Discriminant Analysis based Feature Selection in KDD Intrusion Dataset

Dr.S.Siva Sathya
Department of Computer Science
Pondicherry University,
Puducherry,India.

Dr. R.Geetha Ramani
Department of Computer Science and Engineering
Rajalakshmi Engineering College
(Affiliated to Anna University, Chennai)
Thandalam, Chennai
Tamilnadu, India.

K.Sivaselvi
Department of Computer Science
Pondicherry University,
Puducherry,India.

,

ABSTRACT
Intrusion detection system (IDS) plays a major role in providing network security by analyzing the network traffic log and classifying the records as attack or normal behavior. Generally, as each log record is characterized by a large set of features, an Intrusion Detection System consumes large computational power and time for the classification process. Hence, feature reduction becomes mandatory before attack classification for any IDS. Discriminant analysis is a technique which can be used for selecting important features in large set of features. In this paper, important features of KDD Cup '99 attack dataset are obtained using discriminant analysis method and used for classification of attacks. The results of discriminant analysis show that classification is done with minimum error rate with the reduced feature set.

General Terms
Data Mining, Intrusion Detection.

Keywords
Discriminant analysis, KDD Cup '99 attack dataset, classification, features relevance, minimum error rate, SPSS.

1. INTRODUCTION
With the increased growth of networked systems and applications, the demand for network security is high. Though, there are various ways to provide security such as cryptography, anti-virus, malwares, spywares, etc., it is not possible to provide complete secure systems. So there should be a second line of defense as an Intrusion Detection Systems to detect attacks[5,11]. To identify intruders, differentiating normal user behavior and attack behavior is essential. Efficient IDS can be developed by defining a proper rule set for classifying the network traffic log records into normal or attack patterns. The DARPA KDD Cup '99 dataset has been used by most of the researchers as a test bed for the development of efficient IDS and IPS. Since it is very large with each record composed of 41 features, creation of a rule set is very tedious. More over all features will not be relevant or fully contribute in identifying an attack. So the number of features has to be reduced in order to develop efficient rule set for classification. There are several methods for feature relevance analysis. One such technique is discriminant analysis which is a statistical method for obtaining a reduced feature set.

This paper deals with this statistical method for analyzing the voluminous KDD Cup dataset[6,12]. Even though there are many methods, they have some loopholes like two-way cluster analysis and k-means cluster classify the normal and attack records with high error rate[3]. The hierarchical cluster cannot work with voluminous data. Since the KDD dataset is very large, we cannot use the hierarchical cluster analysis. Hence from the experimental analysis it is found that discriminant analysis classifies the dataset with minimum error rate.

The organization of the rest of the paper is as follows: Section 2 describes the related work in this area. In Section 3 detailed description of Intrusion Detection System, KDD Cup '99 dataset and the features are given. Section 4 presents about Discriminant analysis, Section 5 explains the experimental analysis and results and Section 6 gives the conclusion.

2. RELATED WORK
Several researchers have applied different techniques for feature relevance analysis and some have concentrated on discriminant analysis to identify relevant features in different applications. In 2002, Midori Asak a et al[14] explained in detail about performance of discriminant analysis in intrusion detections and evaluated its classification function. They have obtained information from system logs which is used for their experimental purpose. In 2005, Aarabi et al[15] presented new feature selection algorithm based on discriminant and redundancy analysis to identify feature subsets. They evaluated the performance of this method by extracting features from seizure and non-seizure segments in newborn. In 2008, Mohamed Elgendi et al[16] given an algorithm for intra-class classification which includes an analysis of the R-R time series. Then feature has been extracted and using them, a criterion was created for classification. In 2008, Kun-Ming Yu et al[17] used logistic regression and protocol type for important feature selection to design efficient intrusion detection system. In 2010, Zhiyuan Tan et al[18] used linear discriminant analysis and difference distance map to identify significant features to reduce

the heavy computational cost of an anomaly IDS. In 2011, Fatemeh Amiri et al[19] proposed two feature selection algorithms and compared those with mutual information based feature selection algorithm.

