

Group Leader Teaching Learning Based Intrusion Detection System For Dos Attacks

S. JayaPrakash¹ M.Aramudhan²

¹Research Scholar, Bharathiyar University, Coimbatore, India

*²Associate Professor, Department of Information and Technology, PKIET, Karaikal.
E-mail:Jayaprakash_nagai@yahoo.co.in, aranagai@yahoo.co.in*

Abstract

Developing a novel Intrusion Detection computational technique is a challenging research problem in the areas of security that monitor, identify and discards the abnormal attacks or access to the networks. Network based IDSs examine the activities in the network traffic and raise an alarm whenever a suspicious activity is detected. Existing IDSs are provided with high number of false positives due to imperfect attack recognition, emerging of new attacks and poor features extraction from training dataset. In the proposed IDS, Teaching Learning Based Optimization (TLBO) and Group Leader Teaching Learning Based optimization(GLTLBO) algorithm is applied in the classifier and Genetic Algorithm (GA) is applied for extracting features of the training dataset. In this proposed model, teachers are considered as an attack, whereas learners are assumed as normal traffic. Discussions that are carried out about the attack patterns among the learners helps to observe the extended anomaly behaviors of the attack traffic. This improves the effectiveness of both these algorithms of the IDS in terms of identifying the new attacks and reducing false positives. Simulation results prove that the proposed IDS is better than the existing evolutionary based IDS.

Keywords: GLTLBO, TLBO, GA, IDS, Feature Extraction, Classifier

Introduction

Intrusion is defined as the set of actions that attempt to compromise the confidentiality, integrity and availability of the resources [7]. It can be detected based on deviations from a user's or historical pattern of behavior. Developing a high performance IDS is very complicated research challenges for the researchers and classification accuracy mainly depends on the feature extraction from the dataset. Computational intelligence techniques are used to classify the inward network traffics either as normal or malicious. A lot of computational intelligence approaches have

been proposed by the researchers such as artificial neural network, fuzzy sets, evolutionary computation, expert system approach, rule based approach, artificial immune systems etc. [1,2,13,11,15]. A system that is based on computationally intelligent when it deals with statistical data has error rates that estimated human performance.

IDS are examined with large amount of data that causes slow training, testing process and low detection rate. So, feature extraction is the challenging task in developing IDS [8]. Generally, IDS consists of three phases such as data preprocessing, features extraction and classifier. The tasks that are carried out in preprocessing phases includes (i) identifies the attributes and their value (ii) Convert categorical to numerical data (iii) Data normalization and (iv) compute redundancy check and handle about null value. The feature extraction process is a preprocessing step when constructing IDS, used to reduce the dimensionality of the dataset by removing irrelevant, redundant features and improve the prediction accuracy of the classifier using selected features from the dataset. Classifier module finds the conditions of the traffics that are either legitimate or malicious attack. Classifier is faced with a problem when it has to generate rules with many attributes or features. Obviously, the time required to generate rules is proportional to the number of features. In addition, irrelevant and redundant features can reduce both the predictive accuracy and comprehensibility of the induced rule and degrade the classifier speed. Thus, selecting the most relevant features is necessary, this strategy is implemented to simplify the rules and reduce its computational time while retaining the quality of classifiers, as it represents the original features set.

Teaching Learning Based Optimization (TLBO) algorithm is proposed on the effect of a teacher on learners. There are two phases in this algorithm, the first phase called as a teacher, describes learning from the teacher and second phase called as learners, describes the learning by the interaction between learners. Generally, in learning phase, the normal profile of the traffic is calculated where as a classifier is used to improve detection of new attacks in the detection phase. Existing IDS algorithms are suffering from a high number of false positives and less detection accuracy due to imperfect attack recognition, emerging of new attacks and poor features extraction from training dataset [2].

The rest of this paper is organized as follows. In section 2 discusses the related work based on evolutionary algorithms, section 3 describes the teaching learning based optimization algorithm and Genetic algorithm section 4 describes the implementation of the proposed algorithm using KDDCUP 99 data set. In the last section, discusses the conclusion and future work.

