International Conference on Communication and Information Processing (ICCIP-2019)
Available on: Elsevier-SSRN

Adaptive Jaya Optimization Technique for Feature Selection in NSL-KDD Data Set of Intrusion Detection System

Thupakula Bhaskar*[1], Tryambak Hiwarkar[2], K. Ramanjaneyulu[3]

[1]*Research Scholar,SSSUTMS,Bhopal,M.P ,India.*
[2] *SSSUTMS,Bhopal,M.P ,India,* [3]*PVPSIT,Vijayawada,A.P ,India*

Abstract

Now a days, network traffic is increasing due to the exploding usage of smart devices and the Internet. The intrusion detection work centered on feature selection or decrease because few of the features are irrelevant and excess which results prolonged detection procedure and reduces the performance of an intrusion detection system (IDS). The NSL-KDD data set is a refined variant of its predecessor KDD‟99 data set. The intent of this work is to determine essential selected input features in building IDS that is computationally efficient and amazing. For this standard feature selection jaya optimization method is used. In this paper the NSL-KDD data set is analysed and applied Adaptive Jaya Technique for selecting best features to minimize low false alarm rate & maximize detection rate.

Keywords- Intrusion Detection System, NSL-KDD dataset, feature selection ,Machine Learning.

1. Introduction

In recent world, Internet and computer networks has become a crucial part in various organizations for surviving technological world. Most of the customer's share their personal information in the networks and also many concerns are also depend on the network for their daily business[1].Cyber-attacks the interruption of computers usual working and the loss of an important data via malicious network actions remain fetching more expensive[2].

Nomenclature	
IDS	Intrusion Detection System
DoS	Denial of Service Attack position
U2R	User to Root Attack
R2L	Remote to Local Attack
AJO	Adaptive Jaya Optimization (AJO)

[1] * Corresponding author. Tel.+91-9766466079.
E-mail address: shiridisaibaba22@gmail.com.

Intrusion Detection is a method of detecting, monitoring and analyzing the actions which are measured as destructions for the policies related to security of a network environment [3][4].The idea of identifying attacks in the cyber system on networks through a outline for intrusion detection system (IDS),this is depend on the theory which security damages could be sensed through audit records in monitoring system for abnormal designs of system usage [5].

2. Related Work

An internet is one of the popular one among number of users and various cyber-attacks are generated against internet. A fast and efficient cyber security intrusion detection is a major recent research challenge due to the increasing usage of internet based services. The aim of intrusion detection research is to overcome the drawbacks of existing approaches in internet security. High detection time, low accuracy and low flexibility are the common drawbacks of intrusion detection approaches. Advanced features used by intruders such as IP address spoofing, encrypted payload and dynamic ports which should be determined before any losses occur. So it is needed to diagnose intrusion from intruder by proper feature learning.

2.1 Feature Selection

Feature selection is crucial to rising the efficiency of machine learning algorithms. It is the procedure of selecting a subset of primary features according to definite criteria and is an essential and often used technique in machine learning for dimension reduction. Most of the data contains inapplicable, redundant, or noisy features. Feature selection cut down the number of features, removes inapplicable, extra, or noisy features and brings about perceptible effects on applications: speed up a machine learning algorithm, better learning accuracy and directing to improved model quality [7].

2.2 NSL-KDD Data Set

The NSL-KDD data set recommended to work out few of the built-in problems of the KDDCUP'99 data set.KDDCUP’99 is the largely utilized data set for anomaly detection. Types of Features were in Fig 1. [8].

Type	Features
Nominal	Protocol_type(2), Service(3), Flag(4)
Binary	Land(7), logged_in(12), root_shell(14), su_attempted(15), is_host_login(21),, is_guest_login(22)
Numeric	Duration(1), src_bytes(5), dst_bytes(6), wrong_fragment(8), urgent(9), hot(10), num_failed_logins(11), num_compromised(13), num_root(16), num_file_creations(17), num_shells(18), num_access_files(19), num_outbound_cmds(20), count(23) srv_count(24), serror_rate(25), srv_serror_rate(26), rerror_rate(27), srv_rerror_rate(28), same_srv_rate(29) diff_srv_rate(30), srv_diff_host_rate(31), dst_host_count(32), dst_host_srv_count(33), dst_host_same_srv_rate(34), dst_host_diff_srv_rate(35), dst_host_same_src_port_rate(36), dst_host_srv_diff_host_rate(37), dst_host_serror_rate(38), dst_host_srv_serror_rate(39), dst_host_rerror_rate(40), dst_host_srv_rerror_rate(41)

Fig. 1. Types of Features

The NSL-KDD data contains 41 features and 5 classes that are normal and 4 types of attacks: Dos,Probe, R2L, and U2R. Denial of Service Attack (DoS) is an attack in which the attacker makes some calculation or memory resource excessively busy or too full to hold authorized requests or contradict legitimate users access to a machine. Probing Attack is an effort to gather information close to a network of computers for the apparent purpose of surround its security controls. User to Root Attack (U2R) is a class of use in which the attacker beginning out with access to a normal user account on the system and is able to use some vulnerability to gain root accession to the system. Remote to Local Attack (R2L) happen when an attacker who has the ability to direct packets to a machine over a network [8].

