An Efficient Approach for Classifying Intrusion using Fusion based HMM & Clustering

Chandrima Dutta
Department of Computer Science
Truba Institute of Science & IT
Bhopal, india
Cdutta03@gmail.com

Prof. Amit Saxena
Department of Computer Science
Truba Institute of Science & IT
Bhopal, India
amitsaxena@trubainstitute,ac,in

Dr. Manish Manoria
Director
Truba Institute of Science & IT
Bhopal, India
manishmanoria@trubainstitute.ac.in

Abstract—Intrusion detection system is a method of identifying unnecessary packets that may be creates some damage in the network; hence various Intrusion detection based methods are implemented to provide security in the network traffic flow. Here in this paper an efficient technique of identifying intrusions is implemented using the concept of hidden markov model and then classification of these intrusions is done. The methodology implemented here is applied on KDDCup 99 dataset where the data to be detected is first group some by using clustering approach so that the similar packets are grouped into one and the dissimilar packets are grouped into another. Now some of the important attributes are selected from the dataset and defined as the states of Hidden Markov Model and the probability is calculated from each of the state to other state and finally these probabilities are fused to find the overall probability of the dataset and hence on the basis of threshold probability packets can be classified as low and medium and high intrusions.

Index Terms—Component, formatting, style, styling, insert.

I. Introduction

Protecting networks from computer safety attacks is a vital apprehension of computer security. Comprehensive collection and straight clarification of traffic information are core problems in network traffic anomaly detection. As network traffic may lead to variety of information exchange and sensitive data transfer. Although it is also well known that the dependency of network are also emerging rapidly. Due to this the network traffic circumstance are very vital now a days and it will become more complicated in forthcoming time. Many host-based anomaly detection systems have been proposed to notice server compromises to detect intrusions by monitoring the execution of a program to see if its behavior conforms to a model that describes its normal behavior [1].

On the basis of the natural complexity in characterizing the standard complex routine, the complexity of irregularity uncovering may be categorized as model based and non-model based. According to model based anomaly detectors, it is assumed that an identified model is accessible for the normal behaviour of definite specific aspects of the network and any divergence from the norm is supposed an anomaly. Network behaviours that cannot be characterized by any model for such condition non-model based approaches are used. Non-model based approaches can be auxiliary classified based on the unambiguous implementation and accuracy constraints that have been imposed on the detector.

The HMM (Hidden Markov Model is based on the concept of selecting a number of states with a number of hidden states where the probability from each of the state is computed and a hidden state is stored. It is a model of generating a number of sequence of hidden states as well as the observable sequences of states.

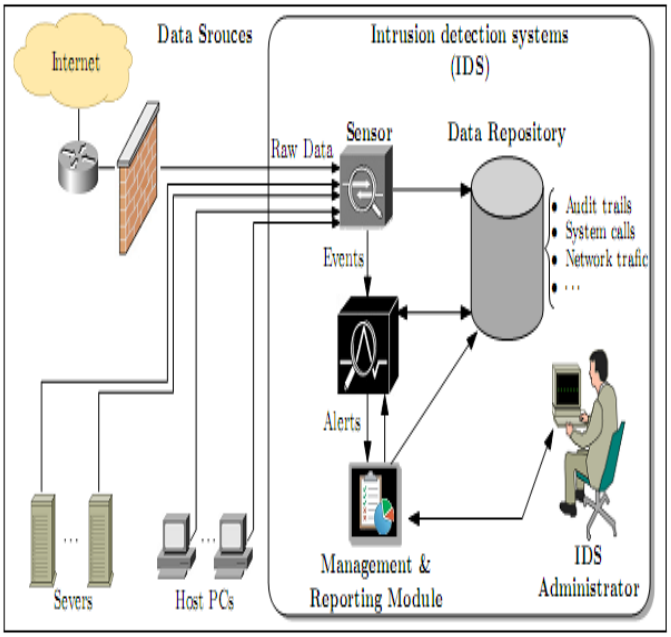

Figure 1. High level architecture of an intrusion detection system.

