



Inas Ali  
Soukaena Hassan

# A Model To Detetct DOS Using Data Mining Classification Algorithms



**LAMBERT**  
Academic Publishing

Inas Ali  
Soukaena Hassan

## A Model To Detetct DOS Using Data Mining Classification Algorithms



LAP LAMBERT Academic Publishing

## **Impressum / Imprint**

Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Alle in diesem Buch genannten Marken und Produktnamen unterliegen warenzeichen-, marken- oder patentrechtlichem Schutz bzw. sind Warenzeichen oder eingetragene Warenzeichen der jeweiligen Inhaber. Die Wiedergabe von Marken, Produktnamen, Gebrauchsnamen, Handelsnamen, Warenbezeichnungen u.s.w. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutzgesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Bibliographic information published by the Deutsche Nationalbibliothek: The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <http://dnb.d-nb.de>.

Any brand names and product names mentioned in this book are subject to trademark, brand or patent protection and are trademarks or registered trademarks of their respective holders. The use of brand names, product names, common names, trade names, product descriptions etc. even without a particular marking in this work is in no way to be construed to mean that such names may be regarded as unrestricted in respect of trademark and brand protection legislation and could thus be used by anyone.

Coverbild / Cover image: [www.ingimage.com](http://www.ingimage.com)

Verlag / Publisher:

LAP LAMBERT Academic Publishing

ist ein Imprint der / is a trademark of

OmniScriptum GmbH & Co. KG

Heinrich-Böcking-Str. 6-8, 66121 Saarbrücken, Deutschland / Germany

Email: [info@lap-publishing.com](mailto:info@lap-publishing.com)

Herstellung: siehe letzte Seite /

Printed at: see last page

**ISBN: 978-3-659-69717-3**

Copyright © 2015 OmniScriptum GmbH & Co. KG

Alle Rechte vorbehalten. / All rights reserved. Saarbrücken 2015

**Inas Ali**  
**Soukaena Hassan**

**A Model To Detetct DOS Using Data Mining Classification  
Algorithms**

## Table of Contents

Subject	Page No.
Table of Contents	1
List of Abbreviations	4
List of Figures	6
List of Tables	8
List of Algorithms	9
<b>Chapter One: General Introduction</b>	
1.1 Overview	10
1.2 Data Mining Applications	11
1.3 Literature Survey	12
1.4 Aim of This Work	14
1.5 Thesis Layout	15
<b>Chapter Two: Data Mining</b>	
2.1 Introduction	16
2.2 Data Mining	17
2.3 Data Mining and Knowledge Discovery in Databases	18
2.4 Input: Concepts, Values, Features, Objects and Datasets	21
2.4.1 Concepts	21
2.4.2 Values	22
2.4.3 Features	22
2.4.4 Objects	23
2.4.5 Datasets	24
2.5 Data Repositories for Data Mining	24
2.6 Data Mining Methods	26
2.6.1 Classification	26
2.6.2 Mining Frequent Patterns, Associations, and Correlations	36
2.6.3 Clustering	40

2.6.4	Prediction	40
-------	------------	----

### **Chapter Three: Security and Intrusion Detection System**

3.1	Introduction	41
3.2	Security Attacks	41
3.2.1	Passive Attack	42
3.2.2	Active Attack	43
3.3	Malicious Software	45
3.3.1	Malware Requiring Host	45
3.3.2	Malware Not Requiring Host	46
3.3.3	Malware Countermeasure Requirements	50
3.4	Intrusion Detection System	50
3.4.1	False Positives and False Negatives	52
3.4.2	Components of Intrusion Detection System	52
3.4.3	Intrusion Examples	53
3.5	Generic Intrusion Detection Model	53
3.6	Classification of Intrusion Detection Systems	55
3.6.1	According to Intrusion Detection Approach	55
3.6.2	According to Data Source of Intrusion Detection System	57
3.6.3	According to Intrusion Detection System Reaction	59
3.6.4	According to Analysis Timing	60
3.7	Intrusion Detection System with Data Mining	61
3.7.1	Development of Data Mining Algorithms for Intrusion Detection	61
3.7.2	Association and Correlation Analysis, and Aggregation to Help Select and Build Discriminating Attributes	62
3.7.3	Analysis of Stream Data	63

### **Chapter Four: Design, Implementation, and Results of the Proposed Model**

4.1	Introduction	64
4.2	General Design of the Proposed Model	65

4.2.1	The Proposed HybD Dataset	65
4.2.2	Preprocessing on the 10% KDD'99 Dataset and Constructing the Proposed HybD Dataset	69
4.2.3	Feature Selection	76
4.2.4	Classifiers Building	82
4.3	Implementation of the Proposed Model	87
4.3.1	Implementation of Preprocessing on the 10% KDD'99 Dataset and Constructing the Proposed HybD Dataset	88
4.3.2	Implementation of Feature selection	91
4.3.3	Implementation of the Classifiers Building	94
4.4	Experimental Work and Results	97

