ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

Light Weight Intrusion Detection System with Wrapper Approach and Optimized Feature Selection

Amol A.Dhiwar, Dr. G.K.Patnaik

Assistant Professor, Dept. of Computer Engg, SNDCOE&RC, Savitribai Phule Pune University, Yeola, Maharashtra, India

Professor & Head of Dept. of Computer Engg., SSBT's COET, North Maharashtra University, Bambhori, Jalgaon, India

ABSTRACT: Features based Intrusion Detection Systems (IDS), mostly used for Denial of Service (DOS) attacks, have low response in terms of intrusion detection because of missing Local Area Network Denial (LAND) and duration features. Hence, precise security of a system is not assured without considering LAND and duration features. In order to minimize DOS attacks and to make the system more secured, it warrants additional features. All the features are having their values that indicate the presence or absence ofan intrusion. An existing genetic algorithm has considered 16 features for intrusion detection but, still some DOS &Remote to Local (R2L) attacks are not covered in it. These attacks are depends on duration &LAND features of dataset. These two features are focused and extracted using genetic algorithm so that detection response of IDS's is improved.

KEYWORDS: Intrusion detection, datasets, Genetic Algorithm, feature extraction, information gain, networkattacks, network anomalies.

I. INTRODUCTION

Intrusion detection is an important method that monitors network traffic and finds network intrusions such as faulty network behaviors, illegal network access and hostile attacks to computer systems. An Intrusion detection system (IDS) detects intrusions and reports it accurately to the proper authority [1].Conventional intrusion prevention strategies, such as firewalls, access control schemes or encryption methods, have failed to prove themselves to effectively protect networks and systems from increasingly sophisticated attacks and malwares.Generally intrusion detection systems are divided into two variations, misuse detection and anomaly detection. Misuse detection depends on the prior representation of specific patterns for intrusions, allowing any matches to it in current activity to be reported. The anomaly based system builds a model of the normal behavior of the system and then looks for anomalous activity such as activities that do not confirm to the established model.The anomaly detection systems are adaptive in nature; it can deal with new attack butit unable to identify all the type of attack.Many researchers have proposed and implemented various models for IDS but it often generates toomany false alerts due to their simplistic analysis.

The Intrusion Detection Systems (IDS) turn out to be the proper solution to an intrusion issues and have become a crucial component of any security infrastructure to detect all possible threats before it induce wide damage. The design and construction of IDS is subjected to many concerns including information collection, information pre-processing, intrusion identification, reporting and response. All the components of IDS compare the audit data with the detection paradigms that model the patterns of intrusive behavior, so that both successful and unsuccessful intrusion attempts are identified. Intelligent IDS is a dynamic defensive system that is capable of adapting to dynamically changing traffic pattern and is present throughout the network, so it helps to detect all types of attacks by considering correlated feature.The existing system uses new approach for light weight intrusion detection that works on decision tree & wrapper approach. In wrapper based approach, after classifier construction, the analysis of specificity and sensitivity is generated. The specificity and sensitivity sense the relationship between the detection rate and error rate, and find the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

optimal features that minimize the wrong alarm rates for IDS. The selected significant features are then used for restructuring and improving the classifier. The performance of the IDS is optimized when the total errors are minimized. Hence the objective of the correct feature selection is to produce precise intrusion detection and minimize the wrong alarm rates. In order to prevent Denial of Service (DOS) attacks, proper feature set need to be extracted from the dataset.

In existing IDS, only 16 features are extracted from dataset with the help of Genetic Algorithm. The genetic algorithm is best optimized one among the rest of algorithmgiven in Table1. But due the features opted in existing system causes a chance to enter an intrusion of DOS type. To design good IDS for DOS type of attacks, it needs to consider LAND(Local Area Network Denial) and duration features. These features are extracted from the dataset using genetic algorithm.

The rest of paper is organized as follows: In Section 2,Related Work with emphasis on variousmethods and frameworks used for intrusion detection is discussed. The Problem Statement is described in Section 3. In Section 4 Proposed Solution is described.

II. RELATED WORK

Siva S. Sivatha Sindhu et.al, in [1], proposed a neural network based IDS for detecting internet-based attacks on a computer network. Neural networks are used to classify and predict current and future attacks. Feed-forward neural network along with the back propagation training algorithm was employed here to detect intrusion. Randomly selected data points form KDD Cup (1999) is used to train and test the classifier. The process of learning the behavior of a given program by using evolutionary neural network based on system call audit data is proposed here.

