Feature Selection in the Corrected KDD -dataset

ZARGARI, Shahrzad <http://orcid.org/0000-0001-6511-7646>

Sheffield Hallam University

Feature Selection in the Corrected KDD-dataset

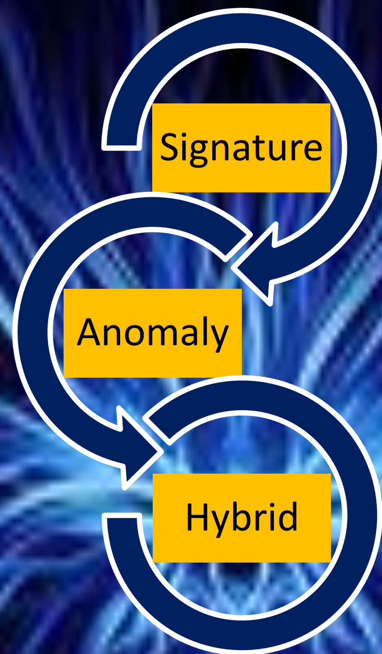
Shahrzad Zargari

Computing Department, Sheffield Hallam University

Contents

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[1] 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 KDD-database Structure

Features

D207295 | SF

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP
1	durati	type	service	flag	src-byt																																					traffic type
2	0	tcp	http	S0																																						apache2.
3	0	tcp	http	S0																																						apache2.
4	0	tcp	http	S0																																						apache2.
5	0	tcp	http	S0																																						apache2.
6	0	tcp	http	S0																																						apache2.
7	0	tcp	http	S0																																						apache2.
8	0	tcp	http	S0																																						apache2.
9	0	tcp	http	S0																																						apache2.
10	0	tcp	http	S0																																						apache2.
11	0	tcp	http	S0																																						apache2.
12	0	tcp	http	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	90	90	1	1	0	0	1	0	0	255	255	1	0	0	0	0.49	0.49	0.49	0	apache2.
13	0	tcp	http	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	91	91	1	1	0	0	1	0	0	255	255	1	0	0	0	0.49	0.49	0.49	0	apache2.
14	0	tcp	http	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	92	92	1	1	0	0	1	0	0	255	255	1	0	0	0	0.49	0.49	0.49	0	apache2.
15	0	tcp	http	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	93	93	1	1	0	0	1	0	0	255	255	1	0	0	0	0.5	0.5	0.48	0	apache2.
16	0	tcp	http	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	94	94	1	1	0	0	1	0	0	255	255	1	0	0	0	0.5	0.5	0.48	0	apache2.
17	0	tcp	http	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	95	95	1	1	0	0	1	0	0	255	255	1	0	0	0	0.51	0.51	0.47	0	apache2.
18	0	tcp	http	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	96	96	1	1	0	0	1	0	0	255	255	1	0	0	0	0.51	0.51	0.47	0	apache2.
19	0	tcp	http	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	97	97	1	1	0	0	1	0	0	255	255	1	0	0	0	0.51	0.51	0.47	0	apache2.
20	0	tcp	http	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	98	98	1	1	0	0	1	0	0	255	255	1	0	0	0	0.52	0.52	0.46	0	apache2.
21	0	tcp	http	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99	99	1	1	0	0	1	0	0	255	255	1	0	0	0	0.52	0.52	0.45	0	apache2.
22	0	tcp	http	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	100	1	1	0	0	1	0	0	255	255	1	0	0	0	0.53	0.53	0.45	0	apache2.
23	0	tcp	http	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	101	101	1	1	0	0	1	0	0	255	255	1	0	0	0	0.53	0.53	0.45	0	apache2.
24	0	tcp	http	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	102	102	1	1	0	0	1	0	0	255	255	1	0	0	0	0.53	0.53	0.44	0	apache2.
25	0	tcp	http	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	103	103	1	1	0	0	1	0	0	255	255	1	0	0	0	0.54	0.54	0.44	0	apache2.

H5 | 0

	A	B	C	D	E	F	G	H	I
1	duration	type	service	flag	src-byte	dst-byte	land	wrong-frg	urgent
2	0	tcp	http	S0	0	0	0	0	0
3	0	tcp	http	S0	0	0	0	0	0

