International Journal of Engineering & Technology, 7 (4.19) (2018) 1011-1015

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Effect of Mutation and Crossover Probabilities on Genetic Algorithm and Signature Based Intrusion Detection System

Mr. Prakash N Kalavadekar [1]*, Dr. Shirish S. Sane [2]

[1] *Research Scholar, K.K Wagh Institute of Engineering Education & Research, Nashik*
[2] *Research Guide, K.K Wagh Institute of Engineering Education & Research, Nashik*
Savitribai Phule Pune University, India
**Corresponding author Email: kprak3004@gmail.com*

Abstract

Conventional methods of intrusion prevention like firewalls, cryptography techniques or access management schemes, have not provided complete protection to computer systems and networks from refined malwares and attacks. Intrusion Detection Systems (IDS) are giving the right solution to the current issues and became an important part of any security management system to detect these threats and will not generate widespread harm. The basic goal of IDS is to detect attacks and their nature that may harm the computer system. Several different approaches for intrusion detection have been reported in the literature. The signature based concept using genetic algorithm as features selection and, J48 as classifier to detect attack is proposed in this paper. The system was evaluated on KDD Cup 99, NSL-KDD and Kyoto 2006+ datasets.

Keywords: *Intrusion Detection, Security, Signature, Features, J48.*

1. Introduction

Security attacks are classified into 2 main branches: passive and active. The passive attackers are usually invisible (hidden) and do tapping of the communication link to gather data or destroy the network functioning parts. Passive attacks are classified as eavesdropping, tampering, traffic monitoring and analysis. Active attacks are used to affects the operations within the network [1].
The performance of networking services will be get degraded or stopped because of these attacks. Active attacks are classified as 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 provides preventing before happening of any attack. Signature based technique can used to protect against the targeted 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 protection answer would instantly use the 'detection' section of the attack to find which parts of the nodes are being compromised.
Mitigation: In this step the affected nodes were removed from the network and securing the network [18].
In any security system, if prevention does not 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 referred as: "any set of actions that plan to compromise the integrity, confidentiality, or handiness of a resource" and intrusion interference techniques like encoding, authentication, access management, secure routing etc. are parts of the initial phase of defense solutions for intrusions. But till there are security systems does not provide fully preventions for intrusions. 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 posses as: "low false positive rate, calculated because the proportion of normalcy variations detected as anomalies, and high true positive rate, calculated because the proportion of anomalies detected". Thus there's plenty a lot of scope for analysis in up detection performance for unknown attacks & detection speed.

2. Motivations 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 & 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 useful to detect.
The researcher sobh says that such systems works like the antivirus systems, which will be useful for only detecting some or all known attacks [18].
These systems used known attack dataset like KDD Cup 99 which contains 41 attributes for each signature of different types (DOS, R2L, U2R, and Probe) attacks [5].

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Mostly internet based attacks can detected by the IDS which are developed using neural network [Malkiand Shun]. The feed forward type neural network with the back propagation training algorithmic was used to determine and predict current and possibly future attacks. For training & 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 SivathaSindhu, 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.

Depend on 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 & 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 & unknown attacks [1] [2] [4].

Neural network based classifier is designed by Koutsoutos, 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 KDD Cup 99 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.

Wenying Fenga, Qinglei Zhangc, Gongzhu Hud, Jimmy Xiangji Huange (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].

3. Implementation Methodologies

Signature based IDS can be trained by using previously known attack pattern. Whenever new record comes to system it compares that pattern with previously known attack pattern and based on comparison decision will be given. Figure 1 shows proposed architecture of Effective IDS, in which signature based detection system will be used for detection of known & unknown attacks**.**

