kddcup.data_10_percent.gz`

The test data: `kddcup.testdata.unlabeled.gz`

Test data with corrected labels: `correcte.gz`

This study used the corrected test data for the data mining

# The KDD-CUP 1999 Structure

DOS: denial-of-service, e.g. syn flood

Probing: surveillance and other probing, e.g.. Port scanning

R2L: Unauthorized access from a remote machine, e.g. guessing password

U2R: Unauthorized access to local superuser (root) privileges, e.g., various "buffer overflow" attacks

The KDD-Cup 1999 dataset

24 attacks types

The test KDD-Cup 1999 dataset

37 attacks types

41 features

The distribution of the attacks in the KDD-Cup 1999 dataset is different from the test KDD-Cup 1999 dataset



# The Features Proposal

Amiri 2011

Olusola  
2011

Tang 2010

Chebrolu  
2005

Keyacik  
2005

## Proposed Features

- 3) Service
- 5) Source bytes
- 6) Destination bytes
- 39) Dst host rerror rate



# The experimental Work

4 Samples from the  
Corrected KDD-CUP  
1999 dataset



WEKA (V.3.7.4)  
Data mining  
software  
(Random Forest  
algorithm)



Including all features

Proposed features by  
this study  
(3,5,6,& 39)

CfsSubsetEval +  
GreedyStepwise  
(2,3,5,& 6)

InfoGainVal + Ranker  
(5,3,23,& 24)



Features suggested by this study (3,5,6,& 39) have higher intrusion detection rates with minimum false positives

# Weka V.3.7.4: Data Mining software in Java



Weka is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

# A Typical Output of Weka

The screenshot shows the Weka Explorer interface. The 'Classifier' tab is active, displaying a 'RandomForest -I 10 -K 0 -S 1' model. The 'Test options' section shows 'Cross-validation' selected with 10 folds and 66% split. The 'Classifier output' section displays the following summary statistics:

Correctly Classified Instances	90351	94.3446 %
Incorrectly Classified Instances	5416	5.6554 %
Kappa statistic	0.9234	
Mean absolute error	0.0034	
Root mean squared error	0.0417	
Relative absolute error	9.0131 %	
Root relative squared error	30.2045 %	
Coverage of cases (0.95 level)	99.6178 %	
Mean rel. region size (0.95 level)	3.021 %	
Total Number of Instances	95767	

Below the summary statistics, a 'Detailed Accuracy By Class' table is shown:

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.998	0	0.991	0.998	0.994	0.999	apache2.
0.999	0	0.999	0.999	0.999	0.999	back.
0.6	0	0.5	0.6	0.545	0.9	buffer_overflow.
0	0	0	0	0	0.5	ftp_write.
1	0	1	1	1	1	guess_passwd.
0.086	0	0.692	0.086	0.153	0.978	httptunnel.
0	0	0	0	0	0.5	imap.
1	0	0.995	1	0.998	1	ipsweep.

The 'Result list' on the left shows the selected model: '21:28:47 - trees.RandomForest'.



# The experimental Work (2)

Including 10 features

4 Samples from the Corrected KDD-CUP 1999 dataset



WEKA (V.3.7.4)  
Data mining software  
(Random Forest algorithm)



Proposed features by this study  
(3,5,6,& 39)+  
(4,14,16,27,28,& 37)

CfsSubsetEval + GreedyStepwise  
(2,3,5 & 6)+  
(8,23,30,34,36,& 4)

InfoGainVal + Ranker  
(5,3,23 & 24)+  
(33,35,2,36,34,& 6)

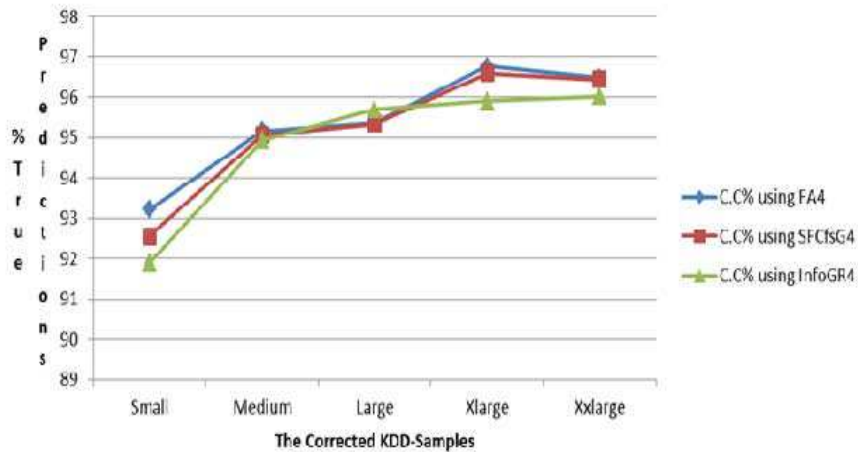
Feature set suggested by InfoGainVal+Ranker has higher intrusion detection rates however, comparing the results of applying only 4 features and 10 features, indicates that the detection rates improve slightly so it is a matter of trade off between increasing dimensionality or detection rate