## **Chapter Five: Conclusions and Suggestions for Future Work**

5.1	Conclusions	105
5.2	Suggestions for Future Work	106
<b>References</b>		108
<b>Appendix A</b>	SNORT	112
<b>Appendix B</b>	The 10% KDD'99 Dataset	114
<b>Appendix C</b>	NB Classifier with AR Method	119

## Table of Contents

Subject	Page No.
Table of Contents	1
List of Abbreviations	4
List of Figures	6
List of Tables	8
List of Algorithms	9
<b>Chapter One: General Introduction</b>	
1.1 Overview	10
1.2 Data Mining Applications	11
1.3 Literature Survey	12
1.4 Aim of This Work	14
1.5 Thesis Layout	15
<b>Chapter Two: Data Mining</b>	
2.1 Introduction	16
2.2 Data Mining	17
2.3 Data Mining and Knowledge Discovery in Databases	18
2.4 Input: Concepts, Values, Features, Objects and Datasets	21
2.4.1 Concepts	21
2.4.2 Values	22
2.4.3 Features	22
2.4.4 Objects	23
2.4.5 Datasets	24
2.5 Data Repositories for Data Mining	24
2.6 Data Mining Methods	26
2.6.1 Classification	26
2.6.2 Mining Frequent Patterns, Associations, and Correlations	36
2.6.3 Clustering	40



2.6.4	Prediction	40
-------	------------	----

### **Chapter Three: Security and Intrusion Detection System**

3.1	Introduction	41
3.2	Security Attacks	41
3.2.1	Passive Attack	42
3.2.2	Active Attack	43
3.3	Malicious Software	45
3.3.1	Malware Requiring Host	45
3.3.2	Malware Not Requiring Host	46
3.3.3	Malware Countermeasure Requirements	50
3.4	Intrusion Detection System	50
3.4.1	False Positives and False Negatives	52
3.4.2	Components of Intrusion Detection System	52
3.4.3	Intrusion Examples	53
3.5	Generic Intrusion Detection Model	53
3.6	Classification of Intrusion Detection Systems	55
3.6.1	According to Intrusion Detection Approach	55
3.6.2	According to Data Source of Intrusion Detection System	57
3.6.3	According to Intrusion Detection System Reaction	59
3.6.4	According to Analysis Timing	60
3.7	Intrusion Detection System with Data Mining	61
3.7.1	Development of Data Mining Algorithms for Intrusion Detection	61
3.7.2	Association and Correlation Analysis, and Aggregation to Help Select and Build Discriminating Attributes	62
3.7.3	Analysis of Stream Data	63

### **Chapter Four: Design, Implementation, and Results of the Proposed Model**

4.1	Introduction	64
4.2	General Design of the Proposed Model	65

4.2.1	The Proposed HybD Dataset	65
4.2.2	Preprocessing on the 10% KDD'99 Dataset and Constructing the Proposed HybD Dataset	69
4.2.3	Feature Selection	76
4.2.4	Classifiers Building	82
4.3	Implementation of the Proposed Model	87
4.3.1	Implementation of Preprocessing on the 10% KDD'99 Dataset and Constructing the Proposed HybD Dataset	88
4.3.2	Implementation of Feature selection	91
4.3.3	Implementation of the Classifiers Building	94
4.4	Experimental Work and Results	97

## **Chapter Five: Conclusions and Suggestions for Future Work**

5.1	Conclusions	105
5.2	Suggestions for Future Work	106
<b>References</b>		108
<b>Appendix A</b>	SNORT	112
<b>Appendix B</b>	The 10% KDD'99 Dataset	114
<b>Appendix C</b>	NB Classifier with AR Method	119

## List of Abbreviations

Abbreviation	Description
2-D	2 Dimensional
Acc	Accuracy
AI	Artificial Intelligence
ANN	Artificial Neural Network
AR	Association Rules
ATM	Automated Teller Machine
AV	Antivirus
CAT	Capillary Agglutination Test
DARPA	Defense Advanced Research Projects Agency
DB	Database
DM	Data Mining
DoS	Denial of Service
DR	Detection Rate
DW	Data Warehouse
ECOS	Simple Evolving Connectionist System
EKG	Electrokymogram
FAR	False Alarm Rate
FN	False Negative
FP	False Positive
FTP	File Transfer Protocol
GR	Gain Ratio
HIDS	Host Intrusion Detection System
HybD	Hybrid Dataset
ID	Intrusion Detection
ID3	Interactive Dichotomizer 3
IDM	Intrusion Detection Model
IDS	Intrusion Detection System
IIS	Internet Information Server
InfoGain	Information Gain
IP	Internet Protocol
KDD	Knowledge Discovery in Databases
KDP	Knowledge Discovery Process
KNN	$k$ -nearest neighbor