Related Work

Genetic Algorithm has been used for network intrusion detection in different ways. Some of the approaches directly use GA for to obtain the classification rules [11,12, 13, 6, 7], while others use different AI methods for possession of the rules, where GA are used to select appropriate features or to determine the optimal parameters of some functions [8, 9,10]. Li [7] represent a technique using GA to detect abnormal network

intrusion. This approach, focus on obtaining classification rules for quantitative and distinct features of network data. Apart from the implementation of rule generation for IDS is given, but results of experiments do not exist. Bridges [9] this method combines both fuzzy data mining techniques and Genetic Algorithm for detection of network anomalies and misuses. The most features are not predicted properly in various existing Genetic Algorithm based IDS. This method uses Genetic Algorithm to recognize the optimal parameters of the fuzzy functions for selecting the features of the relevant network.

Lu [10], In this method classification rules are generated by Genetic Programming. Detection or Classification of intrusions on the network with the help of the fitness function is fine tuned by this method. The time required to train the system with huge data creates a Genetic Programming implementation difficult. Crosbie [11] Different agent techniques and Genetic Programming can be used to detect network intrusions. The set of agents that determine the network behaviors can be found out by an agent who monitors, one parameter of the network audit data and Genetic Programming. Many small autonomous agents can be used in this method which is an advantage and the communication among the agents is a drawback. This system identifies the attacks using ruleset by proceeding Genetic Algorithm, then exploit rules only for R2L and DOS type of attacks. Between these two attacks, one of each is selected. The common performance of the system is less than 60%.

The authors of [11] detected the SYN flooding attacks at leaf routers that connect end hosts to the Internet, which utilizes the normalized difference between the number of SYN packets and the number of FIN (RST) packets in a time interval. If the rate of SYN packets is much higher than that of FIN (RST) packets by a non-parametric cumulative sum algorithm, the router recognizes that some attacking traffic is mixed into the current traffic. Similar work is presented in [2], where the fast and effective method for detecting SYN flood attacks is given. Moreover, a linear prediction analysis is proposed as a new paradigm for DoS SYN flood attack detection. The proposed mechanism makes use of the exponential back off the property of TCP used during timeouts. By modeling the difference of SYN and SYN&ACK packets, it is shown that this approach is able to detect an attack within short delays. Again, this method is used at leaf routers and firewalls to detect the attack without the need of maintaining any state. However, considering the fact that the sources of the attack can be distributed in different networks, there is a lack of analysis for the traffic near the sources and also the detection of the source of SYN flooding attack in TCP based low intensity attacks are missing.

Proposed Work

The general architecture of proposed IDS is shown in Figure-1. The architecture contains two phases (i) Training phase (ii) Testing phase. In the training phase, the KDDCUP 99 datasets were used, Data pre-processing, feature selection using genetic algorithm and classifier using TLBO and GLTLBO are implemented in training stage and DoS patterns are identified. Second stage is testing phase, the captured traffic is evaluated as in training stage, pattern identified, matched with database and a decision

to be taken. New pattern were identified by analyzing the behavior of the traffic, if it was against the legitimate traffic, the pattern was captured and updated in the database.

Data Preprocessing

Data Preprocessing is an important step in the machine learning, computing that eliminates out of range values, impossible data combinations, missing values, etc. Generally, data preprocessing includes learning, normalization, transformation, feature extraction and selection. The output of the data preprocessing is the final training set that extracts knowledge of the testing phase. The following four steps used for data preprocessing:

1. Identifying features and its related values.
2. Converting original feature data value into numeric data value.
3. Applying data normalization based on min-max normalization.
4. Perform similarity check and remove null values.