Network intrusion detection: Network intrusion detection systems listen to network communications. They are acquainted with intrusions which come during the networking environment. Basically a system intrusion exposure system (NIDS) is a service which listens on a network interface looking for suspicious traffic. Network intrusion detection systems are mostly signature based.

Host Based Intrusion Detection: Host intrusion detection systems (HIDS) inhabit on a resource supervised. This resource is mostly a computer server or workstation. HIDS appear at shaped log files, changes in the file system or check for changes in the process table. Their objective is to identify intrusions into a host.

***Signature based intrusion detection*:** Signature based intrusion is based on signatures of known attacks. These signatures are accumulated and evaluated against events or received traffic. If a pattern matches, an alert is generated.

II. Literature Survey

Wenying Feng, Qinglei Zhang, Gongzhu Hu, Jimmy Xiangji Huang proposed a new and efficient technique for the detection of intrusions using the hybrid combination of Support vector machine and Ant colony networks [1]. Since Data mining is a technique for extracting some meaningful information from it so that the proper and quick analysis can be done. Intrusion Detection is one of application of data mining in which the packets to be send from source to destination needs to be filtered and if the packet contains any attack it can be detected with high alarm rate. Here in the paper a new technique of identifying these intrusions using machine learning approach such as Support vector machine. The classification of intrusions in the packets using support vector machine is an efficient way of classifying the attacks in the packet.

Chih-Fong Tsai, yu-Feng Hsu, Chia-Ying Lin,Wei-Yang Lin also proposed a new technique of identifying intrusions by analyzing various intrusion detection techniques their advantages and limitations [2]. The paper summarized all the intrusion detection technique implemented by analyze their various advantages and limitations. The paper discusses and compares 55 related articles from the period of 2000 to 2007 in which various classification and clustering techniques for the detection of intrusions are implemented and analyzed.

Latifur Khan, Mamoun Awad, Bhavani Thuraisingham implemented an efficient technique for the intrusion detection using support vector machine and the clustering using hierarchical clustering [3]. Here in this paper the combination of hierarchical clustering and then classification using support vector machine is proposed which provides high true positive and accuracy as compared to the existing techniques for the detection of intrusions. The input dataset is first clustered into 'N' groups according to the classification classes and then these clustered groups are classified using machine learning approach such as support vector machine. This methodology greatly classifies the dataset and provides high alarm rate for the detection of intrusions.

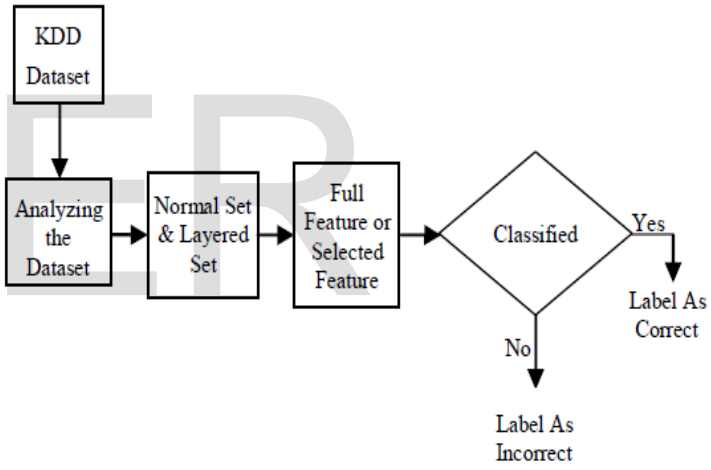

Figure 2. Representation of Layered Approach

Jianfeng Pu and Lizhi Xiao proposed a new technique for the Network Intrusion detection which is based on the concept of Support vector machine and Ant Colony Algorithm [4]. A hybrid combinatorial method of applying support vector machine and ant colony algorithm for the detection of network packets contains anomalous behavior. Support vector machine is used for the selection of important features of the packets that flow in the network.