Abbreviation	Description
LCS	Learning Classifier Systems
Malware	Malicious Software
MAP	Maximum Posteriori
min_conf	Minimum Confidence
min_sup	Minimum Support
MRI	Magnet Resonance Imaging
NB	Naïve Bayesian
NIDS	Network Intrusion Detection System
OS	Operating System
PC	Personal Computer
R2L	Remote to Local
SCAN	Systolic Coronary Artery Narrowing
SQL	Structured Query Language
SVM	Support Vector Machine
TCP	Transmission Control Protocol
TestingHD	Testing Hybrid Dataset
TN	True Negative
TP	True Positive
TrainingHD	Training Hybrid Dataset
U2R	User to Root
UNIX	Uniplexed Information and Computing Service
WWW	World Wide Web

## List of Figures

Figure No.	Title	Page No.
2.1	General Block Diagram of KDD	19
2.2	Relationship among Values, Features, Objects, Dataset, DBs, and DWs	21
2.3	Connection Record (Object)	23
2.4	Relational DB	24
2.5	Typical Architecture of a DW System	25
2.6	Example of Decision Tree Structure	28
3.1	Security Services	42
3.2	Passive Attacks	43
3.3	Active Attacks	44
3.4	Malicious Software	45
3.5	A Generic IDM	54
3.6	Classification of IDSs	56
3.7	Typical location for a NIDS	58
3.8	HIDS	59
4.1	Flowchart of the General Structure of the Proposed Model	66
4.2	The 10% KDD'99 Dataset	88
4.3	The Proposed HybD Training Dataset	90
4.4	Transformed HybD Training Dataset	90
4.5	Selection of the <i>min_sup</i> Threshold	91
4.6	Results of Customized A priori algorithm	92
4.7	Results of Customized ReliefF Measure	93
4.8	Results of Customized GR Measure	93
4.9	Classifiers Construction	94
4.10	Classification with Customized NB Classifiers	95
4.11	Customized ID3 Classifiers: Sets of Classification Rules	95
4.12	Classification with Customized ID3 Classifiers	96

4.13	Classification Results of NB and ID3 Classifiers	101
4.14	DR and FAR of NB and ID3 Classifiers	103
4.15	Accuracy Ratios of ID3 and NB Classifiers	104

## List of Tables

Table No.	Title	Page No.
2.1	Fragment of a Transactional DB	25
2.2	"Play Tennis" Training Dataset	27
2.3	Sample Dataset for NB Classifier	34
2.4	Sample Transactions to demonstrate AR Mining	39
2.5	Computation of Frequent Itemsets	40
4.1	Host-based Features	69
4.2	Selected Features according to AR, ReliefF, GR	98
4.3	Classification Results of NB and ID3 Classifiers	100
4.4	DR and FAR of NB and ID3 Classifiers	102
4.5	Accuracy Ratios of NB and ID3 Classifiers	103

## List of Algorithms

Algorithm No.	Title	Page No.
2.1	The Hunt's Fundamental	29
2.2	ID3	30
2.3	A priori	38
2.4	AR Generation	39
4.1	Convert Notepad To Array	70
4.2	Selecting Records with Specific Classes	71
4.3	Selecting Training&Testing Datasets	71
4.4	Adding Host-based Features to Dataset	72
4.5	Transformation	74
4.6	Feature_Selection	76
4.7	Customized_Apriori	77
4.8	Itemset_Extraction	78
4.9	Extract_Items	78
4.10	Find_Frequent_Items	79
4.11	Construct_New_Itemsets	79
4.12	Customized ReliefF	81
4.13	Customized Gain_Ratio	82
4.14	Classifiers Construction	83
4.15	Customized Naive Bayesian	85
4.16	Customized ID3	86
4.17	Number&Name of Classes	86
4.18	Initial_Entropy	87
4.19	Classification With ID3 Rules	87



# Chapter One

## General Introduction

### 1.1 Overview

With the rapid expansion of computer systems and the significant developments in the new technologies in this domain, the important data are under constant threats of intrusion, which is defined to be any unauthorized access attempt to manipulate, modify, or destroy information, or to render a system unreliable or unusable. All those make the security a critical issue for computer systems [Na06a]. Malware is defined to be a program that performs a malicious function, such as compromising a system security, damaging a system or obtaining sensitive information without user permission [Ga08]. Many methods have been developed to secure the system infrastructure and communication over the Internet such as the use of antivirus (AV), firewalls, encryption, and IDS [Al11].

With the increasing creativity of intrusions, the development of effective IDS is becoming a greater challenge. ID is a set of techniques and methods that are used to detect suspicious activity in computer systems, both at network and computer level. Therefore, the main goal of IDS is to identify unauthorized use, misuse, and external penetrations [Na06a]. DM-based intrusion detection (ID) framework can detect new intrusions accurately and automatically. The DM methods automatically find patterns in the used dataset and use these patterns to detect a set of new intrusions. By comparing detection methods that use DM with a traditional signature based methods; it is seen that, DM-based detection methods are more than doubling the current detection rates for new malwares [Sc01].

DM-based ID techniques generally fall into two main categories: *misuse detection* and *anomaly detection*. In misuse detection techniques, patterns of well-known attacks are used to match and identify known attacks and their variations, but they are not effective against novel attacks that have no matched rules or patterns yet. Anomaly detection techniques, on the other hand, build models of normal behavior, and flag observed activities that deviate significantly from the established normal