

Effective Intrusion Detection Systems using Genetic Algorithm

Mr. Prakash N Kalavadekar, Dr. Shirish S. Sane

Department of Computer Engineering,
K.K. Wagh Institute of Engineering Education and Research, Nashik,
Savitribai Phule Pune University, Maharashtra.
kprak3004@gmail.com, sssane@kkwagh.edu.in

Abstract—Conventional methods of intrusion prevention like firewalls, cryptography techniques, have not proved themselves to completely defend networks and systems from newly generated malwares and attacks. Intrusion Detection Systems (IDS) are useful to find the correct solution to solve the current problems and became an important part of any security network infrastructure to detect these threats without generating any problem to network. The basic purpose of IDS is to detect attacks and their nature that may harm the computer system. Several different approaches for intrusion detection are available as per the literature. These approaches are broadly defined by three ways: i) Signature based approach ii) Anomaly based approach and iii) Hybrid approach that combines signature and anomaly detection approaches. The proposed system works for signature based concept using genetic algorithm as features selection and detection. The system is tested on KDDCup99 and NSL-KDD dataset using Weka3.6 classifiers and implemented classifier.

Index Terms—Attributes, Intrusion, Security, Signature.

I. INTRODUCTION

Security attacks are categorized as passive and active attacks. The passive attackers are usually hidden and do tapping of the communication link to gather data or destroy the network functioning. Passive attacks like eavesdropping, tampering, traffic monitoring and analysis. Active attacks are used to affect the operations within the network [1]. The performance of networking services will be get hampered or stopped because of these attacks. Active attacks are like hole attacks, Denial-of-Service (DoS), jamming, flooding etc. The security solutions for two types of networks (wireless or wired) are as given below: Prevention: It is like precaution taken for protecting network from any attack. Detection: If an attacker break the precautions made by the prevention system, then defending is difficult for such types of attacks. At this point, the detection comes in the picture to find out where the problem is occurred. Mitigation: In this step the actions will be taken on infected nodes to maintain security in the network [18]. In any security system, if prevention fails to stop intrusions, then detection system will be used for further process. Detection means finding suspicious behavior of user during a network communications. In the security set up, IDS offer information to the opposite systems such as identification, location (single node or group of nodes from particular region), time of the intrusion, type of intrusion

(active or passive), specific attack name, OSI layer such as physical, data link, network from where attack is happened. This data would be terribly useful in defense like mitigating and analyzing the results of attacks. So, IDS plays important role in network security. Intrusion is defined as any set of actions that damage the integrity, confidentiality, or normal working of system. The discovery of security keys to the intruders can compromise the security of nodes. So this will break the defined mechanism of preventive security. So the IDS will play the role of disclosure of intrusions for preventing important system resources. The IDS should have low false positive rate and high true positive rate. Thus there is a lot of scope for analysis in detection performance for unknown attacks and detection speed.

II. MOTIVATION AND RELATED WORK

Detection using Misuse or Signatures: -For known attacks signatures database is generated and is used for detecting future attacks. This type of detection methods always gives accurate and efficient finding of attacks which are known with low false positive rate [1]-[5]. The limitation is that it only works for known attack, if any new kind then it will not be useful to detect. The researcher Sobh says that such systems work like the anti-virus systems, which will be useful for only detecting some or all known attacks [18]. These systems use known attack dataset like KDD Cup 99 which contains 41 attributes for each signature of different types (DOS, R2L, U2R, and Probe) attacks [5]. Mostly internet based attacks can be detected by the IDS which are developed using neural network [Malki and Shun]. The feed forward neural network and the back propagation training algorithm were used to determine and predict current and possibly future attacks. For training and testing of classifier KDD Cup (1999) dataset is used. This method is only used for signature detection [4][13]. Sahana Devi K. J., Bharathi gives information of systems based on misuse model like SNORT and Bro [1]. Siva Sivatha Sindhu, S. Geetha and A. Kannan given decision tree based light weight signature based detection (nerotree) using a wrapper approach. As well it used genetic algorithm for optimizing selection of signature features from given 41 features in KDD Cup 99 dataset [5] [13]. Anomaly Detection: -In this behavior modeling is used such that profiles of users are prepared on the

basis of normal operations. The normality score is calculated and used to find certain deviation for declaration of anomaly [1] [7]. It is compulsory to update normal profiles periodically as per the changes in network behavior. These systems are able to detect unknown or any attack which is previously not occurred. According to the nature of the processing of behavioral data Garcia Teodoro had mentioned that it can be divided into three ways of implementation as follows[16] [18].