3. Proposed System Using Adaptive Jaya Technique

In the proposed System the features are selected from the NSL-KDD dataset using Adaptive Jaya optimization[6]. In the proposed system we use the NSL-KDD dataset. Initially the data is loaded and the best features among the data is selected by using Adaptive Jaya optimization technique. The features are given as input for the Modified deep neural network. The deep neural network consists of several layers such as input layer, Hidden layer and Output layer. By using the Adaptive Jaya optimization technique the performance in detection rate is increased and produces less false alarm rate.

3.1. Working principle of Adaptive Jaya optimization technique

In Adaptive Jaya optimization technique[6], consider a standard function of sphere. The objective function is used to calculate the values of Ci to reduce the value of the Sphere function.

$$\max f(C) = \text{mean} \sum_{i=1}^{m} C \quad \ldots\ldots\ldots\ldots(1)$$

The value of standard function is 0 for all Ci values of 0. To illustrate the Adaptive Jaya algorithm, Consider a population size of 5 (i.e. candidate solutions), Protocol_ type, service, flag, Dst_host_count, Dst_host_srv_count, Same_srv_rate are few of the features selected from the NSL Dataset for calculation. The table1 represents the initial population value selected from dataset. To obtain low function value, the lowest value of f(C) is taken as the worst and the highest value of f(C) is taken as the best.

Table 1: Initial population

Candidate	Protocol_ type	Service	Flag	Dst_host _count	Dst_host _srv_count	Same_srv_rate
1	1	50	1	255	10	0.04
2	1	50	1	255	1	0.01
3	1	20	2	134	86	1
4	2	15	2	3	57	1
5	1	61	3	29	86	1
F(c)	1.2	39.2	1.8	135.2	48	0.61
status				Best		Worst

From the initial population, based on the values of the objective function the best and worst solution can be identified. From the above table the Dst_host_count feature is considered as the best and the Same_srv_rate feature as worst. Here the random values selected Protocol_ type, service, flag, Dst_host_count, Dst_host_srv_count, Same_srv_rate for the first iteration are 0.32, 0.43, 0.26, 0.14, 0.17, 0.18 respectively. The values for the table 3 is calculated using the equation 2. In this if $C_{i,j,k}$ represents i^{th} variable for the j^{th} candidate in the k^{th} iteration

$C'i,j,k = Ci,j,k - (a1 * Ci,worst,k)$……(2)

Ci,worst,k->worst value of the candidate

C'i,j,k->updated value

The example for the equation2 is illustrated as

C'1,1,1= C1,1,1-(*a1* *worst)

= 1-(0.32*0.61)=0.8048

C'1,2,1= C1,2,1-(*a2*worst*)

= 50-(0.43*0.61) = 49.7377

The values for other candidates can be calculated in the same way.

The worst value selected from the table 1 is 0.61. The feature Same_srv_rate shows the worst value and the same feature is omitted from the table1 for further calculation.

Table 2

Candidate	Protocol_ type	Service	Flag	Dst_host_count	Dst_host_srv_count
1	1	50	1	255	10
2	1	50	1	255	1
3	1	20	2	134	86
4	2	15	2	3	57
5	1	61	3	29	86

The calculation for the first iteration by assigning the random values are done by using the table2 and equation 2

Table 3

Candidate	Protocol_ type	Service	Flag	Dst_hos_count	Dst_host_srv_count
1	0.8048	49.7377	0.8414	254.9146	9.8963
2	0.8048	49.7377	0.8414	254.9146	0.8963
3	0.8048	19.7377	1.8414	133.9146	85.8963
4	1.8048	14.7377	1.8414	2.9146	56.8963
5	0.8048	60.7377	2.8414	28.9146	85.8963
F(c)	1.0048	38.9377	1.6414	135.1146	47.8963
status	worst			Best	

The updated values of the candidates and the functional value f(c) in first iteration is shown in the above table3.The worst value selected from the table 3 is 1.0048. The feature Protocol_ type shows the worst value and the same feature is omitted from the table3 for further calculation.

Table 4

Candidate	Service	Flag	Dst_host_count	Dst_host_srv_count
1	49.7377	0.8414	254.9146	9.8963
2	49.7377	0.8414	254.9146	0.8963
3	19.7377	1.8414	133.9146	85.8963
4	14.7377	1.8414	2.9146	56.8963
5	60.7377	2.8414	28.9146	85.8963

The calculation for the Second iteration are done by assigning the random values to service, flag, Dst_host_count, Dst_host_srv_count in table 4 are 0.38,0.51,0.46,0.31 respectively in equation 2

Table 5

Candidate	Service	Flag	Dst_host_count	Dst_host_srv_count
1	49.35588	0.328952	254.4524	9.584812
2	49.35588	0.328952	254.4524	0.584812
3	19.35588	1.328952	133.4524	85.58481
4	14.35588	1.328952	2.452392	56.58481
5	60.35588	2.328952	28.45239	85.58481
F(c)	38.55588	1.128952	134.6524	47.58481
status		Worst	Best	

The updated values of the candidates and the functional value f(c) in Second iteration are given in Table 5 . From the above table the feature Flag is considered as the worst and the Dst_host_count as best..The values for the table 5 is calculated in the above manner.