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

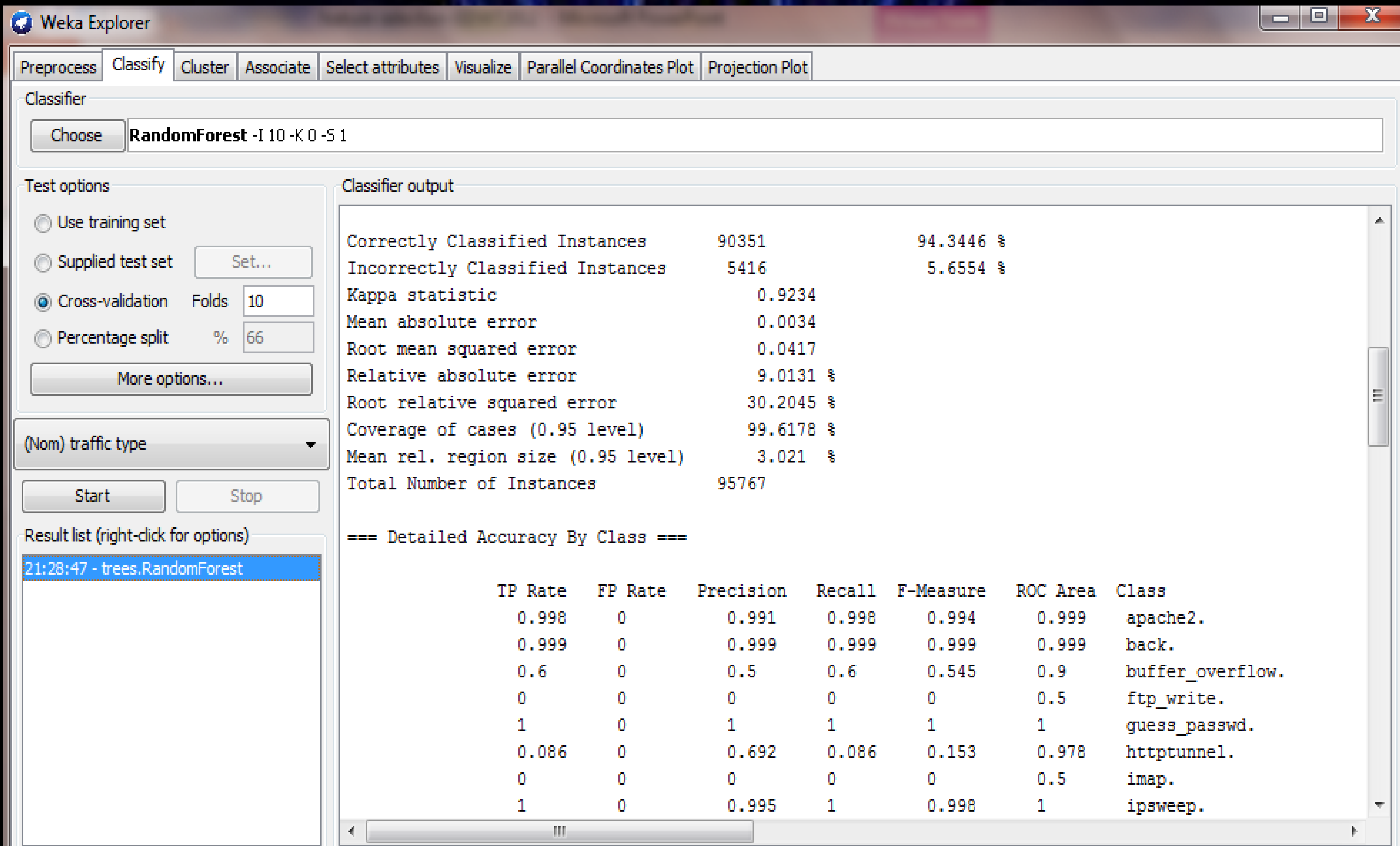
Weka V.3.7.4: Data Mining software in Java

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.

http://www.cs.waikato.ac.nz/~ml/weka//

A Typical Output of Weka



The experimental Work (2)