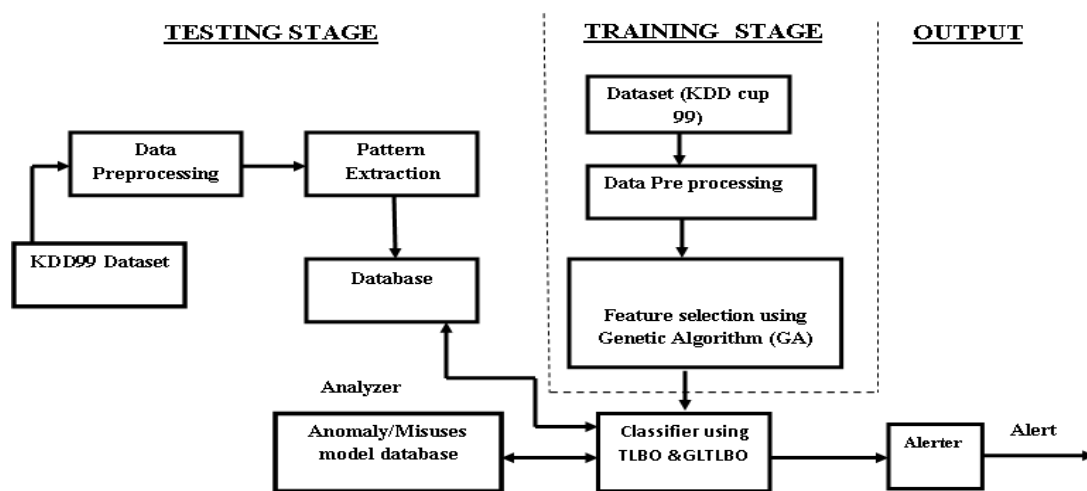


Figure 1: General Implementation Architecture of Intrusion Detection System

Feature Selection Based on Genetic Algorithm

Accuracy of the classifier depends on the selection of optimum feature subsets. Feature selection method [5] mainly used for selecting a subset of features from the original dataset. There are two feature selection methods are already proposed, namely filter and wrapper methods. Filter method was mainly based on general characteristics of data features without involving machine language. These features are ranked based on certain criteria, where features with highest rank values are selected as optimal. The main advantages of filter method are low computational cost without involving any machine language algorithm for feature selection. Frequently used filter method are information gain method. Wrapper method is mainly used for feature subset selection from the dataset based on the objective function and analysis of the performance of feature subset.

In this paper, Genetic Algorithm is used to select an optimal feature subset from the datasets. GA reduces the KDD cup 99 features from 41 attribute to 6 attribute features that are related to the characteristics of DoS attack, which reduces 85% of feature space. The six attributes are protocol_type, src_bytes, dst_bytes, count (No of connection to the same host), srv_count (No of connection requesting same service), serror_rate. KDD'99 dataset contains a huge number of redundant records. 10% portions of the full dataset contain only two types of DoS attacks (Smurf and Neptune). These two types constitute over 71% of the testing dataset which completely affects the evaluation. Brief Steps about Genetic algorithm that selected features from the dataset is shown as algorithm below

- Step-1: Initialize a population of Pre-processed data.
- Step-2: Calculate objective function for each individual.
- Step-3: Selection of individual solution.
- Step-4: Perform matching of a pair of individuals.
- Step-5: Perform mutation operations.
- Step-6: Calculate objective function for newly created population.
- Step-7: If it satisfies stop the operation.
- Step-8: Otherwise, repeat step-3.
- Step-9: Return the best features from KDD 99 dataset that reflects the properties of DoS.

Teaching-Learning-based optimization

Teaching Learning Based Optimization (TLBO) algorithm is proposed by R.V. Rao et al. [4,9] that is based on the effect of a teacher on learners. This algorithm was developed based on the inspiration of teaching learning process and mimics the teaching learning ability of the teacher and learners in the classroom. There are two phases in this algorithm, the first phase called as teacher where learners learn through the teacher. During this phase, a teacher tries to increase the mean result of the classroom. The second phase called as learners where learners increase their knowledge by interaction among themselves. A learner interacts randomly with the learners for enhancing the individual knowledge. A learners learn new things if the other learner has more knowledge than any other learners in the classroom.

In TLBO, a class of learner is considered as the attack and different subjects offered to learners are considered as the patterns of attack traffics. The number of attacks and patterns are denoted as n_a and n . n_a is the number of attacks where n is the pattern of different attack traffics. Let $x_i = (x_{i1}, x_{i2}, \dots, x_{in})^T$, $i=1, 2, 3, \dots, n_a$, be the i^{th} attack where n is the number of attacks considered, let m_j be the mean result of the particular attack, j is the name of the specific attack. The best solution of the attack as $x_{\text{best}} = (x_{\text{best}1}, x_{\text{best}2}, \dots, x_{\text{best}n})^T$ is the exact pattern of the attacks captured from the outcome x_{best} is identified as the pattern for the specific attack.