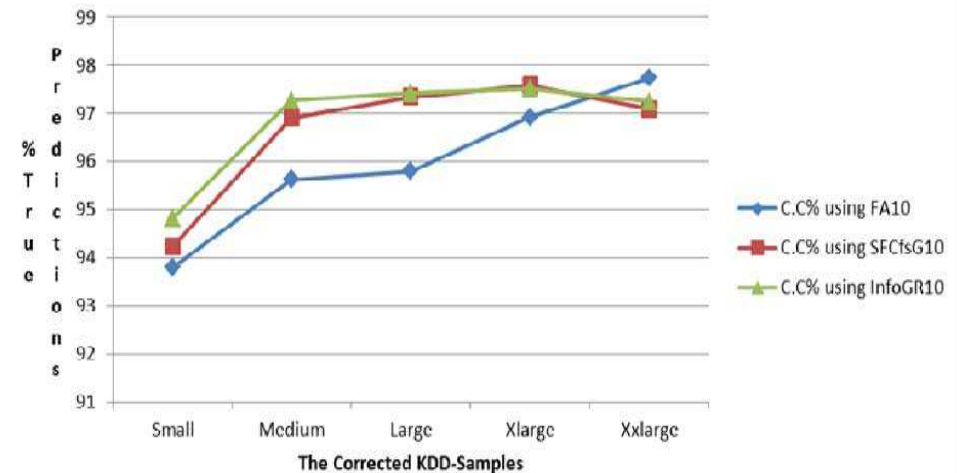


# The result of data mining using different feature subsets

## The results of data mining using three different subsets of features (4)



## The results of data mining using three different subsets of features (10)

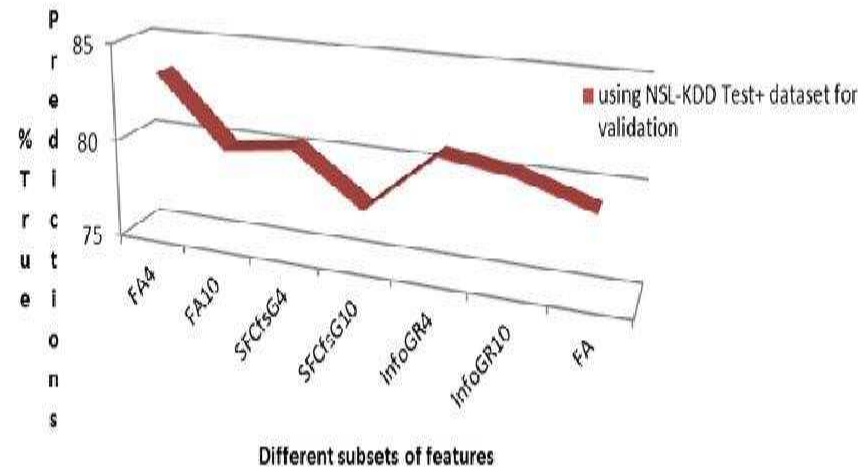


# NSL-KDD<sup>1</sup> Anomaly Dataset

The same experimental work was carried out on NSL-KDD anomaly dataset

The results showed that the proposed features produce higher detection rates than the other two methods of data mining.

The results of data mining (using different subsets of features)



# Conclusions

The statistical analysis of the Corrected KDD-CUP 1999 indicated that feature selection can reduce the high dimensions (curse of dimensionality) of the dataset and computational time while it does not have significant effect on intrusion detection rate.

The proposed subset of features (3,5,6,& 39) can be used in data mining tasks which performed better intrusion detections than the other subsets of features suggested by (CfsSubsetEval + GreedyStepwise) and (InfoGainVal + Ranker).

The subset of 10 features produced by InfoGainVal + Ranker algorithm performed better than the other subsets however, it is a matter of trade off (adding more dimensions) in order to improve the detection rate slightly.

The statistical analysis on NSL-KDD dataset confirmed the above results.

For future work, finding the optimum subset of features to be used in intrusion detection