Kapil Kumar Gupta et.al, in [2], proposed a host based IDS using combination of K-Means clustering and ID3 decision tree learning algorithms for unsupervised classification of abnormal and normal activities I the network. The K-Means clustering algorithm is first applied to the normal training data and it is partition into K clusters using Euclidean distance measure. Decision tree is constructed on each cluster using ID3 algorithm. Abnormality scores value from the K-Means clustering algorithm and decisions rules from ID3 are extracted. Finally anomaly score value is gained using a special algorithm which combines the output of the two algorithms.

Dr. Saurabh Mukherjeea, Neelam Sharma, in[3], proposed FVBRM model for feature selection and make its comparison with three feature selectors CFS, IG and GR. Their experimental result illustrates feature subset identified by CFS has improved Naïve Bayes classification accuracy when compared to IG and GR. Although GR is an extended of IG algorithm, but in analysis, authors have used both the techniques for feature selection and IG performs better than GR. FVBRM method shows much more improvement on classification accuracy with compared to CFS but takes more time. Future work includes customize of FVBRM feature selection method to improve the results for intrusion particularly for U2R attacks with reduced complexity and overheads. Authors have selected 24 features & it gives detection rate of 97.78 %.

Mohammad Sazzadul Hoque, in [4], proposed an Intrusion Detection System by applying genetic algorithm to efficiently detect various types of intrusions. To design and count the performance of their system author have used the standard KDD99 benchmark dataset and obtained appropriate detection rate. To take the calculation of the fitness of a chromosome author has used the standard deviation equation with distance. If their used equation would better or heuristic in this detection process, their detection rate and process would have improved to a great extent, especially the false positive rate would be surely much lower. The problem with this system is that it focuses only on some DOS types of attacks with detection rate of 90.25 percent.

Mostaque Md. Morshedur Hassan, in [5], proposed a method of applying genetic algorithms with fuzzy logic is presented for network intrusion detection system to efficiently detect various types of intrusions. To design and measure the performance of the system he carried out a number of experiments using the standard KDD Cup 99 benchmark dataset and obtained appropriate detection rate. To take analysis of the fitness of a chromosome he used the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

fuzzy confusion matrix where the fuzzy membership value and fuzzy membership function for the complement of a fuzzy set are two different concepts.

Vivek K. Kshirsagar, in [6], proposed various IDS models. Various techniques can be used to implement IDS. In the paper, author has mainly concentrated on signature based i.e. misuse detection system. Anomaly based IDS requires to identify new anomalies based on rules stored in IDS while misuse IDS can find only those attacks whose matching rules are already stored in rule set.

Emma Ireland, in [7], proposed the fuzzy genetic algorithm. It had a higher detection rate than the traditional genetic algorithm that was used in [4]. The genetic algorithm in [4] had a high detection rate for denial of service attacks. When compared with the winning entry of the KDD99 classifier Learning Contest, then it was shown to have a better detection rate for denial of service and user to root attacks. Author showed that the use of genetic algorithms and fuzzy genetic algorithms in intrusion detection are effective ways of detecting attacks.

Snehal A. Mulay, in [8], proposed the novel multiclass SVM algorithm for implementation of Intrusion Detection System. The integration of Decision tree model and SVM model gives better results than any individual methods. But the final results for the system proposed was not available, but authors thought that multi class pattern recognition problems can be solved using the tree structured binary SVMs and the resulting intrusion detection system could be faster than other methods.

Latifur Khan, in [9], proposed reduction techniques using clustering analysis to approximate support vectors in order to speed up the training process of SVM. An author has proposed a method, namely, Clustering Trees based on SVM (CTSVM), to reduce the training set and approximate support vectors. Clustering analysis was used to generate support vectors to improve the accuracy of the classifier.

Amrita Anand and Brajesh Patel, in [10], proposed a method that determines network traffic is potential threat to a network or not, so IDS have a method for differentiating whether it is malicious or not. Hence, this research has introduced a new methodology to identify a fast attack intrusion using time based detection of attack. The method used to identifies anomalies based on the number of connection made in a second. For remaining validation, the methodology is then implemented on a different set of real network traffic. In view of the fact that this research only concentrate on the TCP connection, but the researcher are planned to investigate other protocol and other flag to recognize the fast attack intrusion activity.

Mya Thidar Myo Win, and Kyaw Thet Khaing, in [11], proposed the comparison of the result of detection of attacks with selected features and all features. First, feature relevance is performed by analyzing the nature of selected attack. It analyses the involvement of each feature to classification and a subset of features are selected as relevant features. After that, Random Forest, Naive Bayes and k-nearest neighbor are applied on classification.