3. INTRUSION DETECTION SYSTEM

An intrusion detection system (IDS) can be a device or software application that monitors the network or system activities for malicious attacks or policy violations and reports it to a Management Station[4]. IDS are considered to provide dynamic defense mechanisms to various network security threats. IDS can be divided into two types as network based and host based. Network intrusion detection system (NIDS) detects intrusions by continuously monitoring network traffic by connecting to network hub or switch which is configured for port mirroring, or network tap. NIDS uses sensors to capture all network traffic and to monitor individual packets to identify whether it is normal or attack. An example of a NIDS is Snort [13]. Host-based intrusion detection system (HIDS) uses agent as a sensor on a host that identifies intrusions by analyzing system calls, application logs, file-system modifications (binaries, password files, etc.) and other host activities and state. OSSEC is an example for Host based intrusion detection system [13].

Passive systems are called as Intrusion Detection Systems and reactive systems are known as Intrusion Prevention Systems. IDS detect malicious activity, from a set of log records and alert the user. IPS auto-responds to the suspicious activity by resetting the connection or by reprogramming the firewall to block network traffic from the suspected malicious source. Based on the methodology adopted to identify intrusions, IDS could be classified as: anomaly detection and misuse detection. In anomaly detection, normal user behavior is developed. The anomaly detector monitors incoming packets and check for normal behavior. If it is deviating then it is considered as abnormal or attack. In misuse detection, abnormal behavior is modeled[5]. The misuse detector monitors network segments and check for abnormality. Misuse detector has higher accuracy when compared to anomaly detector because modeling normal behavior is difficult. Commercial IDS are mostly based on misuse detection[5]. The log records usually contain a large number of features which make the task of an Intrusion Detection System very difficult. Hence important features can be derived using some feature reduction algorithm and used for classification of data as normal or attack.

3.1 KDD CUP '99 Intrusion Detection Dataset

The KDD Cup '99 attack dataset is a public repository to promote the research works in the field of intrusion detection[8]. The details of KDD dataset is given in the subsequent section. The dataset has 41 features and 3,11,029 records. The KDD Cup '99 intrusion detection datasets are based on the 1998 DARPA initiative, which provides designers of intrusion detection systems(IDS) with a benchmark on which to evaluate different methodologies[9]. To do so, a simulation is made of a fictitious military network consisting of three 'target' machines running various operating systems and services. Additional three machines are then used to spoof different IP addresses to generate traffic. Finally, there is a sniffer that records all network traffic using the TCP dump format. The total simulated period is seven weeks. Attacks fall into one of four categories: User to Root; Remote to Local; Denial of Service; and Probe [1].

Denial of Service (dos): Excessive consumption of resources that denies legitimate requests from legal users on the system[1].

Remote to Local (r2l): Attacker having no account gains a legal user account on the victim machine by sending packets over the networks[1].

User to Root (u2r): Attacker tries to access restricted privileges of the machine[1].

Probe: Attacks that can automatically scan a network of computers to gather information or find known vulnerabilities[1].

3.2 KDD CUP '99 Features

In 1998, MIT Lincoln Lab developed standard set of data for intrusion detection which can be used by researchers. This dataset is obtained by setting up military environment and it was used in International Knowledge Discovery and Data Mining Tools Contest. TCP dump data are obtained and it was processed as connections. Specifically, "a connection is a sequence of TCP packets starting and ending at some well defined times, between which data flows from a source IP address to a target IP address under some well defined protocol"[6]. The description of the various features is shown in the Table 1.

Table 1. KDD'99 Feature Description

Feature No.	Feature Name	Description
1	Count	number of connections to the same host as the current connection in the past two seconds
2	destination bytes	Bytes sent from destination to source
3	diff srv rate	% of connections to different services
4	dst host count	count of connections having the same destination host
5	dst host diff srv rate	% of different services on the current host
6	dst host rerror rate	% of connections to the current host that have an RST error
7	dst host same src port rate	% of connections to the current host having the same src port
8	dst host same srv rate	% of connections having the same destination host and using the same service
9	dst host serror rate	% of connections to the current host that have an S0 error
10	dst host srv count	count of connections having the same destination host and using the same service

