Detection and Classification of Intrusions using Fusion Probability of HMM

Hemlata Sukhwani
M. Tech. Scholar (CSE)
Oriental Institute of Science and Technology, Bhopal

Shwaita Kodesia
Assistant Professor (CSE)
Oriental Institute of Science and Technology, Bhopal

Sanjay Sharma
Assistant Professor (CSE)
Oriental Institute of Science and Technology, Bhopal

ABSTRACT
Intrusion detection system is a technique of identifying unwanted packets that creates harm in the network; hence various IDS are implemented for the security of network traffic flow. Here in this paper an efficient technique of identifying intrusions is implemented using hidden markov model and then classification of these intrusions is done. The methodology sis applied on KDDCup 99 dataset where the dataset is first clustered using K-means algorithms and then a number of attributes is selected which are used for the detection of intrusion is passed to the HMM, after calculating probability from each of the states, these probabilities are fused to get the resultant final probability and also overall probability is calculated from dataset on the basis of which intrusions are classified as low, medium or high.

Keywords
IDS, Anomaly, HMM, Behavioral Distance

1. INTRODUCTION
The main definition of intrusion detection system starts with the use in hardware or software so that the data or traffic flows through these devices can be filtered since this traffic may contains some unwanted flow data that may harm to the system. Using intrusion detection system the author can not only identify various attacks in the network traffic but can also classify the type of attack. Although there are various techniques implemented for the detection and prevention of network attacks but on the basis of characteristics of the detection of attacks it can be classified as signature based and anomaly based.

1.1 Hidden Markov Model
The HMM Model formed with a definite number of states set. Due to transitions among the states are presiding over by a set of transition probabilities that are associated with each and every state. In an exacting state, a result or observation can be generated as per separate probability distribution associated with the state. It is only the result of finite number of set of transitions, not the state that is able to be seen to an outside viewer. States are hidden to the external observers consequently named as Hidden Markov Model. The Markov Model used for the hidden layer is a first-order Markov Model which indicates that the probability of individual in an exacting state depends no more than on the earlier state. At the same time as in a scrupulous state, the Markov Model is supposed to "emit" a visible keep up a correspondence to that particular state of the Markov model. One of the objectives of using an Hidden Markov Model is to presume from the set of predicated recognizable path that the most likely path in state gap that was trailed by the system. Figure 1 shows the example of HMM [16].

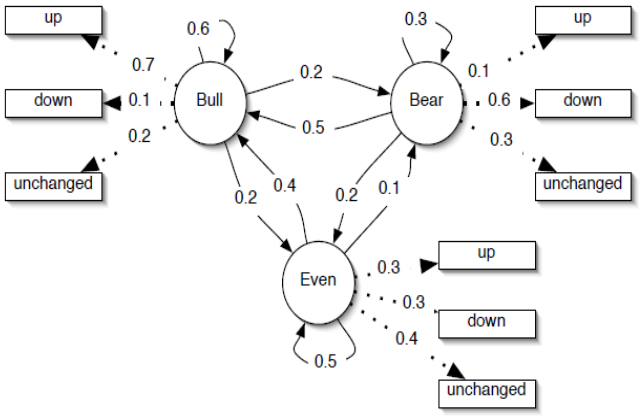

Figure: 1 Example of HMM

There are various techniques that are implemented for the detection of intrusions such as by using of learning [1] which uses the concept of natural immune system, by using the concept of system calls [2], and by using statistical or threshold values [3] and using classification or clustering technique of data mining [4]. From all the techniques implemented for the detection of intrusions HMM is one of the efficient techniques which provide high performance of detection intrusions. But the intrusions detection using HMM takes more training time due to which its efficiency hides.

Depending on the information source considered, an intrusion detection system (IDS) may be either host-based intrusion detection system (HIDS) or network-based intrusion detection system (NIDS).

2. INTRUSION DETECTION AND ITS TYPE
Everywhere IDSs allow for the detection of successful or unsuccessful attempts to compromise systems security. An IDS is an important component of any security infrastructure that complements other security mechanisms. As illustrated in Figure 2, IDS consists of four essential components: sensors, analysis engines, data repository, and management and reporting modules.