The various combinations of features selection algorithms are tried and it is found that the detection rate of each individual algorithm is varied in proportion with the number of features selected. The detection rate is percentage amount of value that notifies an accuracy of IDSs. The Table1.shows details of such feature selection algorithms

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

Table 1. Survey of Different Feature Selection Algorithms

Sr No.	Features \ Algorithms	BestFirst + ConsistencySubsetEval	GeneticSearch + CfsSubset Eval	GeneticSearch + ConsistencySubsetEval	GreedyStepwise + CfsSubsetEval	Ranker + ChiSquaredAttributeEva	RankSearch + CfsSubsetEval	RankSearch + ConsistencySubsetEval	Genetic Algorithm
01	Duration	√	√	√		√		√	
02	Protocol_type		√			√	√	√	√
03	Service	√	√	√		√	√	√	√
04	Flag		√		√	√	√	√	√
05	Src_byte	√	√	√	√	√	√	√	√
06	Dst_byte	√	√			√	√	√	√
07	Land					√	√	√	
08	wrong_fragmwnt		√	√	√	√	√	√	√
09	Urgent								
10	Hot			√	√	√	√	√	√
11	Num_failed_login		√	√		√	√	√	
12	Logged_in		√			√	√	√	√
13	Num_compromised					√	√	√	
14	Root_shell		√	√		√		√	
15	Su_attempted								
16	Num_root			√		√			
17	Num_file_creation					√		√	
18	Num_sell								
19	Num_access_file				√	√			
20	Num_outbound_cmds								
21	Is_hot_login			√					
22	Is_guest_login					√		√	
23	Count	√				√		√	
24	Serror_rate			√		√	√	√	√
25	Rerror_rate					√	√	√	
26	Same_srv_rate		√	√		√	√	√	√
27	Diff_srv_rate		√		√	√	√	√	
28	Srv_count	√		√		√		√	√
29	Srv_serror_rate		√	√		√	√	√	
30	srv_rerror_rate		√	√		√		√	
31	Srv_diff_host_rate		√	√		√		√	
32	Dst_host_count					√		√	
33	Dst_host_srv_count	√		√		√		√	
34	Dst_host_same_srv_rate					√	√	√	
35	Dst_host_diff_srv_rate				√	√	√	√	√
36	Dst_host_same_src_port_rate	√		√	√	√	√	√	√
37	Dst_host_srv_diff_host_rate	√	√	√	√	√	√	√	√
38	Dst_host_serror_rate		√			√	√		
39	Dst_host_srv_serror_rate	√	√			√	√	√	√
40	Dst_host_rerror_rate			√		√			√
41	Dst_host_srv_rerror_rate			√		√		√	
Detection Rate		97.01	98.16	97.86	97.97	98.13	98.4	98.15	98.38

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

III. PROBLEM STATEMENT

An attacks are, illegal activity done by intruders,intrusion that affect the security of system directly. The DOS (Denial of Service) attacks keep the system busy so that others unable to get access of it on time. The attacks are identified using feature set from dataset. The dataset consist of possible training dataset related to the 41 different numbers of features. LAND (Local Area Network Denial) is an attack where intruder keeps local network too busy hence network traffic goes into loop. LAND attack is identified using LAND feature. The duration is another feature that helps to identify DOS attack. For e.g. if duration = 0:0:1, protocol = finger, source_port = 18989, destination_port = 79, source_ip = 99.19.99.19, destination_ip = 192.168.254.10 then it is Neptune attack [6]. An existing light weight intrusion detection system with wrapper approach considered 16 features in association with Genetic Algorithm (GA) for feature selection. The objective of proposed system is to develop a light weight intrusion detection system (IDS), targeting at detecting anomalies in the network. The anomalies are the inaccuracy in the data which is identified using some system.
The algorithm (Genetic Algorithm) used for feature selection is optimized one. The features that have selected in existing system to train the dataset from dataset produced the detection rate of about 98.38 %. But the features like LAND and duration are not considered in existing system that helps to detect DOS type of attacks. Hence using existing solution still there are chances to enter DOS type of attacks. So that to identify the DOS type of attacks with more precise detection, it is required to consider the LAND and duration features with existing algorithm. Since the features LAND and Duration are the most relevant features of LAND attack.

IV. PROPOSED SOLUTION

The Genetic algorithm, used in an existing intrusion detection system is best one, has great optimization value. So in proposed work, genetic algorithm is chosen for feature selection.An additional featuresLAND and Duration are taken into consideration that help to detect DOS attacks so that the performance of system improves. In resultant system, the possibility of the DOS types of attacks to enter into system getsreduced.

