Sheffield Hallam University

Feature Selection in the Corrected KDD -dataset

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

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

https://shura.shu.ac.uk/17048/

This document is the Presentation

Citation:

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 (Unpublished) [Conference or Workshop Item]

Copyright and re-use policy

See http://shura.shu.ac.uk/information.html



Sheffield Hallam University

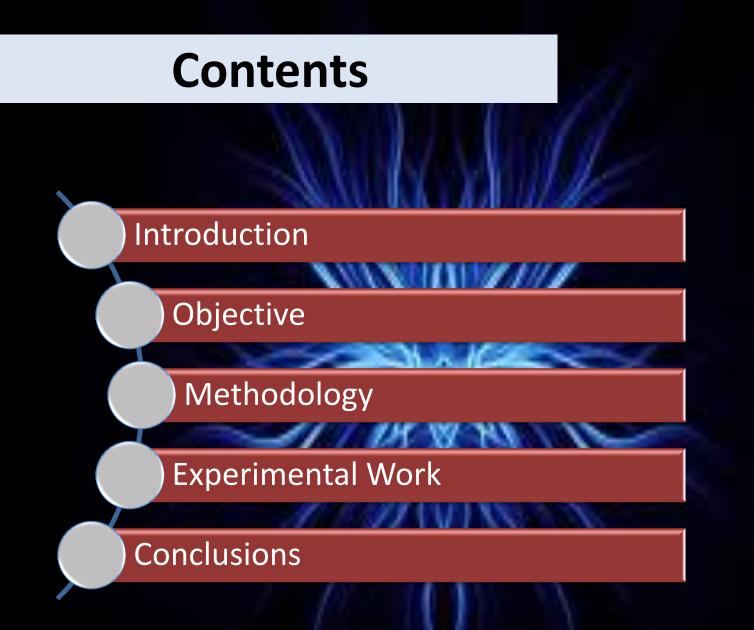
Feature Selection in the Corrected KDD-dataset

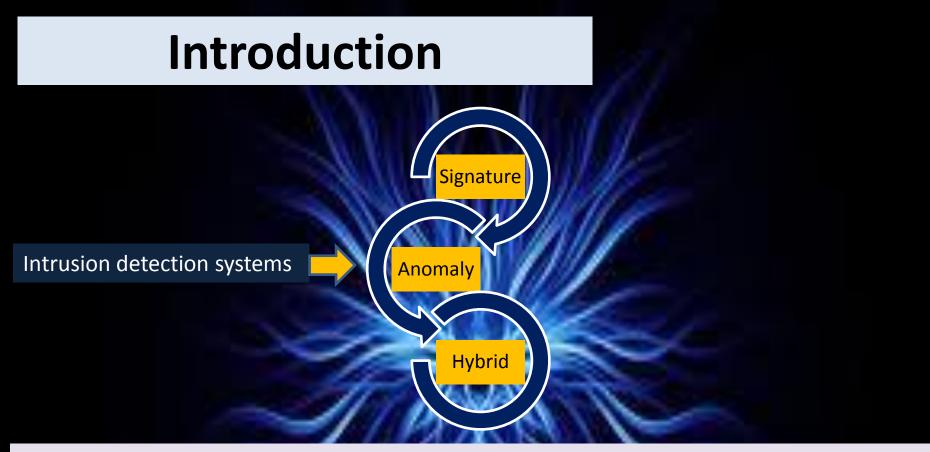




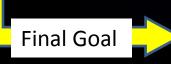
Shahrzad Zargari

Computing Department, Sheffield Hallam University





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.



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 1999¹ 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