11	dst host srv diff host rate	% of connections to the same service coming from different hosts
12	dst host srv rerror rate	% of connections to the current host and specified service that have an RST error
13	dst host srv serror rate	% of connections to the current host and specified service that have an S0 error
14	Duration	Duration of the connection.
15	Flag	Status flag of the connection
16	Hot	number of "hot" indicators
17	is guest login	1 if the login is a "guest" login; 0 Otherwise
18	is host login	1 if the login belongs to the "host"
19	Land	1 if connection is from/to the samehost/port; 0 otherwise
20	logged in	1 if successfully logged in; 0 otherwise
21	num access files	number of operations on access control files
22	num compromised	number of "compromised" conditions
23	num failed logins	number of failed logins
24	num file creations	number of file creation operations
25	num outbound cmds	number of outbound commands in an ftp session
26	num root	number of "root" accesses
27	num shells	number of shell prompts
28	protocol type	Connection protocol (e.g. tcp, udp).
29	rerror rate	% of connections that have “REJ" Errors
30	root shell	1 if root shell is obtained; 0 otherwise
31	same srv rate	% of connections to the same service
32	serror rate	% of connections that have “SYN" Errors
33	Service	Destination service (e.g. telnet, ftp)
34	src bytes	Bytes sent from source todestination
35	srv count	number of connections to the same service as the current connection in the past two seconds
36	srv diff host rate	% of connections to different hosts
37	srv rerror rate	% of connections that have “REJ" errors
38	srv serror rate	% of connections that have “SYN" Errors
39	su attempted	1 if "su root" command attempted; 0 otherwise
40	Urgent	number of urgent packets
41	Wrongfragment	number of wrong fragments

Features are grouped into four categories[3]:

Basic Features: These features are directly obtained from packet headers. Basic features are the first six features provided in feature description section[3].

***Content Features*:** Domain knowledge is applied to assess data portion of the TCP packets. Features like number of failed login attempts are content features[3].

***Time-based Traffic Features*:** These features are designed to capture properties that mature over a 2 second temporal window. One example of such a feature would be the number of connections to the same host over the 2 second interval[3]l.

***Host-based Traffic Features*:** Some probing attacks scan the hosts (or ports) using a much larger time interval than two seconds, for example once per minute. Therefore, connection records were also sorted by destination host, and features were constructed using a window of 100 connections to the same host instead of a time window[3].

The KDD cup ‘99 intrusion detection benchmark consists of three components, which are detailed in Table 2. In the International Knowledge Discovery and Data Mining Tools Competition, only “10% KDD” dataset is employed for the purpose of training[10]. It is a concise from of “Whole KDD”. This dataset has only 22 attack types and they are mostly of denial of service category. They have more number of examples for attack than normal. Whereas “Corrected KDD” dataset provides a dataset with different statistical distributions compared to “10% KDD” or “Whole KDD”. It contains 37 type of attacks. Table 2 gives number of records in each attack category.

Table 2. Basic characteristics of the KDD 99 intrusion detection datasets in terms of number of samples

Dataset	DoS	Probe	u2r	r2l	Normal
“10%KDD”	391458	4107	52	1126	97277
“CorrectedKDD”	229853	4166	70	16347	60593
“WholeKDD”	3883370	41102	52	1126	972780

In this paper Corrected KDD is used for the experiments. There are 37 types of attacks in the dataset with varying percentage of different attacks which is shown in Table 3.