Group Leader Teacher – Learning Based Optimization

Generally, studies of learners are not only following a teacher, but, also discuss with peers, specially learn from skilled peers. Based on this observation, Class Representative (CR) and Group Leader (GL) are selected to guide the learning

process in learner phase. In learner phase, each learner is first assigned to groups and each learner can be a member in more than one group. Each group has one leader and all leaders are under CR. In this proposed work, teachers are considered as attacks, learners identify the patterns that have deviated from the normal. Each learner group has assigned to one type of DoS attack. Each GL identifies the deviated patterns in the group and intimate to the CR. Each group has selected the most deviated traffic patterns

$$Y_i = \text{best of } \{ x_{i-b} \dots x_{i-1}, x_i, x_{i+1} \dots x_{i+b} \}$$

Similarly, if learner p performs better than q, the updating rule of learner is

$$x''_{i-j}(t) = x'_{i,j}(t) + r.(x'_{p,j}(t) - x'_{q,j}(t))$$

Otherwise

$$x''_{i-j}(t) = x'_{i,j}(t) + r.(x'_{q,j}(t) - x'_{p,j}(t))$$

In the above, each type of attack is assigned to one group to detect extended anomalies and intimate the CR. CR analyses further and find any anomalies, it is updated in the rule structure. A decision tree data structure was used for the implementation. The algorithm of GLTLBO is given as below

Step-1: Initializing the problem and algorithm parameters.

Step-2: Establishing the initial population learners.

Step-3: Compute the objective functions.

Step-4: Compare the mean of the population.

Step-5: Determine the best solutions (Teacher).

Step-6: Modify solutions based on the teacher knowledge according to the teacher phase.

Step-7: Update the solutions according to the learner phase and step-3.

Step-8: Go to step-4 until the iteration number arrives at the maximum iteration numbers.

Simulations and Result Discussions

The proposed computational intelligence based Intrusion Detection System was implemented in Mat Lab. During the evaluation, 10 percent labeled data of KDDCUP 99 was used for training the proposed IDS. This dataset contains three types of traffics and six types of DoS attacks about four gigabytes and each traffic record has 41 features names that values facilitate to identify the type category either normal or attack. It contains a total of 24 attack types that fall into four major categories such as Denial of Service (DoS), probe, User to Root (U2R), Remote to User (R2L). DoS attacks are difficult to deal with because they are very easy to launch, difficult to track and also it is not easy to refuse the requests of the attacker. Back, land, Neptune (Syn Flood), Pod (Ping of Death), smurf, teardrop are the six kinds of DoS attacks in KDDCUP 99. Back type of denial of service attack against the Apache web server, an attacker submits requests with URL's containing many front slashes. As the server tries to process these requests, it will slow down and becomes unable to process other requests. Back attack needs to know that requests for documents with more than some

number of front slashes in the URL should be considered an attack. In the Smurf attack, attackers use ICMP echo request packets directed to IP broadcast addresses from remote locations to create a denial-of-service attack. The Land attack occurs when an attacker sends a spoofed SYN packet in which the source address is the same as the destination address. Teardrop occurs due to IP fragmentation re-assembly code does not properly handle overlapping IP fragments. This attack by looking for two specially fragmented IP datagram. The first datagram is a 0 offset, fragment with a payload of size N, with the MF bit on (the data content of the packet is irrelevant). The second datagram is the last fragment (MF == 0), with a positive offset greater than N and with a payload of size less than n. Neptune attack describes that each half-open TCP connection made to a machine causes the 'tcpd' server to add a record to the data structure that stores information about all pending connections. This data structure is of finite size, and it can be made to overflow by intentionally creating too many partially-open connections. Neptune attack can be distinguished from normal network traffic by looking for a number of simultaneous SYN packets destined for a particular machine that are coming from an unreachable host. A host-based intrusion detection system can monitor the size of the tcpd connection data structure and alert a user if this data structure nears its size limit. Ping of Death attack has been reported when the systems reacts in an unpredictable fashion when receiving oversized IP packets. Possible reactions include crashing, freezing and rebooting. Ping of Death can be identified by noting the size of all ICMP packets and flagging those that are longer than 64000 bytes.