- 1] Statistical based: The profile is generated using stochastic behavior of the user and network. The network is monitored and profiles are generated periodically. An anomaly score is calculated with the help of reference profile. The score is checked for a certain threshold and depend on that declaration of the anomaly is done.
- 2] Knowledge based: The history based data of the network with normal and certain attacks condition is used.
- 3] Machine learning based: The system is trained with various patterns as explicit or implicit. The updating is done periodically so as to improve the intrusion detection performance on the basis of the previous results.

Hybrid Approach: -This approach combines signature and anomaly based detection approaches so that advantages of both approaches will improve the performance of the system. This approach works for detection of known and unknown attacks [1] [2] [4]. Neural network based classifier is designed by Koutsoutsos, Christou and Efremidisto give solution. The combinations of more than one neural network are used to detect attacks on web servers. The system is capable of detecting unseen attacks and making categorization. The rule based approach with enhanced C4.5 algorithm is suggested by Prema Rajeswari and Kannan for intrusion detection .This system is capable for detecting abnormal behaviors of internal attackers through classification and decision making in networks [9].

D. Barbara gives sensitivity for signature-based and anomaly-based IDSs with respect to the characteristics of the attacks, training history, services provided, and underlying network conditions. For labeled attacks data mining techniques are also useful to construct classification models [5] [8]. Lee et al. gives information about how to specify rules for anomaly detection with respect to normal problems [18]. Fan et al. further extended Lee et al.'s work to find accurate gaps between known attacks and unknown anomalies [18]. Kai Hwang, Min Cai, Chen, and Min Qin suggest data mining techniques where rule mining was used to design IDS. They have found that how single connection attacks differ from multi connection attacks. They also give information of systems based on misuse model like SNORT and Bro [1]. Gisung Kim, Seungmin Lee, Sehun Kim (2014) done the analysis on a brand new hybrid intrusion detection technique that hierarchically integrates a signature and an anomaly detection model. First, the C4.5 as decision tree is used to produce the signature detection model which decomposes the traditional training information to form small subsets. Associate anomaly detection model is formed using one-class support vector machine (1-class SVM) [2]. The experiments were conducted with the revised version of KDDCup99 data set, as NSL-KDD. By maintaining low false positive rate, their method is better in

detection rate for known and unknown than the conventional methods. The time complexity for the training and testing of the system is also significantly reduced. In one-class SVM, the labeled information is not required, but for real world false positive rate is may increase [21]. Wenying Feng,a,b, Qinglei Zhangc, Gongzhu Hud, Jimmy Xiangji Huangc (2014) take the advantage by combining SVM and CSOACNs(Clustering based on Self-Organized Ant Colony Network) avoiding their weaknesses. The system is evaluated with KDDCup-99 data set and found that CSVAC (Combining Support Vectors with Ant Colony) gives better performance in classification rate and efficiency than only SVM or CSOACN [19].

III. IMPLEMENTATION METHODOLOGY

1. Signature based IDS In Signature based IDS known attack pattern is used to train the system. The new record will be get compared with that attack and based on comparison decision will be given. The following Figure 1 shows architecture of signature based IDS, for the detection of known attacks.

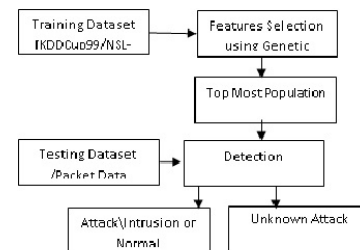


Fig. 1. Framework of EIDS

These IDS can be used to protect network and individual system. Before applying any learning algorithm data processing step is essential. By reducing attribute space a good understandable model can be designed. Feature reduction can be done by two approaches as1] wrapper which uses the learning algorithm to find out the usefulness of features,2] The filter which uses general characteristics of the data. The wrapper approach runs much slower but produces better result than filter.