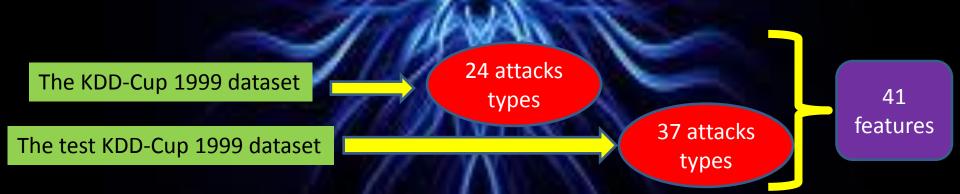
The KDD-CUP 1999 Structure

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

The Features Proposal



Proposed Features

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

The KDD-database Structure

Features

	D207295	•	0	f_{x}	SF								Σ													
	A B	С	D	E	F	GHI	J K L N	/INO	P	Q R S T U V W	X	Y	ZA	A	AB A	AC	AD	AE AF	AG A	AH AI	AJ	AK	AL	AM	AN A	C AP
1 0	duratictype	service	flag	src-byt		H5		- (1	<i>f</i> * 0																traffic type
2	0 tcp	http	S0	_		115		• (<u> </u>	J.4 U																apache2.
3	0 tcp	http	SO				\	D		C				E I	-			F		G					1	apache2.
4	0 tcp	http	S0			A		В		L C		D		E		F			G	Н			apache2.			
5	0 tcp	http	S0		4																			~		apache2.
6	0 tcp	http	S0		1	dura	tion	typ	е	service	TI:	ag	sr	C-k	byte	e	dst	t-byte	e li	and	W	ror	ng	trg	urgen	t apache2.
7	0 tcp	http	S0		_		_				_				•	_		•	_	_			-	-	_	apache2.
8	0 tcp	http	S0		2		0	tcp		http	S)				0		(0	0				0	. (apache2.
9	0 tcp	http	S0							-						-										apache2.
10	0 tcp	http	S0		3		0	tcp		http	S	ר				0		(0	0				0) apache2.
11	0 tcp	http	S0		-											-			-	-				-		apache2.
12	0 tcp	http	S0	0)	0000	000	000	0	00000090	90	1	1	0	0	1	0	0 255 2	255	1 (0 0	0	0.4	9 0.49	0.49	0 apache2.
13	0 tcp	http	S0	0)	0000	000	000	0	00000091	91	1	1	0	0	1	0	0 255 2	255	1 (0 0	0	0.4	9 0.49	0.49	0 apache2.
14	0 tcp	http	S0	0)	0000	000	000	0	00000092	92	1	1	0	0	1	0	0 255 2	255	1 (0 0	0	0.4	9 0.49	0.49	0 apache2.
15		http	S0	0)	0000	000	000	0	00000093	93	1	1	0	0	1	0	0 255 2		1 (0 0	0	0 0.	5 0.5	0.48	0 apache2.
16		http	S0	0)	0000	000	000	0	00000094		1	1	0	0	1	0	0 255 2		1 (0 0	0	0 0.	5 0.5		0 apache2.
17		http	S0	0)	0000		000	0	00000095		1	1	0	0	1	0	0 255 2		1 (0 0		_	1 0.51		0 apache2.
18		http	S0	0)	0000		000	0	00000096		1	1	0	0	1	0	0 255 2		1 (0 0			1 0.51		0 apache2.
19		http	S0	0)	0000	000	000	0	00000097		1	1	0	0	1	0	0 255 2		1 (0 0			1 0.51		0 apache2.
20		http	S0	0)	0000		000	0	0000098		1	1	0	0	1	0	0 255 2		1 (0 0			2 0.52		0 apache2.
21		http	S0	0		0000		000	0	0000099		1	1	0	0	1	0	0 255 2		1 (0 0		_	2 0.52		0 apache2.
22	0 tcp	http	S0	0)	0000	000	000	0	0 0 0 0 0 0 100	100	1	1	0	0	1	0	0 255 2	255	1 (0 0	0	0.5	3 0.53	0.45	0 apache2.
23		http	S0	0)	0000	000	000	0	0 0 0 0 0 0 101	101	1	1	0	0	1	0	0 255 2		1 (0 0	0	0.5	3 0.53	0.45	0 apache2.
24	0 tcp	http	S0	0)	0000	000	000	0	0 0 0 0 0 0 102		1	1	0	0	1	0	0 255 2		1 (0 0	0	0.5	3 0.53	0.44	0 apache2.
25	0 tcp	http	S0	0)	0000	000	000	0	000000103	103	1	1	0	0	1	0	0 255 2	255	1 (0 0	(0.5	4 0.54	0.44	0 apache2.