Based on the description above, the following rule structure derived from the KDDCUP 99 dataset and it was given in the table 1. In this proposed model, the hidden related information about the features was observed. Learners discussed among others, about possible potential variations in traffic records which helps to realize the prior knowledge of anomalous behaviors in advance. This proposed computational technique facilitates prompt detection and distinction of possible individual traffic records from the crowd. There are 97277 normal and 3, 91,450 DoS attacks traffic records in 10 percent labeled KDDCUP 99 data set. 2,80,790 smurf, 107201 Neptune, 2203 back, 979 teardrop, 21 land and 264 pod are in the 10 percent labeled KDDCUP 99 . After removing duplicated instances class, 97277 normal, 641 smurf, 51820 Neptune, 994 back, 19 land, 918 teardrop, 206 pod are the traffic records considered for training the proposed IDS. The rule structure of six types of Dos attacks in KDDCUP 99 dataset is shown in table 1. After TLBO, the extended rule set identified with respect to each attack is shown in table 2. Effectiveness of the IDS is evaluated by its ability to make correct predictions. Events are successfully labeled as normal and attacks. False positives refer to normal events being predicted as attacks. False negatives are attack events incorrectly predicted as normal events. Detection Accuracy (DA) is defined as the ratio of the sum of true negative and positive rate and sum of true and false positive and negative rate.

Table 1: Rule Structure of Dos Attacks in KDDCUP99 datasets

Sl. No.	Attack Description	Attack Type
1	protocol= ICMP, Service=ecr_i,src_byte=1032, flag=SF, host_count=255	smurf
2	protocol=tcp,service=private or ctf, flag=SO or SF, serror_rate=1	neptune
3	protocol=tcp,service=http, flag = SF or STFR,src_byte=54540, dst_byte=7300 or 8314, same_srv_rate=1,srv_count>=5	back
4	protocol=UDP,service=SF,src_byte=28,wrong fragment=3, dst_host_count=255	teardrop
5	protocol=tcp,service=finger,flag=SO,land=1,srv_count=2, dst_host_srv_error_rate >=0.17	land
6	Protocol=ICMP,service=ecr_i,flag=SF,src_byte=1480,wrong fragment=1,dst_hostcount=255,dst_host_diff_srv_rate =0.02	Pod

Table 2: Extended Rule set observed from the proposed Techniques

Sl.No	Attack Description	Attack Type
1	If (Duration <3)and (protocol_type=icmp) and (dst_byte=125016) Then Buffer_overflow	smurf
2	if {the connection has following information: source IP address 124.12.5.18; destination IP address:130.18.206.55; destination port number: 21; connection time: 10.1 seconds } then {stop the connection}	Neptune
3	protocol=tcp,service=http, flag = SF or RSTFR,src_byte=54540, dst_byte=7300 or 8314, same_srv_rate=1,srv_count>=5	back
4	If (source_bytes > 265616) and(source_bytes <= 283618) then Warezmaster Attack	teardrop
5	If (Duration 0 to 25) and (protocol_ type = tcp and UDP) and (service=ftp OR private OR other domain)	land
6	If (duration < 10 seconds) of an FTP connection/session, there are many Hot indicators (hot > 20) being set by a logged user then it is highly likely that is being executed	Pod

Table 3: Results Obtained from the Simulation

Test Data	Training Data	Test data	Deduction accuracy (%)			
			PSO-GA	FUZZY CLUSTERING	TLBO	GTLBO
Normal	97277	60255	99.0	99.2	99.5	99.7
Smurf	641	400	87	96	83	98
Neptune	51820	20500	95	98	96	99.5
Back	994	714	96	96	98	97
Teardrop	918	300	94	96	96	96
Land	19	07	99.4	99	94	95
Pod	206	101	99.4	98	99.5	99.5

Simulation results show that performance variations among evolutionary algorithms that used as computational intelligence in IDS are less. Clustering based algorithms performance is better compared to non-clustered. Results reveal that no evolutionary algorithm performs better for all types of DoS attacks. Simulation results of the proposed techniques are shown in table 3. Compared to the existing, proposed technique is efficient. It reduces more false negative compared to the existing work that reveals in the simulation results in table 3.