1] Probe Attack The aim of such types of attack is to acquire information of targeted network by using external source. Hence, the duration of connection and source bytes features of basic connection level are significant while features like number of files creations and number of files accessed are not important to detect probe attacks.

2] DoS Attack The purpose of DoS attacks to stop the services forcefully by flooding it with illegitimate requests. So for the DoS attack, traffic features likes connections having same destination host and same service and packet level features such as the source bytes and percentage of packets with errors are important. The feature logged in or not is not important for detecting DoS attacks.

3] R2L Attack The R2L attacks required the network level and the host level features so they are difficult to detect. For

detecting R2L attacks the network level features as duration of connection and numbers of failed login attempts are important.

4] U2R Attack For the detection of U2R attacks semantic details are required which are not easy to capture at an earlier stage. These attacks always target an application which required content. For U2R attacks, features like number of file creations and number of shell prompts invoked are important and protocol and source bytes can be ignored.

Thus, from the total 41 features, 5 features for Probe attack, 9 features for DoS attack, 14 features for R2L attack and 8 features for U2R attack are important. So to get these numbers of optimized features GA (Genetic Algorithm) is used.

TABLE I
MOST RELEVANT CLASS FEATURES

Class label	Relevant features
Back	5,6
Land	7
neptune	3,4,5,23,26,29,30,31,32,34,36,37,38,39
Pod	8
Smurf	2,3,5,6,12,25,29,30,32,36,37,39
teardrop	8
Satan	27
ipsweep	36
Nmap	5
portsweep	28
normal	3,6,12,23,25,26,29,30,33,34,35,36,37,38,39
guesspasswd	11,6,3,4
ftppwrite	9,23
Imap	3,39
Phf	6,10,14,5

A. Algorithm Steps

1. Initialize the population randomly with the size of each chromosome to be 41.

Each gene value in the chromosome can be zero or one. The zero means presence of feature and one means not presence of feature.

2. Initialize $x = 0.2$, $y = 0.4$, $z = 0.4$. N =total number of records in the training set, $P_c = 0.01$ and $P_m = 0.011$.

3. for each chromosome in the new population

A] Do uniform crossover and mutation operation to the chromosome with probability values of P_c and P_m .

B] Evaluate Fitness = $x \cdot (1/\text{Number of Ones}) + y \cdot \text{Sensitivity} + z \cdot \text{Specificity}$

4. If (Current fitness Previous fitness ; 0.0001) then Quit.

5. Select the top best chromosomes into new population using tournament selection.

6. If number of generations is not reached, go to line 3.

IV. EXPERIMENTAL WORK

1] KDD Cup -99 Dataset The KDD Cup 99 dataset was prepared by DARPA. The database was generated using TCP/IP

raw data in local area network with multiple attacks. It consist of 4,898,431 connections records from that 3,925,650 are attacks. For each TCP/IP connection, 41 various quantitative and qualitative features were extracted. Each record of KDD Cup 99 dataset has 41 attributes from which some attributes are irrelevant and redundant. Because of these attributes it may add noise and affect the accuracy of proposed systems. Using a dataset with a large number of attributes require more time for training and detection processes, so the performance can be degraded. So data mining tools like Weka or Oracle Data mining can be used to find relevant attributes. The Minimum Description Length (MDL) algorithm will be used for ranking the attributes in a dataset based on their significance. It is found that 14 attributes out of the 41 attributes of the KDD Cup 99 dataset have an importance value above zero and the rest have an importance of zero [12].