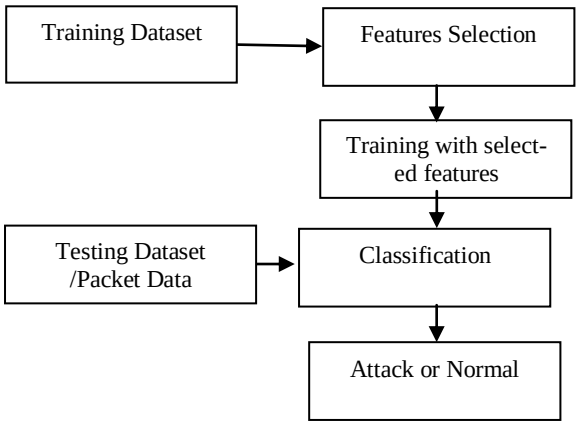


Figure 1: Framework of Signature based IDS

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 as

1] Wrapper which required the learning algorithm to find out the importance of features 2] the filter which uses general characteristics of the data. The filter approach runs faster than wrapper approach but wrapper produces better result than filter.

As per the survey, from the total 41 features of signature based dataset, Probe attack required 5, DoS attack required 9, R2L attack required 14 and U2R attack required 8 which are important features to detect these attacks. So to get these numbers of optimized features GA (Genetic Algorithm) is used.

Features selection using Genetic Algorithm (GA) requires taking care of encoding & fitness function. For IDS, a binary encoded fixed length string can be used where the gene value will be 0 or 1 which will decided from the number of features. So, each individual chromosome with fixed length in population represents the given features set.

Fitness Function: - The fitness feedback is required to evaluate feature subset which is represented in GA population which will be helpful for enhancing detection rate and accuracy of the IDS.

Algorithm Steps:
Input- Binary encoded string which is having length n (where n is the number of features), population size, generations count, Uniform crossover probability (Pc), Mutation probability (Pm), Empty solution.
Output- Selected important features.

1. Initialize the population with chromosome which has size n.
2. In the chromosome each gene value can be '0' or '1'. (0- means feature value zero and 1- means feature value other than zero)
3. Initialize Pc and Pm, Maximum Fitness.
4. While ((current_fitness - previous_fitness) > 0.001) {
a. With the specified probability Pc & Pm do uniform crossover and mutation operations.
b. Increment fitness value if solution bits match with gene bits.
5. Using tournament selections find the best of chromosomes into new population. }
6. Display the solution with selected features.

4. Results and Analysis

4.1. Datasets description

The KDD Cup 99 dataset, NSL-KDD dataset and Kyoto 2006+ dataset have been commonly used in the literature to evaluate the performance of various IDS. Mostly the KDD Cup 99 dataset is used in different IDS to evaluate the performance. Every datasets having the different data sizes and numbers of features, because of that it is possible to validate different feature selection methods.
The KDD Cup 99 consists of five different classes, which are normal and four types of attack (i.e., DoS, Probe, U2R and R2L). Either normal or an attack label is used for each record and having 41 different quantitative and qualitative features.
Tavallaee had proposed NSL-KDD as a new revised version of the KDD Cup 99. The NSL-KDD had solved few problems of the KDD Cup 99 dataset like redundant records in KDD Cup 99 data. Song had presented the Kyoto 2006+ dataset which includes real traffic data of three years between November 2006 and August 2009. The data was collected using honeypots and regular servers at Kyoto University. The following Table 1 gives details about number of features & number of records in each dataset.

Table 1: Datasets Details

Dataset	# Features	Records
KDD Cup 99 10%	41	494021
NSLKDD	41	125973
Kyoto 2006+ (1-3 Nov 2007)	24	237718
Kyoto 2006+ (27-31 Aug 2009)	24	777110

4.1. Experimental Setup

In all experiments filter based feature selection is using genetic algorithm. The GA is applied on three selected datasets using different uniform crossover probability (Pc) and mutation probability (Pm) values to get the important features as shown in Table 2.
The proposed feature selection algorithm is evaluated using J48 as classifier from Weka 3.8.1. The KDD Cup 99, NSL-KDD and Kyoto 2006+ datasets are used to evaluate the performance of IDS. The data of 27-31 August 2009 and 1-3 Nov 2007 was selected from Kyoto 2006+ for the experiments. To evaluate the detection performance a 10-fold cross-validation is used.