Including 10 features

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

4 Samples from the Corrected KDD-CUP 1999 dataset

+

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

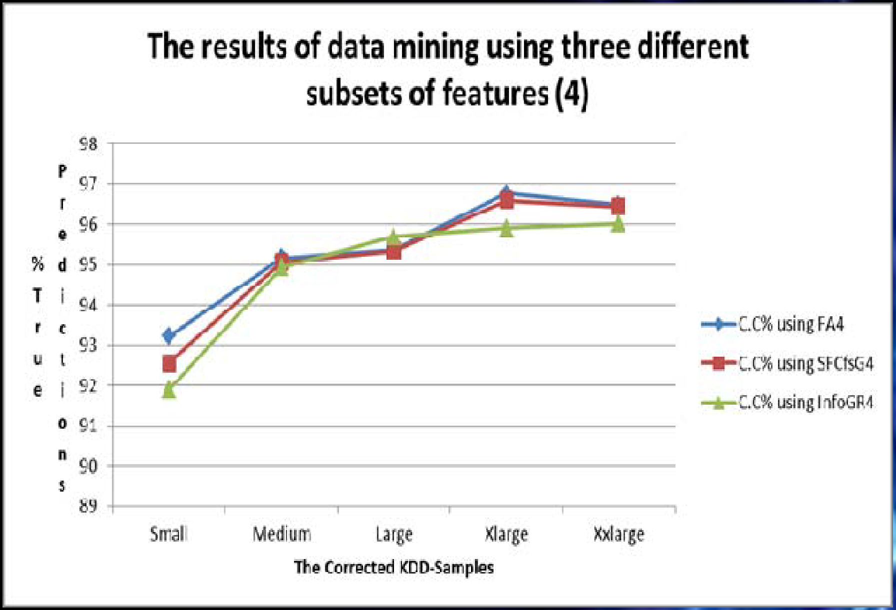
+

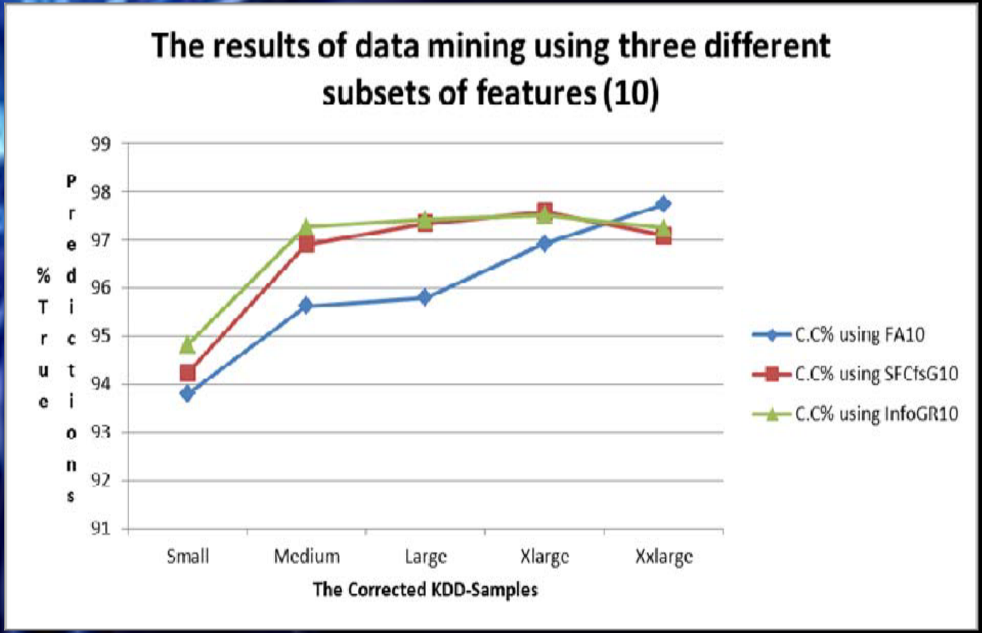
CfsSubsetEval + GreedyStepwise (2,3,5 & 6)+ (8,23,30,34,36,& 4)

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

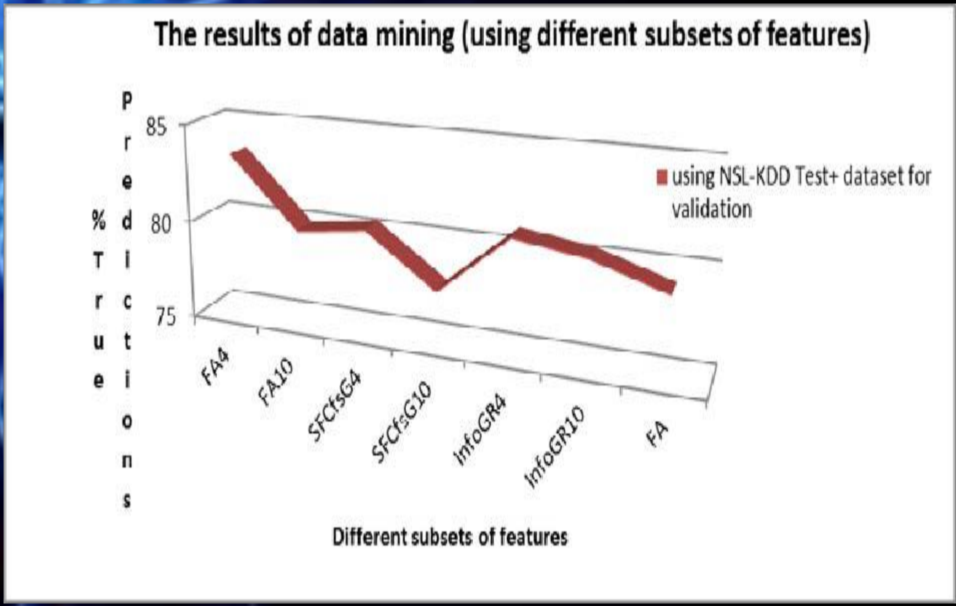




NSL-KDD[1] Anomaly Dataset

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.



1)Tavallaee et al, 2009, <http://iscx.ca/NSL-KDD/>

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