The following four categories are defined for the attacks:

TABLE II
CATEGORIES WITH ATTACK NAMES

Category	Name of Attacks
DOS	'neptune', 'back', 'smurf', 'pod', 'land', 'teardrop'
U2R	'bufferoverflow', 'loadmodule', 'rootkit', 'perl'
R2L	'warezclient', 'multihop', 'ftppwrite', 'spy', 'imap', 'guesspasswd', 'warezm'
PROBE	'portsweep', 'satan', 'nmap', 'ipsweep'

2] NSL-KDD revised version of KDDCup -99 The mostly used dataset in various IDS is KDD Cup 99. But as per the research done using statistical analysis on this dataset, it is found that its affect on the performance of the various approaches of IDS. So solution is a new dataset, NSL-KDD, which is revised version of KDD Cup 99 data set. The NSL-KDD overcomes some of the pitfalls of the original KDD Cup 99 dataset are as follows:

A] The redundant records are removed from the training dataset because of that no biased results will be generated by classifiers for more frequent records.

B] The proportion of number of selected records from each difficulty level group is maintained with the percentage of records in the original KDD Cup 99 dataset. So that accuracy of different machine learning methods is efficient.

C] The numbers of records in the training and testing datasets are sufficient for the experiments, so that results of different research works will be consistent and comparable.

NSL-KDD consists number of records are given in follow- ing Table 2.

3] Experiment and Testing of Signature Based Detection As per Figure 1, we selected different training files based on KDDCup-99 and NSL-KDD dataset. Genetic algorithm is applied on selected population with number of generations to get top most population which is further used for detection process. The Brute Force Single-Keyword Pattern Matching algorithm is used for detecting known and unknown attack. Further results are tested using Weka 3.6 with different classifiers and 14 selected attributes as duration, protocoltype , ser-

TABLE III
NUMBER OF RECORDS IN NSL-KDD

Types of attacks	Training Dataset	Testing Dataset
Normal	67343	9711
DOS	45927	7456
PROBE	11656	2421
U2R	52	200
R2L	995	2756

vice, flag, srcbytes, dstbytes, wrongfragment, numfailedlogins, rootshell, numroot, rerrorrate, srverrorrate, dsthostdiffsrtrate, dsthosterrorrate. The Table IV gives number of features selected by different methods and detection accuracy.

TABLE IV
FEATURES SELECTION BY DIFFERENT METHODS AND DETECTION ACCURACY

Name of Method	Features	Accuracy
Ranker	34	98.15
bestfit+ConsistencySubsetEval	11	97.01
grredystepwise+attrsubseval	9	97.97
CFS Substate (Random Forest Weka)	15	98.88
Reference paper (NeuroTree)	16	98.38
PCA (Neural Network BPA)	8	77.16
Nil	41	98.47

Some sample output of GA:

tcp,http,SF,181,5450,0,0,1,0,0,normal.,

Generation: 1 Fittest: 14

Generation: 2 Fittest: 14

Solution found!

Genes in current chromosomes:

000000000100000010111100000001011100001111

Current Specificity: 2

Current Sensitivity: 14.966666666666667

Generation: 1 Fittest: 14

Generation: 2 Fittest: 14

Solution found!

The Table V gives the testing result using Weka and 14 selected features on different dataset files.

Performances of implemented system have been evaluated using accuracy, detection rate, false positive rate and F-measure metrics and are defined by

$$Accuracy = (TP + TN) / (TP + TN + FN + FP) \quad (1)$$

$$DetectionRate = TP / (TP + FN) \quad (2)$$

$$FalsePositiveRate = FP / (FP + TN) \quad (3)$$

where, True Positive (TP) is the number of actual attacks classified, True Negative (TN) is the number of actual normal records classified as normal ones, False Positive (FP) is the