Benefits of using genetic algorithm for intrusion detection are: I). Genetic algorithms are intrinsically parallel.Because of multiple offspring, it explores the solution space in multiple directions at once. II) Parallelism allows genetic algorithm to implicitly evaluate many schemas at once. It suited to solving problems where space of potential solution is truly huge. III) Genetic algorithm based systems can be re-trained easily. It improves GA's possibility to add new rules and evolve intrusion detection system.
The proposed system is subdivided into four major steps as given below

Step I: Preprocessing of network traffic pattern (removal of duplicate data)
The problem associated with the dataset is that it contains some amount of duplicate records. The effect of occurrence of duplicate records causes the learning algorithm to be biased towards frequent records and unbiased towards rare records. As the percentage of records for R2L (Remote to Local) class is very less in original dataset, the learning algorithm is unbiased towards R2L records. These duplicate records are removed in order to improve the detection accuracy.

Step II: Feature extraction (Use of GA)
The feature extraction isdone by genetic algorithm. While selecting the features for processing, it should be minimum numbers as possible, because selection of good features is very important. The feature selection depends on two important factors, encoding technique used and fitness function. The value of fitness function is set in such a way that our most concerned features get selected for processing.

Step III: Post-processing (normalization)
Normalizationis introduced to make the approach more flexible and allow for different analysis or classification approaches. Before proceeding to evaluate the performance of the classifier, the discrimination capability of the proposed features analyzed.

Step IV - Classification of traffic patterns (Use of classifier)
In proposed system, Bayesiansmethod is used as a classifier. Initially, a network structure is defined with a fixed number of data inputs, hidden nodes and outputs.Figure1.Shows the functional block diagram of a proposed system.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

Figure1.clearly indicates the flow of proposed system. The proposed system takes dataset as an input and redundant data is removed in order to get correct input. The process of eliminating the duplicate records is also called as preprocessing. The processed data is now apply to genetic algorithm, then use of GA to extract the features from the input data. Finally the extracted features are test on the test data so that the appropriate result is achieved that is to identification of whether the data is having the intrusion or not.Genetic Algorithm works on an individual called chromosome [11] and evolves the group of chromosomes to a population of quality individuals. Each chromosome represents a technique to solve the problem. A fitness function used there for each rule that is a measure of each rules implementation. The evolution of population starts from an initial population of selected chromosomes which gradually improve the fitness value.

The three genetic operators selection, crossover, mutation are applied to each individual during the generation process. A group of suitable chromosomes are selected using a fitness function initially eliminating the other individuals. The process continuous by selecting a number of individuals and making pairs each other. The chromosome pair generates one off-string which exchanges their genes around selected cross points. Finally, some individuals are identified and the mutation operations are applied on it. The sub attacks are recognized with respect to the fitness criteria by selecting the best-fit chromosomes capable of detecting the attacks from every population.

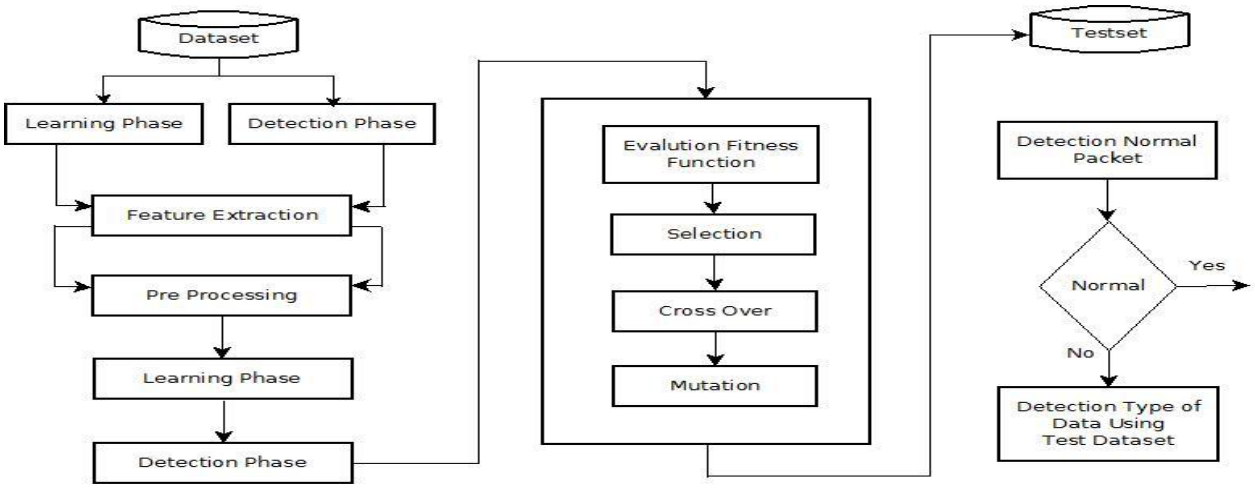

Figure1. Block Diagram of Proposed System

Proposed Algorithm:

