A Unified Approach for Real Time Intrusion Detection using Intelligent Data Mining Techniques

Naveen N C ,
Associate Professor,
Dept of ISE,
R V College of Engineering,
Bangalore and Research Scholar, Dept of CSE, SRM University, Chennai

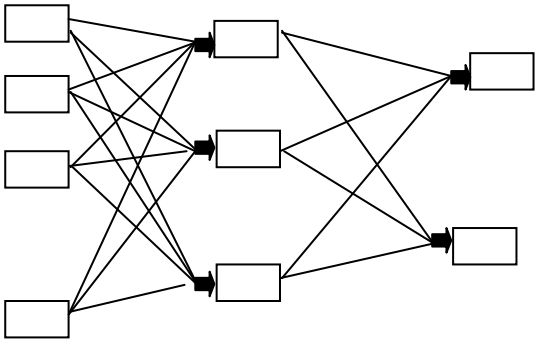
Dr. R. Srinivasan,
Professor, Department of Computer Science and Engineering, R N S I T, Bangalore.

Dr. S. Natarajan,
Professor, Department of Information Science and Engineering, P E S I T, Bangalore

ABSTRACT
In the recent days, there is a rapid increase in the usage of intelligent data mining approaches to predict intrusion in local area networks. In this paper, an approach for Intrusion Detection System (IDS) which embeds an expert system making data mining technique behave intelligently is proposed. Intrusion Detection System (IDS) is considered as a system integrated with intelligent subsystems, which completes the distributed solution procedure on the basis of exchanging large data and information. Any intelligent process self regulates and self-controls itself in the event of intrusion. The system however requires complete information of the intrusion mechanisms and generates appropriate decisions for preventing from further attacks. The combination of methods is intended to give better performance of IDS systems, and make the detection more effective. The result of the evaluation of the new design has produced a better output in terms of efficiency in detection and reduction of false alarm rate from the existing problems. In this paper we present improved architecture along with implementation details. A proper justification for claiming the proposed approach as a better method is also endorsed. The challenging research trends in the field of Data Mining involving Intrusion Detection methods is also discussed at the latter part of the paper.

Keywords - Data Mining, Intrusion Detection System, WEKA, Neural Networks, SLFN

1. INTRODUCTION
The next generation challenges and future directions in building intelligent techniques for the construction of efficient and reliable intrusion detection systems is a real challenge. An intrusion can be defined as "any set of actions that attempt to compromise the integrity, confidentiality or availability of a resource" [1]. Over the years, researchers and designers have used many techniques to design Intrusion Detection Systems [IDS]. But, there have been one or more issues with the existing IDS. Current anomaly detection methods are mainly classified as Statistical Anomaly Detection, Detection Based on Neural Network and Detection Based on Data Mining, etc.
The IDS for the anomaly detection should first learn the characteristics of normal activities and abnormal activities, and then the IDS detect traffic that deviate from normal activities. Anomaly detection tries to determine whether deviation from established normal usage patterns can be flagged as intrusions [6]. Anomaly detection techniques is based on the assumption that misuse or intrusive behavior deviates from normal system procedure [7]. The advantage of anomaly detection is that it can detect attacks that are never seen before. But the disadvantage of anomaly detection is ineffective in detecting insiders' attacks.
In [8], a strategy that effectively combined strategies of Data Mining and Expert Systems were used to design IDS. This technique appeared to be promising, but still with structural and performance problems. Also, combining multiple techniques in designing IDS is a recent research trend and needs further improvement.
Data Mining makes use of algorithms to extract useful information, patterns and trends often previously unknown. Recently researchers have used Data Mining techniques for counter terrorism applications, detect unusual patterns and fraudulent behavior etc. In this research work WEKA an open source tool is used to carry out effective Data Mining and extract useful information. The challenge in using these techniques is to detect and/or prevent attacks and eliminate False Positives and False Negatives as much as possible [10].
Attribute selection in Intrusion Detection using Data Mining algorithm involves the selection of a subset of attributes *d* from a total of *D* original attributes of dataset, based on a given optimization principle [11]. Attribute selection methods search through the subsets of attributes, and try to find the best one among the completing 2^N candidate subsets according to some evaluation function. Many Data Mining algorithms like Decision Tree, Naïve Bayesian Classifier, Neural Network, Genetic Algorithm, and Support Vector Machines etc are useful for classifying Intrusion Detection datasets [3-5].
When Neural Networks is used it is clear that the learning speed of the network is generally slower and a research challenge is to improve performance of learning algorithm. For approximation in a finite training set, Huang and Babri [12] showed that a Single-Hidden Layer Feed Forward Neural Network (SLFN) with at most N hidden nodes and with almost any nonlinear activation function can exactly learn N distinct observations. As suggested one may not adjust the input weights of all the hidden nodes but can randomly be assigned. Feed Forward Neural Networks is one of the most popular research methods which consist of one input layer, one or more hidden layers, and one output layer sending the network output to external environment. In this research we apply this method and the result is encouraging and the learning is fast.