Conclusion and Future Works

In this paper, new computational techniques were proposed by extracting the role of teachers and learners in the classroom. The proposed method performs the classification task and extracts required knowledge using TLBO and GLTLBO. The proposed systems are high reliability and adequate interpretability, and are comparable with several well-known algorithms such as Fuzzy clustering. Results on intrusion detection data set from KDD cup-99 repository show that the proposed approach would be capable of classifying intrusion instances with a high accuracy rate in addition to the adequate interpretability of extracted rules. The results of GLTLBO are better than fuzzy clustering technique. In future work, Weighted Teacher Learner based Optimization technique will be used and their performance will be compared with some existing weight based algorithms.

References

- [1] Ariu, D., Giacinto, G., and R., Perdisci, 2007. "Sensing attacks in computers network with hidden Markov models", *Machine Learning and Data Mining in Pattern Recognition* 4571, pp. 449–463.

- [2] Arun Raj Kumar, P., and Selvakumar, S., 2012, "Identifying the type of high rate flooding attacks using a mixture of expert systems", *I.J. Computer Network and Information Security*, pp. 1-16.
- [3] Chang-Huang Chen, Group Leader Dominated Teaching-Learning Based Optimization, Department of Electrical Engineering, Tungnan University, Taiwan.
- [4] Crosbie, Mark, Gene Spafford, 1995, "Applying Genetic Programming to Intrusion Detection", *In Proceeding of 1995 AAAI Fall Symposium on Genetic Programming, Cambridge, Massachusetts*, pp.1-8.
- [5] Eberhart, R.C., and Shi, Y., 2001, "Particle swarm optimization: developments and applications and resources", *Proceedings of the 2001 Congress on Evolutionary Computation Seoul, South Korea*, pp. 81-86.
- [6] Gong R.H, Zulkemine.M., Anolmaesumi.P., 2005, "A Software Implementation of a Genetic Algorithm Based approach to Network Intrusion Detection", *Proceedings of the SNPD/SAWN'05*, pp.19-27.
- [7] Heady, R., Luger, G., Maccabe, A., Servilla, M., 1990. "The architecture of a network level intrusion detection system", *Computer Science Department, University of New Mexico*.
- [8] Jun Wang, Xu Hong, Rong-rong Ren, Tai-hang. Li., 2010, "A Real-time Intrusion Detection System based on PSO-SVM", *International workshop on Information Security and Application, Qingdao, china*, pp. 319-321.
- [9] Khaled Sellami, Rachid Chelouah, Lynda Sellami, Mohamed Ahmed Nacer, 2011, "Intrusion Detection Based on Swarm Intelligence using mobile agent". *International Conference on Swarm Intelligence*, pp.1-3.
- [10] Li.W., 2004, "Using Genetic Algorithm for Network Intrusion Detection", *Proceedings of the United States Department of Energy Cyber Security Group*.
- [11] Pervez, I. Ahmad, A., Akram, Swati, S.U., 2007, "A comparative analysis of artificial neural network technologies in intrusion detection systems", *WSEAS Transaction on Computers*, 6, pp.175-180.
- [12] Rao, R.V., Savsani, V.J., Vakharia, D.P., 2011, "Teaching-learning-based optimization: A Novel method for constrained mechanical design optimization problems", *International journal on computer-aided design*, pp.303-315.
- [13] Shelly Xiaonan.Wu., Wolfgang Banzhaf, 2010, "The use of computational intelligence in intrusion detection systems: review". *Journal of Applied Soft Computing*, 10, pp.1-35.
- [14] Wang.M.H. Zhang.D., Shin. K.G., 2002, "Detecting SYN flooding attacks in Proceedings of Annual Joint Conference of the IEEE Computer and Communications Societies", *INFOCOM*, 3, pp.1530-1539.
- [15] Zhang Fu, Marina Papatriantafidou, Philippas Tsigas, 2012, "Mitigating Distributed Denial of Service Attacks in Multiparty Applications in the Presence of Clock Drifts", *IEEE transactions on dependable and secure computing*, 9, pp.401-413.