An IDS monitors the activity of a target system through a data source, such as system call traces, audit trails, or network packets. Relevant information from these data sources are captured by IDSs sensors, synthesized as events, and forwarded to the analysis engine for on-line analysis or to a repository for off-line analysis.

The analysis engine contains decision-making mechanisms to discriminate malicious events from normal events. It may include anomaly, misuse, or hybrid detection approaches (described next). Outputs from analysis engines include specific information regarding manifestation of suspicious events. This information is stored in a repository for forensics analysis. A management and reporting module receives events that could indicate an attack from the analysis engine, raises an alarm to notify human operators, and reports the relevant information and the level of threat. The management module controls operations of IDS components, such as tuning decision thresholds of the analysis engine and updating the data repository.

The figure 2 shown below shows the various components of intrusion detection and their various levels.

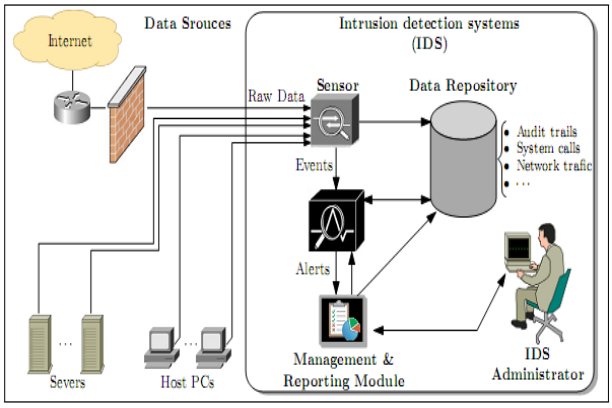

Figure: 2 High level architecture of an intrusion detection system

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), [6-9]. 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 [5]. 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. [8] [10-12].

3. NEED OF IDS

IDS can be illustrate as a specific tool that knows how to read and understand the contents of log files from firewalls, routers, servers, and other network devices. Moreover, a database of identified attack signatures is stored by IDS and it can compare activity patterns, network traffic which observed in the logs. At that position, the IDS can raise the alerts or alarms, take different kinds of usual action ranging from shutting down Internet links or specific servers to launching back traces, and establish additional active efforts to recognize attackers and actively gather facts of their suspicious activities. Specifically, intrusion detection can be describe as a detection of illegal use of or attacks on a system or set of connections. An IDS is needed for detection and determent of such attacks or unauthorized access of systems, networks, and related resources.

4. INTRUSION DETECTION USING HIDDEN MARKOV MODEL

In HMM, the probability with which a given sequence is generated from a model can be calculated using forward-backward procedure and an optimal model can also be built from a collection of sequences using Baum-Welch reestimation formulas. If normal behavior is modeled into an HMM, we can determine whether current behavior is normal or not by comparing the evaluation value of current behavior sequence against the model's threshold for normal behavior. Each HMM determines whether current sequence is abnormal from the measure's point of view it is responsible for and participates in final decision

The figure 3 shown below shows various states of HMM and transitions. This model is used for finding the hidden layer is a first-order Markov Model [15].

It shows the basis model of the HMM which contains a set of initial states a number of hidden states and observed output states.

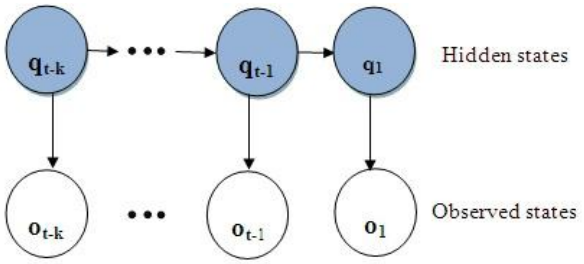

Figure: 3 Simple HMM Architecture

- Graphical Model
- Arrows indicate probabilistic dependencies
- Circles indicate states.

5. PROPOSED METHODOLOGY

The proposed technique is based on the grouping of normal and abnormal behavior in the dataset and then on the basis of the depends attribute in the dataset through intrusions can be detected transition of state and their various probabilities can be calculated and then these probabilities are fused to get a threshold probability which is then compared to classify the type of intrusion in the dataset.