Input Neurons **Hidden Layer** **Output Neurons**

Fig. 1: A Single Layer Feedforward Network (SLFN)

2. RELATED WORK

Different types of neural works exist and three main approaches are used for training the Feed Forward Neural Network

1. Gradient Descent based (e.g. Back Propagation (BP) method): In this the effect of initial weight selection on Feed Forward Networks learning simple functions with the Back Propagation technique is demonstrated, through the use of Monte Carlo techniques. The magnitude of the initial condition vector (in weight space) is a very significant parameter in convergence time variability [13].
2. Standard optimization method based (e.g. Support Vector Machines, SVMs [14], for a specific type of SLFNs, the so-called Support Vector Network). Rosenblatt [15] suggested a learning mechanism where only the weights of the connections from the last hidden layer to the output layer were adjusted.
3. Least-square based (e.g. Radial Basis Function (RBF) network learning [16]), utilizes nonlinear optimization of the first layer parameters which is beneficial only when a minimal network is required to solve a given problem.

As a learning technique, Feed Forward Network has demonstrated well in resolving Regression and Classification problems. Compared to Back Propagation algorithm SLFNs provide a better result if the number of instances is greater than the number of features. Earlier work emphasized that data can be obtained by three ways; by using real traffic, using sanitized traffic and also, using simulated traffic, but IDS are tested mainly on a standard dataset [10]. But in real time fast response to external events within an extremely short time is highly demanded and expected. Therefore, an alternative algorithm to implement real time learning is highly demanded for critical applications with fast changing environments. Even for offline applications, speed is still a need, and a real-time learning algorithm that reduces training time and human effort to nearly zero would always be of considerable value.

4. REAL TIME SYSTEM ARCHITECTURE

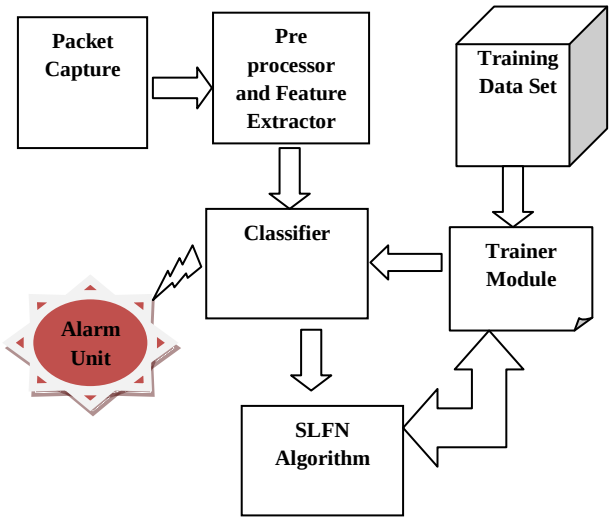


Fig. 2: Real Time System Architecture

Mining data in real time is a big challenge. For capturing the packets in real time we have used JPCAP and WINPCAP tool to intercept the information that is being transmitted. JPCAP provides facilities to:

- Capture raw packets live from the wire.
- Save captured packets to an offline file, and read captured packets from an offline file.
- Automatically identify packet types and generate corresponding Java objects (for Ethernet, IPv4, IPv6, ARP/RARP, TCP, UDP, and ICMPv4 packets).
- Filter the packets according to user-specified rules before dispatching them to the application.
- Send raw packets to the network