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

🕑 Weka Explorer	name and distances in the surface of						- 0 X							
Preprocess Classify Cluster Associate S	elect attributes Visualize Parallel Coordinate	es Plot Projection Plo	t											
Classifier														
Choose RandomForest -I 10 -K 0 -5	pose RandomForest -I 10 -K 0 -S 1													
Test options	Classifier output													
Use training set							A							
	Correctly Classified Instances	90351		94.3446 %										
Supplied test set Set	Incorrectly Classified Instance	es 5416		5.6554 %										
Oross-validation Folds 10	Kappa statistic													
Percentage split % 66	Mean absolute error 0.0034													
Percentage spin 78 00	Root mean squared error	0.043												
More options	Relative absolute error		9.0131 %											
	Root relative squared error	30.204	45 %				=							
(Nom) traffic type 🗸 🗸	Coverage of cases (0.95 level)													
(tony durine type	Mean rel. region size (0.95 le	vel) 3.021	3.021 %											
Start Stop	Total Number of Instances	95767												
Result list (right-click for options)	=== Detailed Accuracy By Class ===													
21:28:47 - trees.RandomForest														
1	TP Rate FP Rat	te 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.	-							
	•						F.							

The experimental Work (2)

Proposed features by this study **Including 10** (3,5,6,& 39)+ features (4,14,16,27,28,& 37) WEKA (V.3.7.4) CfsSubsetEval + 4 Samples from the Data mining GreedyStepwise **Corrected KDD-CUP** software (2,3,5 & 6)+ 1999 dataset (Random Forest (8,23,30,34,36,& 4) algorithm) 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

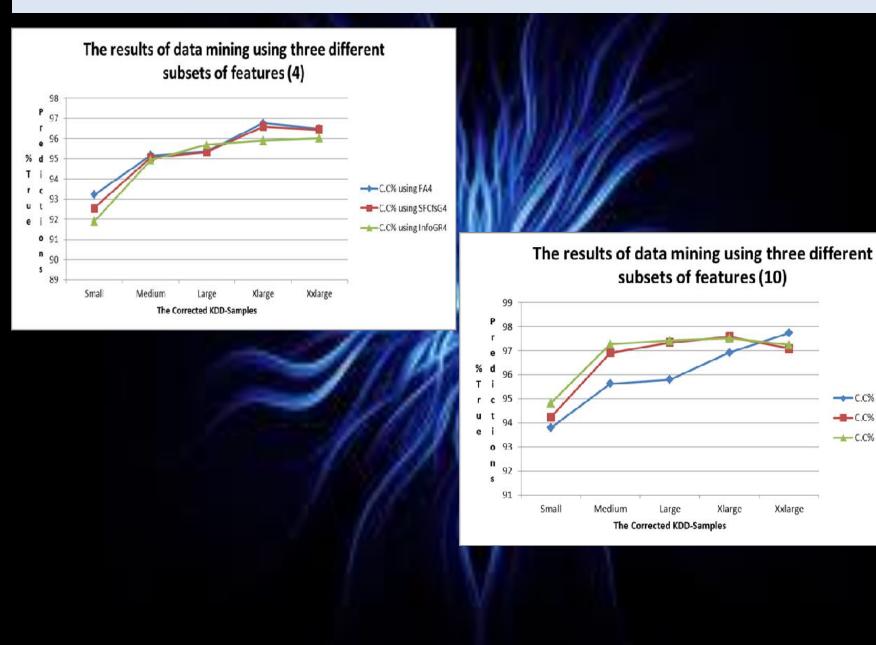
The result of data mining using different feature subsets

C.C% using FA10

C.C% using SFCfsG10

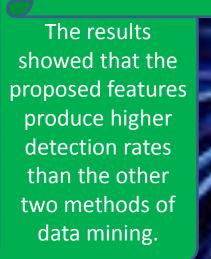
Xlarge

Xxlarge

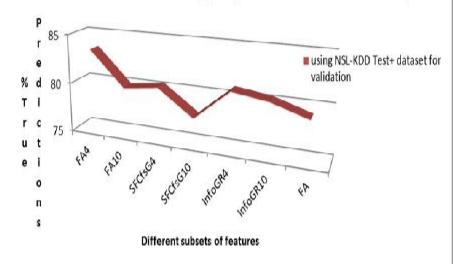


NSL-KDD¹ Anomaly Dataset

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



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



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