Shelly Xiaonan Wu, Wolfgang Banzhaf has given a brief overview of intrusion detection systems and their various limitations and advantages [5]. The paper summarizes the various computational intelligence that may use for the detection of network intrusions in the network or packets. The various artificial intelligence techniques such as Fuzzy and

swarm for the detection of network anomalies can be discussed and analyzed their various advantages and issues.

Qinglei Zhang and Wenying Feng also proposed the same technique for the detection of intrusions using the hybrid combinatorial method of Ant Colony Algorithm and Support Vector Machine [6]. Here in this paper two supervised techniques are combined for the detection of intrusions in the packet. The various experimental results performed on the network packets shows that the hybrid combination is better in performance as compared to the existing Support Vector machine.

S. Janakiraman, V. Vasudevan introduced a new technique for the detection of intrusions using the concept of Ant Colony in Distributed Systems [7]. For the Security of Computer Networks Distributed Intrusion Detection plays a vital role. For the better working of Intrusions as compared to the conventional intrusion detection systems alerts are provided in order to increase the performance of the distributed systems. This paper presents an intelligent learning approach using Ant Colony Optimization (ACO) based distributed intrusion detection system to detect intrusions in the distributed network.

Snehal A. Mulay, P.R. Devale, G.V. Garje proposed a new and efficient technique for the detection of intrusions using combination of Support vector machines and Decision Tree [8]. Support Vector machine is a supervised learning approach which is used for the binary classification so that multiple class problems can be solved easily and quickly, Decision tree based support vector machine can be used for solving multi-class problems more efficiently. By using the combination method of SVM with decision tree can decreases the training and testing time as well as system efficiency also increases.

Wenke Lee proposed a new and efficient framework for the building of intrusions in DARPA Datasets [9]. Here in the paper a new Data mining framework is implemented for the building of the detection of Intrusions in the system. The main scenario of the methodology is to provide the auditing programs so that a set of features from each of the network is applied on the intrusion and anomalies for the categorization of the normal and abnormal activities.

The configuration of analysis engines, update of data repository and response to alerts are among the responsibilities of the IDS administrator. When alerts are raised, the IDS administrator should prioritize and investigate incidents to refute or confirm that an attack has actually occurred. If an intrusion attempt is confirmed, a response team should react to limit the damage, and a forensic analysis team should investigate the cause of the successful attack. An IDS may include a response module that undertake further actions either to prevent an ongoing attack or to collect additional supporting information – it is often referred to as intrusion prevention system (IPS) or intrusion detection and prevention system (IDPS), [10-13]. IDSs are typically categorized depending on their monitoring scope (or location of the sensors) into network-based and host-based intrusion detection systems. They are also classified based on the detection methodology (employed by the analysis engine) into misuse and anomaly detection. More detailed taxonomies have been also developed, which further classify IDSs according to their architecture (centralized or fully distributed), behavior after attacks (passive or active), processing time (on-line or off-line), level of inspection (stateless or state full), etc. [12] [14-16].

Network-based IDSs (NIDSs) monitor the network traffic for multiple hosts by capturing and analyzing network packets for patterns of malicious activities. An NIDS is typically a stand-alone device that can control a network of arbitrary size with a small number of sensors [17]. NIDSs capture network traffic in promiscuous mode by connecting to a hub, network switch configured for port mirroring, or network tap [18].

Host-based IDSs (HIDSs) are designed to monitor the activity of a host system, such as a mail server, web server, or an individual workstation. HIDSs identify intrusions by analyzing operating system calls, audit trails, application logs, file-system modifications (e.g., password files, access control lists, executable files), and other host activities and state [19] [20] HIDSs are typically software-based systems which should be installed on every host of interest.

III. Proposed Methodology

The various parameters used in HMM such as :

1) 'N' represents the no. of states in the model.
2) There are various individual states in the model as S={S1,S2,S3…..Sn}
3) State a particular instant of time't' is q_t.
4) 'M' is the number of distinct observation symbols per state; these observation symbols correspond to the physical output of the system being modeled.
5) Various individual symbols are denoted as V={V1,V2,…..Vm}.
6) 'A' is represented as the probability of distribution during the transition of states.
7) 'B' is represented as probability of distribution of the observational symbols. It can be represented as:

$$B = b_{jk}$$

8) $'\pi'$ can be represented as probability distribution of initial state of transition and can be represented as :

$$\pi = \{\pi_i\}n$$

9) The various sequences of the state's OO=OO1, OO2, OO3….OOT', known as indirect observation of the hidden states and T' can be represented as the number of total number of observations taken.