JPCAP is based on libpcap/winpcap, that is implemented in C and Java. WINPCAP is the industry-standard tool for link-layer network access in Windows environments: it allows applications to capture and transmit network packets bypassing the protocol stack, and has additional useful features, including kernel-level packet filtering, a network statistics engine and support for remote packet capture.
In the Pre processor and Feature Extractor phase, packets collected are preprocessed and effective attributes are selected for further processing**.** Classification is a very common Data Mining task. Classification maps data into predefined groups or classes. It is often referred to as supervised learning because the classes are determined before examining the data. The classifier module will generate the decision tree that can be used for detection. The performance of an Intrusion Detection model depends on its detection rates (DR) and False Positives (FP) [9]. DR is defined as the number of intrusion instances detected by the system divided by the total number of the intrusion instances present in the dataset. FP is an alarm, which rises for something that is not really an attack. It is preferable for an intrusion detection model to maximize the DR and minimize the FP. ID3 Decision Tree algorithm is used to construct the classifier model.
The proposed SLFN algorithm is used and applied to the decision tree to improve the performance of the learning rate and the new attacks found are updated to the training set accordingly for further processing. WEKA tool is used to analyze the audit data, judges that it is a normal behavior, abnormal behavior or aggressive behavior and responds to the result obtained by the operation behavior and finally reports

the result to the manager in a comprehensible form. The output of IDS alarms the network security officer or automated Intrusion Prevention System (IPS).

3.1 Implementation of SLFN Algorithm [2]

Given a training set $X=\{[x_i,t_i]|x_i \in R^n, t_i \in R^m, i=1,2,\ldots,N\}$ activation function g(x), and hidden node number N

Step 1 : Randomly assign input weight wi and bias bi, i=1,2,…N

Step 2 : Calculate the hidden layer output matrix H

Step 3 : Calculate the output weight β

$\beta = H^{\dagger}T$ where $H^{\dagger}$ is the Moore-Penrose generalized inverse of matrix H and $T = [t_1,\ldots t_N]^T$.

Standard SLFNs with N hidden nodes and activation function g(x) are mathematically modeled as

$$\sum_{i=1}^{N} \beta_i g_i(x_j) = \sum_{i=1}^{N} \beta_i g(w_i \cdot x_j + b_i) = o_j, j = 1..n$$

where $w_i = [w_{i1}, w_{12}, \ldots w_{in}]^T$ is the weight vector connecting the i^{th} hidden node and the input nodes, $\beta_i = [\beta_{i1}, \beta_{i2}, \ldots \beta_{im}]^T$ is the weight vector connecting the i^{th} hidden node and the output nodes and bi is the threshold of the i^{th} hidden node $w_i \cdot x_i$ denotes the inner product of the w_i and x_i.

For finding the Moore-Penrose matrix MATLAB is used. The original matrix H is passed as a parameter from the Java application and the resultant matrix $H^{\dagger}$ is obtained.

The experiment was carried out on a real data stream called "Intrusion Dataset", which is collected from the server in real time using JPCAP and WINPCAP tools as shown in Fig.5. Data that was collected is stored in a file where data contains the details of the network connections, such as protocol type, Source IP, Destination IP, Source Port and Destination Port and number of bytes in the source etc.