Table 2: Number of Features selected with different Pc & Pm values

Dataset	Pc	Pm	#Features	Selected Features
KDD Cup 99	0.9	0.001	25	f1,f2,f3,f4,f5,f6, f10,f12,f13,f14,f16,f17, f23 f24,f29,f32,f33,f34,f36,f37,f38,f39,f40, f41
	0.7	0.001	16	f2,f3,f4,f5,f6, f12, f23, f24, f29, f31, f32, f33,f34,f36,f37,f39
	0.6	0.001	14	f2,f3,f4,f5,f6, f12, f23, f24, f29, f31, f32, f33,f34,f36
NSL-KDD	0.9	0.001	14	f1,f2,f3,f4,f5,f6, f23, f24, f29,f32, f33, f34, f35,f36
	0.6	0.001	12	f2,f3,f4,f5,f6,f23,f24,f29,f32, f33, f34, f35
Kyoto 2006 + (27-31 August 2009)	0.9	0.0001	6	f2,f14,f16,f17,f19,f20
	0.6	0.0001	7	f2,f14,f15,f16, f17,f19,f20
Kyoto 2006 + (1-3 Nov 2007)	0.7	0.0001	6	f2,f14,f15,f16,f19,f20
	0.2	0.0001	6	f2,f14,f16,f17,f19,f20

4.3. Performance Evaluation

Performance of implemented system has been evaluated using accuracy, detection rate, false positive rate which are calculated as given below.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP} \quad (1)$$

$$\text{Detection Rate} = \frac{TP}{TP+FN} \quad (2)$$

$$\text{False Positive Rate} = \frac{FP}{FP+TN} \quad (3)$$

where, True Positive (TP) is the number of actual attacks classified as attacks, True Negative (TN) is the number of actual normal records classified as normal ones, False Positive (FP) is the number of actual normal records classified as attacks, False Negative (FN) is the number of actual attacks classified as normal or unknown records.
The F-measure is a harmonic mean between precision and recall.

$$F-\text{measure} = \frac{2(\text{Precision}*\text{Recall})}{\text{Precision}+\text{Recall}} \quad (4)$$

The precision is the proportion of predicted positives values which are actually positive. The precision value directly affects the performance of the system. A higher value of precision means a lower false positive rate and vice versa. The precision is given by (5).

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

The recall is to indicate the proportion of the actual number of positives which are correctly identified. The recall is defined as:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (6)$$

4.3. Discussion of results

The Table 3, 4, 5 and 6 shows the performance of classifier with different features selected by the proposed GA. The feature selection shows the enhancement in the classification performance of IDS. The results are improved in terms of low computational cost and high detection rate. Table 3-6 shows the results of classification with detection rates, false positive rates and accuracy rates. By observing results from these tables the detection system combined with the GA is getting an accuracy rate of 99.95%, 99.39%, 99.60% and 99.75% for KDD Cup 99, NSL-KDD, Kyoto

2006+(27-31 Aug 2009) and Kyoto 2006+(1-3 Nov 2007) respec- tively.

Table 3: Classification performance based on the KDD Cup 99 (494021)

Method (#Features)	Build Time (sec)	Accuracy	DR	FPR	Precision	Recall	F-measure
J48 + All (41)	40.35	99.92	99.99	0.00	99.90	99.99	99.94
J48 + GA (25)	63.83	**99.96**	**100**	**0.00**	**100**	**100**	**100**
J48 + GA (16)	40.39	99.95	99.99	0.00	99.99	99.99	99.99
J48 + GA (14)	30.56	99.94	99.99	0.00	99.99	99.99	99.99
J48 + GA (12)	**26.33**	99.95	99.99	0.00	99.99	99.99	99.99

Table 4: Classification performance based on the NSL-KDD (125973)

Method (#Features)	Build Time (sec)	Accuracy	DR	FPR	Precision	Recall	F-measure
J48 + All (41)	74.45	99.75	**99.98**	**0.2**	**99.70**	**99.98**	**99.70**
J48 + GA (14)	21.67	99.39	99.40	**0.2**	99.40	99.40	99.40
J48 + GA (12)	**12.50**	99.34	99.30	**0.2**	99.30	99.30	99.30