The proposed methodology based on hidden markov model using behavioral distance contains the following parameters as $\gamma = \{A, B, \pi\}$

Here N can be represented as the hidden states let us take it as 5. Here M can be represented as observations to be taken. if the hidden states are taken as SS1, SS2, and SS3 and for SS4, and SS5 the value is 2.

The probability distribution of the state transition is

$A = \{a_{ij}\}$

Where

$a_{ij} = P[q_{t+1} = S_j | q_t = S_i]$, 1<=j<=5 and 1<=i<=5

The probability distribution of observation symbol in state j, $B = \{b_j(k)\}$

Where

$b_j(k) = P[V_k \ at \ t | q_t = S_i]$, 1<=j<=5, 1<=k<=6 if j=1,2 or 3 else 1<=k<=2

The figure shows the transition of different states, where the links connected represents transition probability of the states.

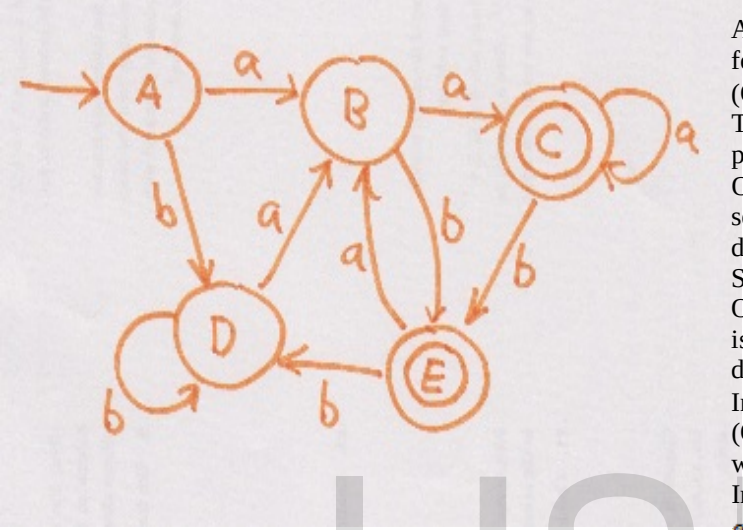

Figure 3. Example of Transition of States using HMM

Welch proposed a hidden markov model which contains and starts with the estimate initial state and likelihood value is used to find the local maxima value.

States	Initial State Distribution Value (π_i)
1	0.000482
2	0.249375
3	0.079432
4	0.36835
5	0.246293

States	1	2	3	4	5
1	0.1658	0.1356	0.2659	0.1853	0.157
2	0.0465	0.2850	0.1759	0.2845	0.1745
3	0.0275	0.1385	0.2759	0.1773	0.2942
4	0.5843	0.1853	0.0649	0.0023	0.0046
5	0.1395	0.1548	0.1844	0.2753	0.2352

After parameter estimation step, Forward Procedure is applied for training HMM. The forward variable:=P (OO1,OO2,OO3,OO4,OO5, qt = Si |λ) …… (1)

The forward variable 'P' indicates the probability of the partial observation sequence OO1, OO2, OO3, OO4, and OO5, and the state Si at time t, given the model. λ Observation sequences OO1, OO2, OO3, OO4, and OO5 represent the discrete observation symbol number of the state's SS1, SS2, SS3, SS4, and SS5 respectively. Thus, in our case values of OO1, OO2, OO3 ranges from 1 to 6 and for OO4 and OO5 it is either 1 or 2. Steps involved in the Forward Procedure are described using equations (2), (3), and (4):

Initialization of the forward variable value α t(i) = π i * bi (OO1) ..... (2)

where 1 <= i <= 5

Induction step of the Forward Procedure $\alpha_{(t+1)}(j) = \left[\sum_{i=1}^{5} \alpha_t(i) * a_{ij}\right] * b_j(O_{t+1})$……..(3)

Where 1<=t<=T-1 and 1<=j<=5.