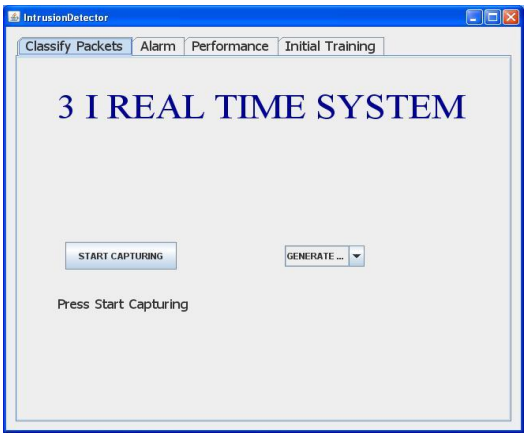


Fig.3 Initial Screen

4. EXPERIMENTAL RESULTS

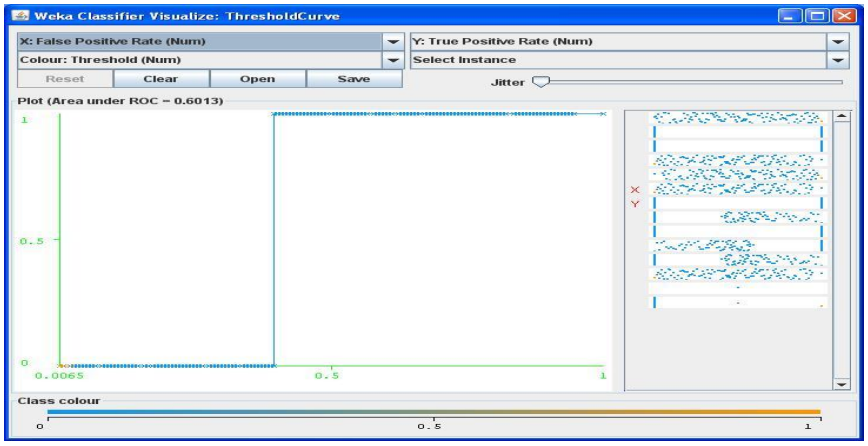


Fig. 4 WEKA tool for analysis

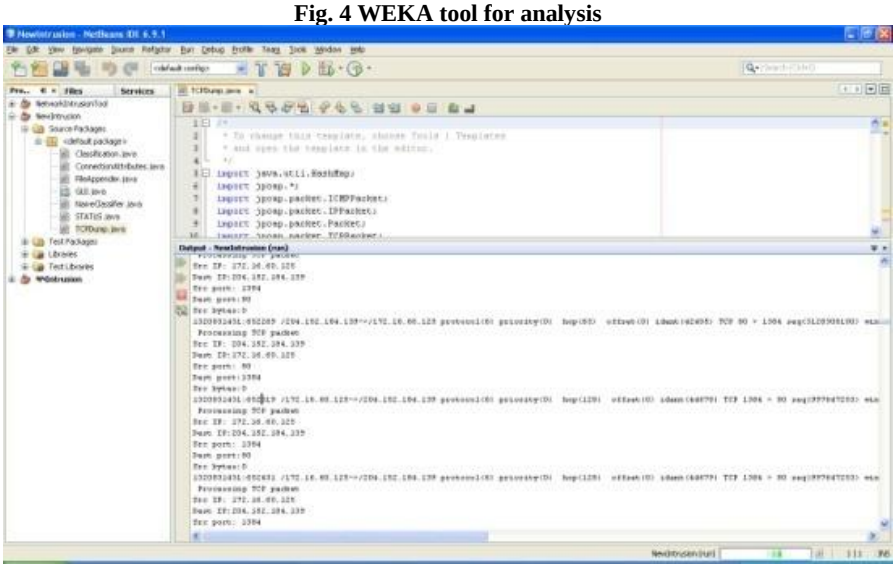


Fig. 5 Packet Capture Screen

Each data sample in the dataset represents attribute value of a class in the network data flow, and each class is labeled either as normal or as an attack with exactly one specific attack type. In total, features can be used in the training dataset and each connection can be categorized into two main classes as shown in Table I.

Table I: Types of Attack Classes

2 Main Attack Classes	Attack Classes
Probing	Ipsweep, Nmap, Portsweep
Denial of Service (DOS)	Neptune, Smurf, Teardrop

This project has two phases namely learning and classifying training data. The data is pre processed and the training data set is updated considering the fields in the packets as shown in Table II.

Table II: Description of packet variables considered

Protocol Type	Service	Source Bytes	Dst Bytes	Count	dst_host_same_src_port_rate	dst_host_srv_diff_host_rate	dst_host_serror_rate	dst_host_srv_serror_rate	dst_host_rerror_rate	Type of Attack
Udp	7	15	0	1	0	0	0	0	0	TEARDROP
Udp	7	100	0	1	0	0	0	0	0	NORMAL
Udp	7	200	0	1	0	0	0	0	0	IPSWEEP
Udp	7	300	0	1	0	0	0	0	0	PORTSWEEP
Udp	7	400	0	1	0	0	0	0	0	NEPTUNE
Udp	7	500	0	1	0	0	0	0	0	SMURF

In the second phase the data is collected online and analyzed to detect intrusion. Table III and Table IV gives the analysis done for the data collected.