Table 3. Attack Frequency

Sl. No.	Attack Name	Count	Percentage
1	apache2	794	.3
2	back	1098	.4
3	buffer_overflow	22	.0
4	ftp_write	3	.0
5	guess_passwd	4367	1.4
6	httptunnel	158	.1
7	imap	1	.0
8	ipsweep	306	.1
9	land	9	.0
10	loadmodule	2	.0
11	mailbomb	5000	1.6
12	mscan	1053	.3
13	multihop	18	.0
14	named	17	.0
15	neptune	58001	18.6
16	nmap	84	.0
	normal	60593	19.5
17	perl	2	.0
18	phf	2	.0
19	pod	87	.0
20	portsweep	354	.1
21	processtable	759	.2
22	ps	16	.0
23	rootkit	13	.0
24	saint	736	.2
25	satan	1633	.5
26	sendmail	17	.0
27	smurf	164091	52.8
28	snmpgetattack	7741	2.5
29	snmpguess	2406	.8
30	sqlattack	2	.0
31	teardrop	12	.0
32	udpstorm	2	.0
33	warezmaster	1602	.5
34	worm	2	.0
35	xlock	9	.0
36	xsnoop	4	.0
37	xterm	13	.0
	Total	**311029**	**100.0**

4. DISCRIMINANT ANALYSIS

Discriminant analysis is a statistical technique used to build a predictive model of group membership based on observed characteristics of each case[6,11]. The purpose of Discriminant Analysis is to classify objects (graduate, undergraduate, etc.,) based on attribute set which describe the objects (e.g. age, gpa, etc.,)[12]. The first purpose is feature selection and the second purpose is classification. In discriminant analysis, the dependent variable (Y) is the group and the independent variables (X) are the object features that might describe the group. The dependent variable is discrete variable and the independent variables can be discrete or continuous.

Linear discriminant model can be used for groups that are linearly separable(i.e. the groups can be separated by a linear combination of features that describe the objects). If there are only two features, the separators between objects group will become lines. If the features are three, the separator is a plane and if the number of features (i.e. independent variables) is more than 3, the separators become a hyper-plane.

The functions are generated from a sample of cases for which group membership is known; the functions can then be applied to new cases with measurements for the predictor variables of unknown group membership[2,6,11].

5. FEATURE RELEVANCE USING DISCRMINANT ANALYSIS FOR KDD CUP '99 ATTACK DATASET

This section gives the experimental results for the classification of attacks in the KDD cup '99 Attack dataset. SPSS tool is used to perform the Discriminant analysis on the training dataset to obtain the important features for the classification process[7]. Since the classification problem is visualized as a two class categorization problem, the KDD cup dataset is fragmented into 37 subsets. The KDD dataset is divided into many subsets. Each subset contains records of normal and specific attack. Each subset is analyzed with the discriminant analysis for identifying the important features for specific attack.

Each subset comprises of the data records of specific attack type and normal. The result includes the classification results obtained for all 37 subsets.

This work comprises of 2 phases:

1. Feature reduction using discriminant analysis
2. Classification with reduced feature set

In this experiment important features which discriminates dataset labels (i.e. either as normal or any type of attack) are identified using discriminant analysis. This analysis result gives a set of features for each subset which is sufficient to group the attack and normal records. These features are considered as relevant features for each attack. For example, 'back' attack can

be identified with features 'hot', 'num file creations' and 'is guest login' instead of all 31 features. Similarly for all attacks related features are given. The number of features required to identify an attack varies with each attack.

The features reduction algorithm is as follows:

Begin

{

For each attack j

do

{

initialize R = { } where R is a feature subset.

initialize F = { set of all 41 features}

Do a discriminant analysis on the KDD dataset with F using Mahalanobis distance in stepwise statistics

// The output of the above step returns the discriminant value of the features along with their ranking and classification and misclassification rate//

Set C=classification rate and M=misclassification rate

for i= 1 to 41

Select highest ranked feature F_k

$R = \{R \cup F_k\}$

$F = \{F - F_k\}$

Do a discriminant analysis on the KDD dataset with R using Mahalanobis distance in stepwise statistics

If(current classification rate>=C AND current misclassification

Rate<=M)

Return R ={selected features for attack j}

}

}

End

The final output of this method provides important features for identifying every attack. The output of the above algorithm that gives the relevant features for all the 37 attack are shown in the table 4.