Termination step of the Forward Procedure $P(O|\lambda) = \sum_{i=1}^{5} \alpha_t(i)$……..(4)

Thus, P (O |λ) is the sum of all the π t (i) values.

Now the probability distribution of each of the state is computed and then fused the probability distribution from each of the state to get an average probability distribution, which is then compared with each of the individual state and then according to the probability distribution the anomalies are classified as normal, medium or high type of anomaly.

IV. Result Analysis

The table shown below is the fused probability that can be estimated using hidden markov model. The fused probability can vary with the number of states selected for the HMM.

No. of states/attributes	Fused Probability
2	0.52
3	0.51
4	0.629
5	0.62
6	0.58
7	0.546
8	0.63
9	0.639
10	0.571

11	0.551
12	0.591
13	0.61
14	0.63

Table 1 Fused Probabilities based on states

The table shown below is the comparison of HMM based Detection Ratio and the proposed Fusion based HMM on the basis of number of packets that are flowing in the network. Hence on the basis of these attributes the detection ratio can be computed as total number of packet flow to the total number of packets in which exactly the intrusions are detected. Also the difference in performance of the HMM and the fusion based HMM methodology is shown in the table.

No. of Packets	Detection Ratio Existing	Detection Ratio Proposed	Detection Difference Between Existing and Proposed System
10	87%	98%	11%
20	89%	98%	9%
50	92%	99%	7%
100	93%	99%	6%
200	95%	99%	4%
500	96%	99%	3%

Table 2 Analysis of Detection Ratio

The table shown below is the comparison of HMM Accuracy (%) and proposed fusion based HMM on the basis of number of instances available in the dataset. Hence on the basis of these attributes the Accuracy can be computed as total number of instances in which exactly the intrusions are detected. Also the difference in performance of the HMM and the proposed fusion based HMM methodology is shown in the table.

No. of instances	Existing Work	Proposed Work
50	86.4	96.5
60	86.8	96.7
70	87	97
80	88	97.2
90	88.2	97.5
100	88.4	97.8
110	88.6	98
120	89	98.6
130	89.4	98.8
140	89.7	98.6
150	90.1	98.4

Table 4 Comparison of Accuracy

The figure shown below is the comparison of HMM based Detection Ratio and the proposed Fusion based HMM on the basis of number of packets that are flowing in the network. Hence on the basis of these attributes the detection ratio can be computed as total number of packet flow to the total number of packets in which exactly the intrusions are detected.

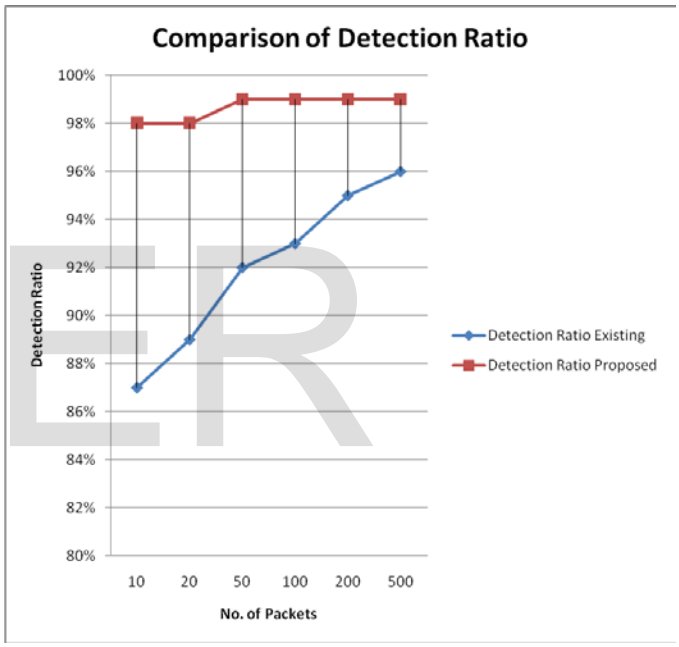

Figure 4. Analysis & comparison of Accuracy

The table shown below is the comparison of HMM Error Rate and proposed fusion based HMM on the basis of number of packets that are flowing in the network. Hence on the basis of these attributes the Error rate can be computed as total number of packet flow to the total number of packets in which exactly the intrusions are detected.