Table 5: Classification performance based on the Kyoto 2006 + (27-31 Aug 2009) (151958)

Method (#Features)	Build Time (sec)	Accuracy	DR	FPR	Precision	Recall	F-measure
J48 + All (21)	23.39	98.88	98.90	1.5	97.90	98.90	98.40
J48 + GA (17)	14.76	**99.60**	**99.60**	**0.5**	**99.40**	**99.60**	**99.45**
J48 + GA (06)	**2.55**	99.51	99.50	0.65	99.40	99.50	99.35

Table 6: Classification performance based on the Kyoto 2006 + (1-3 Nov 2007) (237718)

Method (#Features)	Build Time (sec)	Accuracy	DR	FPR	Precision	Recall	F-measure
J48 + All (21)	12.55	**99.92**	**99.90**	**0.1**	**99.90**	**99.90**	**99.90**
J48 + GA (17)	26.53	99.68	99.70	0.3	99.70	99.70	99.70
J48 + GA (06)	**12.39**	**99.75**	99.80	0.2	99.70	99.80	99.70

The J48-IDS is computationally efficient when used with proposed feature selection algorithm. The time (sec) required to build model is shown in Figure 2. It is observed that less time is required to build model using GA than using all features for KDD Cup 99 and NSL-KDD datasets.