Table 4. Reduced feature set for each attack after discriminant analysis

Attack Names	Relevant Features
apache2	5,6,12,15,29,32,37,38
back	16,24,17
buffer_overflow	22,26,27,30
ftp_write	40
guess_passwd	23,28,33
httptunnel	12,27,29
imap	6,12
ipsweep	11
land	13,38
loadmodule	30
mailbomb	6,8,12
mscan	3,5,6,9,12,13,15,29,31,32
multihop	10,17,18,21,23,27
named	14,18
neptune	3,5,6,12,13,15,31,32
nmap	13
normal	8,9,12,13,17,18,19,28,29,40,41
perl	11,30
phf	12,15,30
pod	20,28,41
portsweep	6,12,15,29,37
processtable	9,10,12,13,14,15,20,28,33,38
ps	18,27
rootkit	18,27,40
saint	11,12,15,31
satan	5,15,31
sendmail	23,30
smurf	7,10,20,28
snmpgetattack	20,28,33,35,36
snmpguess	5,10,20,33,35,36
sqlattack	30
teardrop	41
udpstorm	9
warezmaster	7,10,17,34
worm	5,8,17,20,28,36
xlock	23,34
xsnoop	18
xterm	18,27,40

The KDD Cup '99 dataset already has a field called 'class', which specify the actual group. The analysis will classify them into groups based on the important features obtained by discriminant analysis. Classification and misclassification are based on actual and predicted group membership. In the first

subset, the classification rate and the misclassification rate against apache Vs normal are 99.7% and 0.3% respectively. The classification rate and the misclassification rate against normal Vs apache are 99.9% and 0.1% respectively. Similarly we can infer classification results for other subsets using Table 5.

Table 5. Classification Results of normal Vs Specific attack

Subsets	**Classification (%)**	**Misclassification (%)**
Apache2	99.7	0.3
Normal	99.9	.1
Back	99.4	.6
Normal	99.9	.1
Buffer_overflow	68.2	31.8
Normal	100	0
Ftp_write	33.3	66.7
Normal	100	0
Guess_passwd	100	0
Normal	94.4	5.6
Httptunnel	96.8	3.2
Normal	99.8	.2
Imap	100	0
Normal	99.9	.1
Ipsweep	97.1	2.9
Normal	99.9	0.1
Land	100	0
Normal	100	0
Loadmodule	100	0
Normal	100	0
Mailbomb	100	0
Normal	99.9	0.1
Mscan	95.7	4.3
Normal	99.8	.2
Multihop	22.2	77.8
Normal	100	0
Named	17.6	82.4
Normal	100	0
Neptune	100	0
Normal	100	0
Nmap	100	0
Normal	100	0
Perl	100	0
Normal	99.9	.1
Phf	100	0
Normal	99.9	.1
Pod	100	0
Normal	99.4	.6
Portsweep	100	0
Normal	99.8	.2
Processtable	97.2	2.8
Normal	99.9	.1
Ps	31.3	68.8
Normal	100	0
Rootkit	23.1	76.9
Normal	100	0
Saint	81.9	18.1
Normal	99.9	.1
Satan	99.8	.2
Normal	99.8	.2
Sendmail	35.3	64.7
Normal	100	0
Smurf	100	0
Normal	99.4	.6
Snmpgetattack	100	0
Normal	80.3	19.7
Snmpguess	100	0
Normal	80.2	19.8
Sqlattack	100	0
Normal	100	0
Teardrop	66.7	33.3
Normal	99.9	.1
Udpstorm	50	50
Normal	99.9	.1
Warezmaster	95.2	4.8
Normal	99.1	.9
Worm	100	0
Normal	98.4	1.6
Xlock	33.3	66.7
Normal	100	0

Xsnoop	50	50
Normal	100	0
Xterm	76.9	23.1
Normal	100	0

6. CONCLUSION

This paper has taken up the KDD'99 intrusion dataset to extract the most relevant feature subset for identifying a network traffic log record as normal or attack. For the feature relevance analysis, Discriminant analysis has been used along with a greedy selection method for selecting the features in the subset, based on their discriminant value. The relevant feature set is used for the classification of the entire dataset with normal and attack record. It is found that this analysis gives good classification rate and minimum error rate when compared to the classification done using the full feature set, thereby reducing the burden of the IDS in working with a large feature set.

The Discriminant analysis visualizes the problem as two-class categorization. Future research work can be done using an appropriate analysis method, to view the problem as a multi-class categorization.