No. of states/attributes	Existing Work	Proposed Work
2	0.52	0.23
3	0.62	0.21
4	0.68	0.32
5	0.57	0.31

6	0.53	0.27
7	0.6	0.26
8	0.57	0.31
9	0.47	0.25
10	0.67	0.26
11	0.73	0.25
12	0.53	0.32
13	0.62	0.34
14	0.61	0.32

Table 5 Comparison of Error Rate

The figure shown below is the comparison of HMM Error Rate and proposed fusion based HMM on the basis of number of packets that are flowing in the network. Hence on the basis of these attributes the Error rate can be computed as total number of packet flow to the total number of packets in which exactly the intrusions are detected. Also the difference in performance of the HMM and the fusion based HMM methodology is shown in the figure.

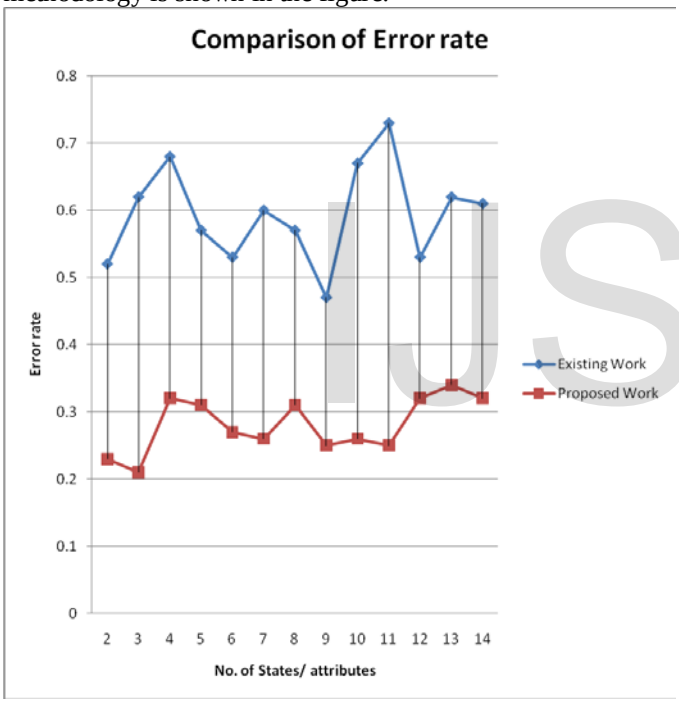

Figure 5 Analysis & comparison of Error Rate

V. Conclusion

Intrusion Detection System is the detected of some anomalous or some unwanted packets that may create some harm to the network. Although there are various techniques implemented for the detection of intrusions in the packet. The existing technique implemented for the detection of Anomalous behavior in the network using Hidden Markov Model (HMM) is efficient in terms of Accuracy and Error rate and detection Ratio, but there are some limitations which needs to be improve further.

Hence a new and efficient technique is implemented here for the detection of intrusions in using Fusion based Hidden Markov Model which provides better results as compared to the existing HMM based intrusion detection. The experimental result shows the performance of the proposed methodology. The fusion based HMM provides efficient Detection Ratio as well provides better accuracy as compared to the HMM based detector.

References

[1] Wenying Feng, Qinglei Zhang, Gongzhu Hu, Jimmy Xiangji Huang,"Mining Network data for intrusion detection through combining SVM's with ant colony networks", *Future Generation Computer Systems 37 (2014) 127-140*", Elsevier 2014.

[2] Chih-Fong Tsai, yu-Feng Hsu, Chia-Ying Lin,Wei-Yang Lin," Intrusion Detection by machine learning: A Review", *Expert Systems with Applications 36 (2009) 11994-12000*,Elsevier 2009.