Our system is divided into two main phases: the precalculation phase and the detection phase. First algorithm depicts major steps in precalculation phase, where a set of chromosome iscreated using training data. Thesechromosome set is then used in the next phase for the purposeof comparison.
The Major steps in pre calculation is as follows

1. Range = 0.125
2. for each training data
3. If it has neighboring chromosome within Range
4. Merge it with the nearest chromosome
5. Else
6. Create new chromosome with it
7. End if
8. End for

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

The major steps of detection phase, where a population is being created for a testdata and going through some evaluation processes (selection, crossover, mutation) the type ofthe test data is predicted. The precalculated set of chromosome is used in this phase to find outfitness of each chromosome of the population.

1. Initialize the population
2. CrossoverRate = 0.13, MutationRate = 0.45
3. While number of generation is not reached
4. For each chromosome in the population
5. For each precalculated chromosome
6. Find fitness
7. End for
8. Assign optimal fitness as the fitness of that chromosome
9. End for
10. Remove some chromosomes with worse fitness
11. Apply crossover to the selected pair of chromosomes of the population
12. Apply mutation to each chromosome of the population
13. End while

V. RESULTS

As the proposed system is in progress, the results are not come yet. Hence results section is not discussed in this work.

REFERENCES

1. Siva S. Savtha Sindhu, S.Geetha and A.Kanan, 'Decision tree based light weight intrusion detection using a wrapper approach', Elsevier's Expert System With Application, Vol.39, pp. 129-141,2012.
2. Kapil Kumar Gupta, Baikunth Nath and Ramamohonarao Kotagiri, 'Layered Approach Using Conditional Random Fields For Intrusion Detection', IEEE Transactions On Dependable and Secure Computing, Vol.7, Issue 1, pp. 35-49,2010
3. Dr. Saurabh Mukherjee and Nilam Sharma, 'Intrusion Detection Using Naive Bayes ClassifierWith Feature Reduction', Elsevier's Procedia Technology, Vol.4,pp.119-128,2012.
4. Mohammaed Sazzadul Hoque, Md Abdul Mukit and Md. Abu Naser Bikas, 'An Implementation Of Intrusion Detection System Using Genetic Algorithm', International Journals Of Network Security and Its Application,Vol.4,No.2,pp.109-120,2012.
5. Mostaque Md. Morshedur Hassan, 'Network Intrusion Detection System Using Genetic Algorithm and Fuzzy Logic', International Journals of Innovative Research in Computer and Communication Engineering, Vol.1,Issue 7,pp.1435-1445,2013
6. Vivek K. Kshirsagar, Sonali M. Tidke and Swati Vishnu, 'Intrusion Detection System using Genetic Algorithm and Data Mining : An Overview', International Journals Of Computer Science and Informatics, Vol.1,Issue 4,pp.91-95,2012
7. Emma Ireland, 'Intrusion Detection Using Genetic Algorithm and Fuzzy Logic', UMM CSci Senior Seminar Conference, 2013.
8. Snehal A.Muley, P.R.Devale and G.V.Garje,'Intrusion Detection Sytem Using Support Vector Machine and Decision Tree', International Journals Of Computer Applications, Vol 3, Issue 3,pp.40-43,2010
9. Latifur Khan ,Mamoun Awad and Bhavani Thuraisingham, 'A New Intrusion Detection System Using Support Vector Machine and Hierarchical Clustering', The VLDB Journal, Vol.16,pp.507-521,2007.
10. Amrita Anand and Brajesh patel , 'An Overview on Intrusion Detection System and Types of Attacks It Can Detect Considering
11. Different Protocols', International Journal of Advanced Research in Computer Science and Software Engineering, Vol.2,pp.94-
12. 98,2012.
13. Mya Thidar Myo Win, and Kyaw Thet Khaing, 'Detection and Classification of Attacks in Unauthorized Access', International Conference on Advanced engineering and Technology,pp.345-349,2014