7. ACKNOWLEDGEMENT

This work is a part of the AICTE funded project titled 'Bio-inspired Intrusion Response System through feature relevance Analysis on Attack Classification', Under the Research Promotion Scheme (RPS), Ref. No: 8023/BOR/RID/RPS-59/2009-10.

8. REFERENCES

[1] J. H. Güneş Kayacýk, A. Nur Zincir-Heywood, Malcolm I. Heywood. Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets**.**

[2] Jianhua Sun, Hai Jin, Hao Chen, Zongfen Han, and Deqing Zou. 2003. A Data Mining Based Intrusion Detection Model. Lecture Notes in Computer Science. Intelligent Data Engineering and Automated Learning. Springer publications. Volume 2690, 677-684.

[3] Khaled Labib**.** 2004. Computer Security and Intrusion Detection**.** ACM Crossroads. Volume 11(1): 2.

[4] Scarfone, Karen; Mell, Peter. 2007. Guide to Intrusion Detection and Prevention Systems (IDPS). Computer Security Resource Center (National Institute of Standards and Technology).

[5] Theuns Verwoerd, Ray Hunt**.** 2002**.** Intrusion detection techniques and approaches. Computer Communications. Volume 25, 1356-1365.

[6] Seema Jaggi and P.K.Batra, "SPSS: An Overview".

[7] SPSS Inc., "SPSS 13.0 Base User's Guide".

[8] Knowledge discovery in databases DARPA archive and Task Description. http://www.kdd.ics.uci.edu/databases/kddcup99/task.html.

[9] The 1998 intrusion detection off-line evaluation plan, MIT Lincoln Lab, Information Systems Technology Group.http://www.11.mit.edu/IST/ideval/docs/1998/id98-eval-11.txt.

[10] S. Hettich, S.D. Bay. 1999. The UCI KDD Archive. Irvine, CA: University of California, Department of Information and Computer Science. http://kdd.ics.uci.edu.

[11] David W. Stockburger, Multivariate Statistics: Concepts, Models, and Applications. http://www.psychstat.missouristate.edu/multibook/mlt03.htm.

[12] G. David Garson, Discriminant Function Analysis. 2008. http://www2.chass.ncsu.edu/garson/pa765/discrim.htm.

[13] Intrusion Detection System. .http://www.webopedia.com/TERM/I/intrusion_detection_system.html.

[14] Midori Asak a, Takefumi Onabura, T adashi Inoue, Shigeki Goto. 2002. Remote Attack Detection Method in IDA: MLSI-Based Intrusion Detection using Discriminant Analysis. Proceedings of the 2002 Symposium on Applications and the Internet (SAINT.02), IEEE.

[15] A. Aarabi, F. Wallois, R. Grebe. 2005. Feature Selection Based on Discriminant and Redundancy Analysis Applied to Seizure Detection in Newborn. Proceedings of the 2nd IEEE EMBS. Conference on neural Engineering, Arlington, Virginia.

[16] Mohamed Elgendi, Mirjam Jonkman, Friso De Boer. 2008. PREMATURE ATRIAL COMPLEXES DETECTION USING THE FISHER LINEAR DISCRIMINANT. Proceedings of the 7th IEEE International Conference on Cognitive Informatics (ICCI'08).

[17] Kun-Ming Yu and Ming-Feng Wu and Wai-Tak Wong. 2008. Protocol-Based Classification for Intrusion Detection. WSEAS TRANSACTIONS on COMPUTER RESEARCH. Volume 3(3).

[18] Zhiyuan Tan, Aruna Jamdagni, Xiangjian He, Priyadarsi Nanda. 2010. Network Intrusion Detection Based on LDA for Payload Feature Selection. IEEE Globecom 2010 Workshop on web and pervasive security.

[19] Fatemeh Amiri, MohammadMahdi Rezaei Yousefi , Caro Lucas , Azadeh Shakery NasserYazdani. 2011. Mutual information-based feature selection for intrusion detection systems. Journal of Network and Computer Applications. Volume 34, 1184–1199.