[3] Latifur Khan, Mamoun Awad, Bhavani Thuraisingham," A new intrusion detection system using support vector machine and hierarchical clustering", *The VLDB Journal (2007) 16:507-521*,2007.

[4] Jianfeng Pu, Lizhi Xiao," A Detection of Network Intrusion Based on SVM and Ant Colony Algorithm", *National Conference on Information Technology and Computer Science (CITCS)*, 2012.

[5] Shelly Xiaonan Wu, Wolfgang Banzhaf," The use of Computational Intelligence in intrusion detection systems: A Review", *Applied Soft Computing 10*, Elsevier, 2010.

[6] Qinglei Zhang and Wenying Feng," Network Intrusion Detection by Support Vectors and Ant Coloy", *Proceedings of the 2009 International Workshop on Information Security and Application (IWISA 2009) Qingdao, China, November 21-22*, 2009.

[7] S. Janakiraman, V. Vasudevan," ACO based Distributed Intrusion Detection System", 2008.

[8] Snehal A. Mulay, P.R. Devale, G.V. Garje," Intrusion Detection System using support vector machine and Decision Tree", *International Journal of Computer Applications (0975-8887)*, Volume 3-No.3, June 2010.

[9] Wenke Lee, "A Data Mining Framework for building Intrusion Detection Models", 1999.

[10] Ghorbani Ali A., Lu Wei, Tavallaee Mahbod, Ghorbani Ali A., Lu Wei, and Tavallaee Mahbod, 2010. Intrusion response. Jajodia Sushil, editor, Network Intrusion Detection and Prevention, volume 47 of Advances in Information Security, pages 185–198. Springer US ISBN 978-0-387-88771-5

[11] Rash Michael, Orebaugh Angela D., Clark Graham, Pinkard Becky, and Babbin Jake, 2005. Intrusion Prevention and Active Response: Deployment Network and Host IPS. Syngress.

[12] Scarfone Karen and Mell Peter, February 2007. Guide to intrusion detection and prevention systems (IDPS). Recommendations of the National Institute of Standards

and Technology sp800-94, NIST, Technology Administration, Department of Commerce, USA, 2007.

[13] Stakhanova Natalia, Basu Samik, and Wong Johnny, 2007. A taxonomy of intrusion response systems. International Journal of Information and Computer Security, 1(1/2):169–184.

[14]Tucker C.J., Furnell S.M., Ghita B.V., and Brooke P.J., 2007. A new taxonomy for comparing intrusion detection systems, Internet Research, 17:88–98.

[15] Lazarevic Aleksandar, Kumar Vipin, and Srivastava Jaideep, 2005. Intrusion detection: A survey. Kumar Vipin, Srivastava Jaideep, and Lazarevic Aleksandar, editors, Managing Cyber Threats, volume 5 of Massive Computing, pages 19–78. Springer US ISBN 978-0-387-24230-9

[16] Estevez-Tapiador Juan M., Garcia-Teodoro Pedro, and Diaz-Verdejo Jesus E., 2004. Anomaly detection methods in wired networks: A survey and taxonomy. Computer Communications, 27(16):1569–1584. ISSN 0140-3664.

[15] Northcutt Stephen and Novak Judy, 2002. Network Intrusion Detection: An Analyst's Handbook. New Riders Publishing, Thousand Oaks, CA, USA, 3rd edition ISBN 0735712654

[16] Peng Tao, Leckie Christopher, and Ramamohanarao Kotagiri, April 2007. Survey of network-based defense mechanisms countering the DoS and DDoS problems. ACM Computing Surveys (CSUR), 39(1):1–42 ISSN 0360-0300.

[17] De Boer Pieter and Pels Martin, 2005. Host-based intrusion detection systems. Technical Report 1.10, Faculty of Science, Informatics Institute, University of Amsterdam.

[18] Vigna G. and Kruegel C., December 2005. Host-based intrusion detection systems. Big-doli H., editor, Handbook of Information Security, volume III, Wiley.