A HMM is characterized by the following elements:

1. N, the number of states in the model.
2. M, the various observations symbols according to the number of states i.e. the discrete alphabet size.
3. A, the state transition probability distribution. In this case is N *N matrix.
4. B, the observation symbol probability distribution. In this case is N*M matrix.
5. P, the initial state distribution. Each element pi is the probability that the initial state is the i-th state.

Here the concept of Classifier Fusion which fuses the outputs of an ensemble of classifiers to produce a single output.

The figure 4 shown below is the flow chart of the proposed methodology. Here KDDCup 99 dataset is used which contains 42 attributes and a number of instance values. The dataset also contains a number of intrusions such as Neptune, smurf and nmap attack [14]. Here clustering is applied on the input dataset which clusters the dataset into two groups cluster-0 and cluster-1 which denotes normal and abnormal values of the dataset. Hidden Markov Model is then applied on the abnormal that data only in result to that it generates probability.

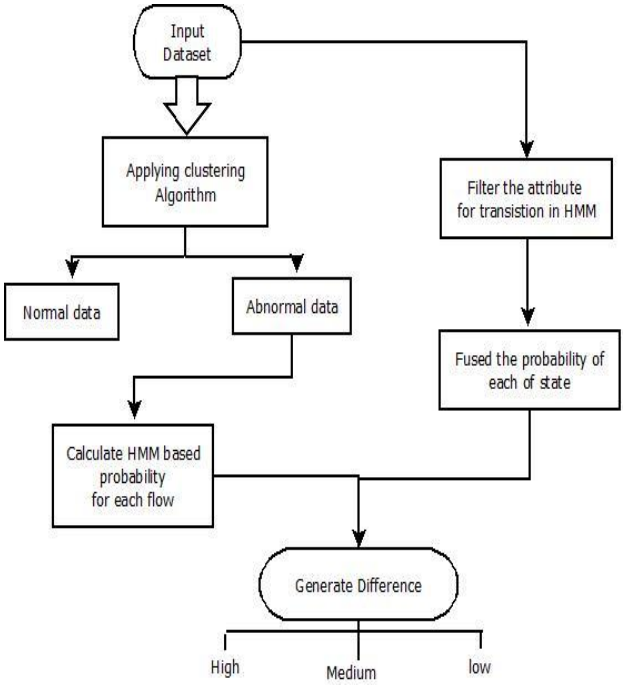

Figure: 4 Outline of the proposed Methodology

The probability calculated here using hidden markov model based on behavioral distance for the entire attributes means hidden markov model contains 42 states on the basis of which probability is calculated. On the other hand the number of attributes is selected which are fully dependent through which intrusions can be calculated are chosen from the dataset and hidden markov model is applied on these attributes to get the probability from each of the attribute and these probabilities are fused to get the resultant probability. Now the difference is computed from the two generated probabilities and the final probability is calculated and compared with the threshold probability to classify the type of intrusion as low, medium or high.

The algorithm contains the following steps:

1. Take an input dataset in which intrusion is detected.
2. Apply clustering algorithm for the grouping of normal and abnormal data in the dataset based on dependent attributes.
3. Compute probability for each data flow in the dataset according to the behavioral distance based Hidden Markov model.
4. For each attributes that are selected for the transition states in the HMM.
5. Calculate probability for each of the state in the HMM, calculates the probability distribution of getting the system call in the process from one state to another.
6. The probability for different states is calculated and the probabilities are fused to get a resultant fused probability.
7. The difference probability of each packet flow in the dataset and final fused result gives probability that can be used for the classification of intrusions.

6. RESULT ANALYSIS

The table 1 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.

Table: 1 Fused probability based on states on Dataset-1

No. of states	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

The table 2 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.

Table: 2 Fused Probability based on states on Dataset-2

No. of states	Fused Probability
2	0.75
3	0.761
4	0.773
5	0.78
6	0.785
7	0.788
8	0.792
9	0.796
10	0.799
11	0.81
12	0.82
13	0.84
14	0.85

The figure shown below is the fused probability from a given set of attributes in the dataset.