TABLE V
WEKA RESULT ON DIFFERENT DATASET INSTANCES

Dataset (Instances)	Feature	Classifier	DR
1000 (KDDCup99)	14	NeuroTree	98.04
1360 (KDDCup99)	14	RandomForest (Weka)	100
1360 (KDDCup99)	14	NaiveBayes (Weka)	98.30
1360 (KDDCup99)	14	DecisionStump (Weka)	82.20
1360 (KDDCup99)	14	RandomTree (Weka)	100
1360 (KDDCup99)	14	REPTree (Weka)	99.34
14952(KDD20)	14	RandomForest (Weka)	99.02
14952(KDD20)	14	NaiveBayes (Weka)	71.34
14952(KDD20)	14	DecisionStump (Weka)	88.20
14952(KDD20)	14	RandomTree (Weka)	99.02
14952(KDD20)	14	REPTree (Weka)	99.02
125973(KDDTest21)	14	RandomForest (Weka)	99.16
125973(KDDTest21)	14	NaiveBayes (Weka)	39.09
125973(KDDTest21)	14	DecisionStump (Weka)	83.20
125973(KDDTest21)	14	RandomTree (Weka)	99.16
125973(KDDTest21)	14	REPTree (Weka)	98.84
25192(NSLKDDTest)	14	RandomForest (Weka)	99.12
25192(NSLKDDTest)	14	NaiveBayes (Weka)	42.71
25192(NSLKDDTest)	14	DecisionStump (Weka)	83.30
25192(NSLKDDTest)	14	RandomTree (Weka)	99.12
25192(NSLKDDTest)	14	REPTree (Weka)	98.40

number of actual normal records classified as attacks, False Negative (FN) is the number of actual attacks and normal records classified as unknown records. The F-measure is a harmonic mean between precision and recall.

$$Precision = TP / (TP + FP) \quad (4)$$

$$Recall = TP / (TP + FN) \quad (5)$$

$$F-measure = (2(Precision * Recall)) / (Precision + Recall) \quad (6)$$

The classification performance of implemented system is shown in Table VI.

TABLE VI
PERFORMANCE CLASSIFICATION FOR ATTACKS BASED ON DATASETS

Dataset	Instances	DR	FPR	Accuracy	F-measure
KDDCup2	1360	99.85	0.00	99.85	99.92
KDDCupTest	11850	90.25	0.00	90.25	94.87
NSLKDDTest	125973	94.93	0.00	97.66	97.39
KDDCup10%	494021	99.73	0.00	99.78	99.86

The below graph shows the number instances used in testing and achieved detection rate as per given in Table VI. It shows that detection rate is better for KDDCup than NSLKDD dataset. The Table 6 shows number of top most population

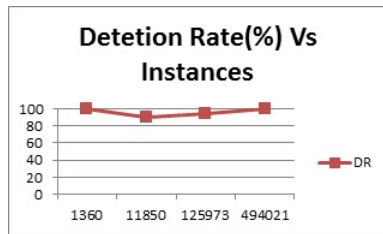


Fig. 2. Detection rate vs Number of Instances

TABLE VII
TOP MOST POPULATION

Dataset	Top Population	Detection Rate
KDDCup2	18	99.85
KDDCupTest	187	90.25
NSLKDDTest	187	94.93
KDDCup10%	6447	99.73

generated by GA which is used for detection. The graph shows attacks detected as DoS, Probe, U2R and R2L.

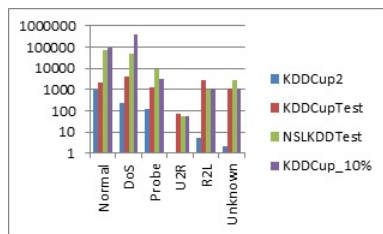


Fig. 3. Number of Attacks Detected

V. CONCLUSION

The current IDS technologies are not sufficient enough to provide a reliable detection rate so the rate can be improved by using different methodologies. The information presented gives how the features selection plays important role to increase the speed of detection and detection accuracy. The top most population generated using GA can be used for further classification using standard method.