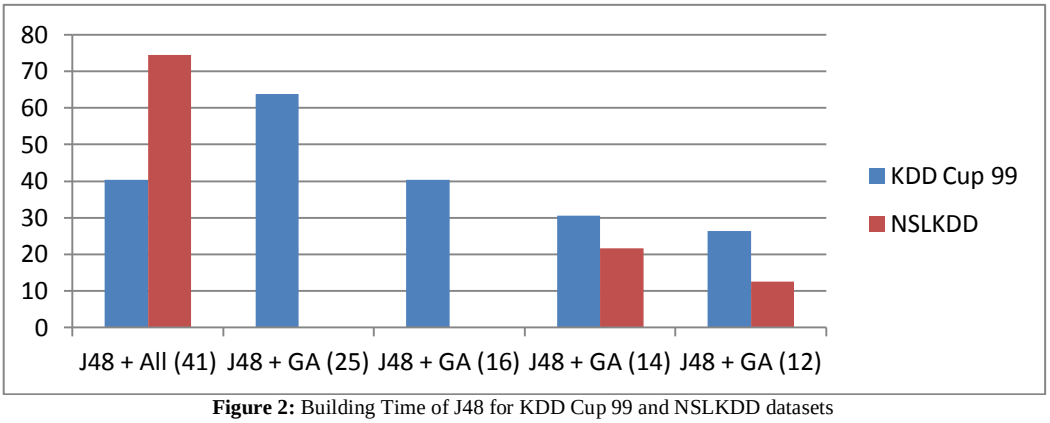


Figure 2: Building Time of J48 for KDD Cup 99 and NSLKDD datasets

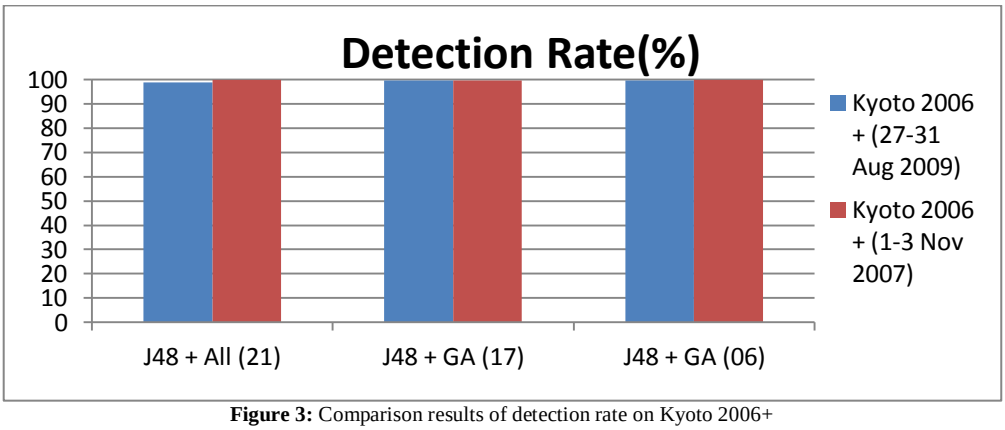


Figure 3: Comparison results of detection rate on Kyoto 2006+

5. Conclusion

Since the current IDS technologies are not sufficient enough to provide a reliable detection rate so work should be carried on to improve the rate. The paper gives that how the features selection is an important to reduce training time and maintain detection rate with accuracy. The filter based features selection genetic algorithm is proposed and used with J48 as classifier method. The evaluation of proposed system was done using three well known datasets as KDD Cup 99, NSL-KDD and Kyoto 2006+. The classification accuracy, detection rate, false positive rate and F-measure achieved by proposed system is up to the mark with existing detection approaches. It can be concluded that the pro-

posed IDS achieved good performance in detecting attacks on all datasets used for experimentation. Performance shown by proposed feature selection algorithm is encouraging till further enhancement can be done.

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] Mohammed A. Ambusaidi, Priyadarshi Nanda "Building an intrusion detection system using a filter-based feature selection algorithm", IEEE Transactions on computers, November 2014.

[3] 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.

[4] S. Jajodia L., Popyack D. Barbara, J. Couto and N. Wuy. Adam, "Detecting Intrusions by data mining ", Technical report, Workshop Information Assurance and Security, USA, 2001.

[5] Bharathi M. Sahana Devi K. J.,"Hybrid intrusion detection with weighted signature generation",Technical report, Dept of CSE, Chickballapur,2011.

[6] 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.

[7] 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

[8] Dr. SaurabhMukherjeea, Neelam Sharma," Intrusion Detection using Naive Bayes Classifier with Feature Reduction", Procedia Technology, 119 – 128, 2012.

[9] Bertrand Portier, Froment-Curtil," Data Mining Techniques for Intrusion Detection", The University of Texas at Austin, Dr. Ghosh - EE380L Data Mining Term Paper, Spring 2000.

[10] 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.

[11] Nahla Ben Amor, Salem Benferhat," Naive Bayes vs Decision Trees in Intrusion Detection Systems" , SAC'04, March 14-17, Nicosia, Cyprus,2004.

[12] 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.

[13] 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.

[14] 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).

[15] TaisirEldos, Mohammad KhubebSiddiqui and AwsKanan,"The KDD99 Dataset: Statistical Analysis for Feature Selection",Journal of Data Mining and Knowledge Discovery ISSN: 2229-6662 & ISSN: 2229-6670, Volume 3, Issue 3, pp.-88-90, 2012.

[16] YisehacYohannes,JohnHoddinott,"Classification and Regression Trees: An Introductin",International Food Policy Research Institute,2033 K Street, N.W.Washington, D.C., U.S.A, 2006

[17] PeymanKabiri and Ali A. Ghorbani, "Research on Intrusion Detection and Response:A Survey", International Journal of Network Security, Vol.1, No.2, PP.84–102, Sep. 2005.

[18] Wenke Lee and Salvatore J. Stolfo, "Data Mining Approaches for Intrusion Detection", 7th USENIX Security Symposium, 1998.

[19] Ismail Butun, Salvatore D. Morgera, and Ravi Sankar, "A Survey of Intrusion Detection Systems inWireless Sensor Networks", IEEE Communications Surveys & Tutorials, 2013.

[20] 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.

[21] Kapil Kumar Gupta, BaikunthNath, Senior Member, IEEE, and Ramamohanarao Kotagiri, 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.

[22] Prakash Kalavadekar, Dr.Shirish Sane "Effective Intrusion Detection Systems using Hybrid Approach"International Journal of Exploring EmergingTrends in Engineering, Voume 3 Issue 2 Mar-Apr-2016

[23] Prakash Kalavadekar, Dr.Shirish Sane "Effective Intrusion Detection Systems using Genetic Algorithm",International Journal on Emerging Trends in Technology, Voume 4, Special Issue July-2017, pp.8315-8319.