Table III: Number of examples considered

Attack Types	Training Examples	Testing Examples
Normal	83453	74563
Denial of Service	12783	10435
Probing	1348	3221
Total Examples	97854	88219

Table IV: Comparison Results

Method	Without using SLFN	Use of SLFN algorithm
Normal	0.08	0.03
Denial of Service	0.06	0.04
Probe	0.21	0.13

5. DISCUSSION AND CONCLUSION

With the advent of new technologies building IDS has become more complex. Designing the IDS for a real time has become even more challenging. Real time learning capability of Neural Networks is the need whenever a new threat is faced, where a new knowledge map has to be built. In this paper a simple and efficient SLFN is used which resulted in a system that can detect attacks faster compared to other methods. As a learning technique, SLFN has demonstrated good potential in resolving Regression and Classification problems. With the use of proposed Huang's algorithm the results are truly encouraging and perform better compared to gradient-based and/or iterative approaches. When the number of pattern classes are large (say, larger than 10), the training time cost is most likely higher. In general, with the use of Soft Computing Paradigms like (Neural Networks, Expert Systems, Agents, Bayesian Networks, Fuzzy Logic, Immune Systems and Genetic Algorithms) IDS may detect intrusions where tedious time-consuming trials of other algorithms can be prevented.

6. REFERENCES

[1]. Huang G-B, Zhu Q-Y, Siew C-K (2004) Extreme learning machine: a new learning scheme of feedforward neural networks. In: Proceedings of international joint conference on neural networks (IJCNN2004), vol 2, Budapest, Hungary, 25–29 July 2004, pp 985–990

[2]. Guang-Bin Huang_, Qin-Yu Zhu, Chee-Kheong Siew, Extreme learning machine: Theory and applications, 2006

[3]. G.-B. Huang, Q.-Y. Zhu, C.-K. Siew, Real-time learning capability of neural networks, Technical Report ICIS/45/2003, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, April 2003

[4]. Li K, Huang G-B, Ge SS (2010) Fast construction of single hidden layer feedforward networks. In: Rozenberg G, Back T, Kok JN (eds) Handbook of natural computing. Springer, Berlin, Mar 2010

[5]. Dewan Md. Farid, Jerome Darmont, Nouria Harbi, Nguyen Huu Hoa, and Mohammad Zahidur Rahman, Adaptive Network Intrusion Detection Learning: Attribute Selection and Classification, World Academy of Science, Engineering and Technology 60 2009

[6]. Dewan Md. Farid, Nouria Harbi, Emna Bahri, Mohammad Zahidur Rahman, Chowdhury Mofizur Rahman, Attacks Classification in Adaptive Intrusion Detection using Decision Tree, World Academy of Science, Engineering and Technology 63 2010

[7]. Ahmad Ghodselahi, A Hybrid Support Vector Machine Ensemble Model for Credit Scoring, International Journal of Computer Applications (0975 – 8887) Volume 17– No.5, March 2011

[8]. A.S. Sodiya, H.O.D. Longe and A.T. Akinwale, A new two-tiered strategy to intrusion detection. Information Management & Computer Security, 12 1 (2004), pp. 27–44.

[9]. LI ZHUOWEI, A Framework For Systematic Design, Analysis And Evaluation Of Intrusion Detection Systems, A thesis submitted to the Nanyang Technological University, 2007

[10]. Iftikhar Ahmad, Azween Abdullah and Abdullah Alghamdi, Towards the selection of best neural network system for intrusion detection, International Journal of the Physical Sciences Vol. 5(12), pp. 1830-1839, 4 October, 2010

[11].MuamerN. Mohammad, Norrozila Sulaiman, Osama Abdulkarim Muhsin, A novel intrusion detection system by using intelligent data mining in weka environment,WCIT 2010

[12]. G.-B. Huang, H.A. Babri, Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions, IEEE Trans. Neural Networks 9 (1) (1998) 224–229.

[13]. John F. Kolen Jordan B. Pollack, Back Propagation is Sensitive to Initial Conditions, Laboratory for Artificial Intelligence Research

[14]. Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, A Practical Guide to Support Vector Classification, 2010

[15]. Rosenblatt F (1962) Principles of neurodynamics: Perceptrons and the theory of brain mechanisms. Spartan Books, New York

[16]. Lowe D (1989) Adaptive radial basis function nonlinearities and the problem of generalization. In: Proceedings of first IEEE international conference on artificial neural networks, pp 171–175