REFERENCES

- [1] Min Cai, Kai Hwang and Min Qin Hybrid intrusion detection with weighted signature generation over anomalous internet episodes, IEEE Transactions on Dependable And Secure Computing, Vol.4 No.1, Jan-March 2007.
- [2] Gisung Kim, Seungmin Lee, Sehun Kim A novel hybrid intrusion detection method integrating anomaly detection with misuse detection, Expert Systems with Applications, Elsevier Ltd, 2014.
- [3] S. Jajodia L., Popyack D. Barbara, J. Couto and N. Wuy. Adam, Detecting Intrusions by data mines, Technical report, Workshop Information Assurance and Security, USA, 2001.
- [4] Bharathi M. Sahana Devi K. J., Hybrid intrusion detection with weighted signature generation, Technical report, Dept of CSE, Chickballapur, 2011.
- [5] Siva S. SivathaSindhu, S. Geetha, A. Kannan" Decision tree based light weight intrusion detection using a wrapper approach", Expert Systems with Applications 39 129-141, 2012.
- [6] Kapil Kumar Gupta, BaikunthNath, RamamohanaraoKotagiri," Layered Approach Using Conditional Random Fields for Intrusion Detection" IEEE Transactions on Dependable and Secure Computing, Vol.4 No.1, Jan-March 2010
- [7] Dr. Saurabh Mukherjeea, Neelam Sharma," Intrusion Detection using Naive Bayes Classifier with Feature Reduction", Procedia Technology, 119 128, 2012.
- [8] Bertrand Portier, Froment-Curtill," Data Mining Techniques for Intrusion Detection", The University of Texas at Austin, Dr. Ghosh - EE380L Data Mining Term Paper, Spring 2000.
- [9] L. PremaRajeswari, KannanArputharaj," An Active Rule Approach for Network Intrusion Detection with Enhanced C4.5 Algorithm", I. J. Communications, Network and System Sciences, 4, 284-359 Published Online, November 2008.
- [10] Nahla Ben Amor, Salem Benferhat," Naive Bayes vs Decision Trees in Intrusion Detection Systems" , SAC04, March 14-17, Nicosia, Cyprus, 2004.
- [11] Ahmed H. Fares and Mohamed I. Sharawy," Intrusion Detection: Supervised Machine Learning", Journal of Computing Science and Engineering, Vol. 5, No. 4, pp. 305-313, December 2011.
- [12] AdetunmbiA.Olusola,, AdeolaS.Oladele and Daramola O.Abosede,"Analysis of KDD 99 Intrusion Detection Dataset for Selection of Relevance Features", Proceedings of the World Congress on Engineering and Computer Science 2010, Vol I WCECS 2010, San Francisco, USA, October 20-22 2010.
- [13] MahbodTavallaee, EbrahimBagheri, Wei Lu and Ali A., Ghorbani,"A Detailed Analysis of the KDD CUP 99 Data Set, Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009).
- [14] TaisirEldos, Mohammad KhubebSiddiqui and AwsKanan, The KDD99 Dataset: Statistical Analysis for Feature Selection", Journal of Data Mining and Knowledge Discovery ISSN: 2229-6662 and ISSN: 2229-6670, Volume 3, Issue 3, pp.-88-90, 2012.
- [15] YisehaeYohannes, JohnHoddinott,"Classification and Regression Trees: An Introduction", International Food Policy Research Institute, 2033 K Street, N.W. Washington, D.C., U.S.A, 2006
- [16] Peyman Kabiri and Ali A. Ghorbani, Research on Intrusion Detection and Response: A Survey, International Journal of Network Security, Vol.1, No.2, PP.84102, Sep. 2005.
- [17] Wenke Lee and Salvatore J. Stolfo, Data Mining Approaches for Intrusion Detection, 7th USENIX Security Symposium, 1998.
- [18] Ismail Butun, Salvatore D. Morgera, and Ravi Sankar, A Survey of Intrusion Detection Systems in Wireless Sensor Networks, IEEE Communications Surveys and Tutorials, 2013.
- [19] WenyingFeng, Quinglei, Gongzhu Hu, Jimmy Xiangi Huang, Mining Network data for intrusion detection through combining SVMs with ant colony networks, Future Generation Computer Systems, Elsevier, 2013.
- [20] Kapil Kumar Gupta, BaikunthNath, Senior Member, IEEE, and RamamohanaraoKotagiri, Member, IEEE, Layered Approach Using Conditional Random Fields for Intrusion Detection, IEEE Transactions on Dependable and Secure Computing, Vol. 7, No. 1, January-March 2010.
- [21] Prakash Kalavadekar. Dr. Shirish Sane Effective Intrusion Detection Systems using Hybrid Approach International Journal of Exploring Emerging Trends in Engineering, Volume 3 Issue 2 Mar-Apr-2016.