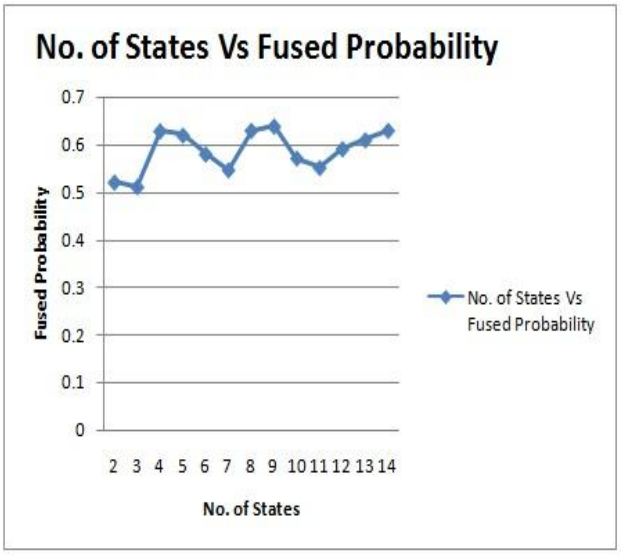

Figure: 5 Fused Probability based on states on Dataset

The table 3 shown below is the comparative analysis of false alarm rate for hidden markov model and Fusion based HMM.

Table: 3 Comparison of False Alarm

	No. of False Alarm	
No. of Requests	**HMM**	**FHMM**
50	5	15
60	6	18
70	25	30
80	10	15
90	10	19
100	5	8

7. CONCLUSION

The proposed methodology implemented here for the detection and classification of intrusion using HMM provides efficient results. The technique implemented here uses the attributes of the dataset as the states to predict the probability and finally probabilities from all the states is fused to get the final output probability which is then compares to the probability predicted from all the states of the dataset. The methodology not only detects the intrusion but also classifies the intrusion as low, medium or high intrusion.

8. ACKNOWLEDGMENT

I would like to express my deep gratitude to Asst. Prof. Miss Shwaita Kodesia and Asst. Prof. Mr. Sanjay Sharma, Computer Science Department, Oriental Institute of Science and Technology, Bhopal, my research supervisors, for their guidance, enthusiastic encouragement and useful critiques of this research work.

9. REFERENCES

[1] S. Forrest, P. D'haeseleer, and P. Helam, "An immunological approach to change detection: Algorithms, analysis and implications". In Proceedings of the IEEE Symposium on Security and Privacy, IEEE Computer Society, vol. 110, 1996

[2] C. Warrender, S. Forrest, and B. Pearlmutter, "Detecting intrusions using system calls: Alternate data models," In Proceedings of the IEEE ISRSP. IEEE Computer Society, 1999, pp. 133 – 145.

[3] S. Forrest, S. A. Hofmeyr, A. Somayaji. and T. A. Longstaff, "A sense of self for unix processes," In Proceedings of the IEEE ISRSP, 1996, pp120 – 128.

[4] Warusia Yassin, Nur Izura Udzir1, Zaiton Muda, and Md. Nasir Sulaiman "Anomaly based intrusion detection through K-means clustering and naives bayes classification". Proceedings of the 4th International Conference on Computing and Informatics, ICOCI 2013 28-30 August, 2013 Sarawak, Malaysia.

[5] A. K. Ghosh, and A. Schwartzbard, "A study in using neural networks for anomaly and misuse detection," In Proceedings of the 8th USENIX Security Symposium, 1999.

[6] Shui Yu, Wanlei Zhou, Robin Doss, Weijia Jia,"Traceback of DDoS Attacks Using Entropy Variations" IEEE/ACM Tran. ON Parallel and Distributed Systems" vol. 22, no. 3, March 2011.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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

[13] 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.

[14] Mohammad Khubeb Siddiqui and Shams Naahid "Analysis of KDD CUP 99 Dataset using Clustering based Data Mining",International Journal of Database Theory and Application.

[15] Megha Bandgar, Komal dhurve, Sneha Jadhav,Vick Kayastha, Prof. T.J Parvat, "Intrusion Detection System using Hidden Markov Model (HMM)", IOSR Journal of Computer Eng.(IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278- 8727Volume 10, Issue 3 (Mar. - Apr. 2013), PP 66-70 www.iosrjournals.org.

[16] Huang et. al. Spoken Language Processing. Prentice Hall PTR