Table 6 : The final value after updating the table 5

Candidate	Service	Dst_host_count	Dst_host_srv_count
1	49.35588	254.4524	9.584812
2	49.35588	254.4524	0.584812
3	19.35588	133.4524	85.58481
4	14.35588	2.452392	56.58481
5	60.35588	28.45239	85.58481

The same process is repeated to select the best among the 41 features.

3.2 Process Flow

In this methodology, we suggest an effective intrusion detection model with adaptive Jaya Optimization (AJO) to concurrently do parameter initialization and feature selection

Step 1: The databases which we used for intrusion detection is collected from network traffic dataset.

Step 2: Next, data analytic method is enhanced by developed algorithms for obtained dataset. For that, the dataset should be separated into training and testing.

Step 3: Adaptive Jaya Optimization (AJO) to simultaneously do parameter setting and feature selection. A faster computational convergence of the optimal solution, of the original Jaya algorithm is modified by adding a weight parameter. As a result, the Adaptive Jaya optimization better the searching ability , as well as reducing the number of the searching agents, number of iterations and computational burden.

4. Results

The execution of proposed methods has been measure by doing experiments with the NSL-KDD dataset. Then Adaptive Jaya Technique is executed on anaconda tensorflow platform. The feature selection is performed on 41 features and best 17 features were selected on best score after 24th iteration. Every iteration least score feature was eliminated. The selected best features are used as an input to the modified neural network to detect the various kinds of attacks in cyber security.

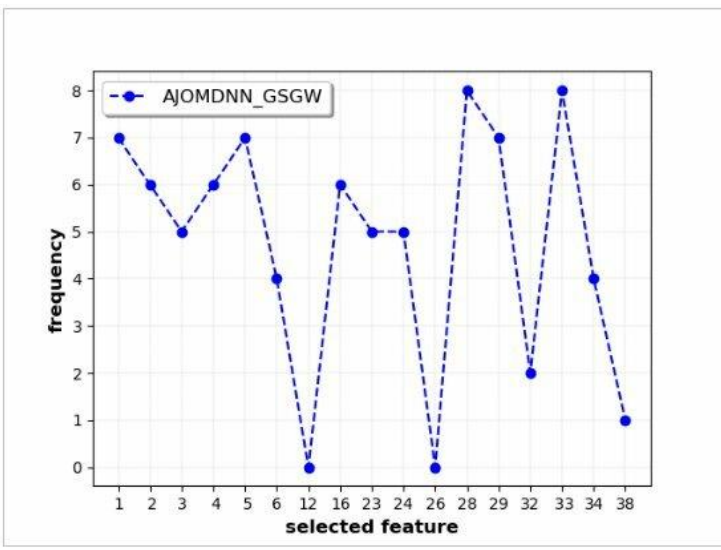

Fig. 2. Selected Features

The above Fig:2 describes the frequency of the selected features used in the proposed system. The x-axis represents the selected features from NSL- KDD dataset using AJO.

5. Acknowledgments:

"**A Machine Learning Based Hybrid Intrusion Detection System in Cyber Security**" has been a subject with tremendous scope to research upon, which leads to explore new heights in the field of Computer Science & Engineering, and its miscellaneous applications. I 'm thankful to my Research guide: Dr.Tryambak Hiwarkar & Research Co-guide:Dr.K. Ramanjaneyulu whose guidance helped me to work successfully. Their guidance will always encourage me to do work perfectly and professionally.

References

[1] Bamakan, Seyed Mojtaba Hosseini, Huadong Wang, Tian Yingjie, and Yong Shi. "An effective intrusion detection framework based on MCLP/SVM optimized by time-varying chaos particle swarm optimization." Neurocomputing 199 (2016): 90-102.

[2] Deshpande, P., Sharma, S.C., Peddoju, S.K. et al. Security and service assurance issues in Cloud environment. Int J Syst Assur Eng Manag (2018) 9: 194. https://doi.org/10.1007/s13198-016-0525-0

[3] Ashfaq, R.A.R., Wang, X.Z., Huang, J.Z., Abbas, H. and He, Y.L., 2017. Fuzziness based semi-supervised learning approach for intrusion detection system. Information Sciences, 378, pp.484-497.

[4] Masud, Mehedy, Bhavani Thuraisingham, and Latifur Khan. Data mining tools for malware detection. Auerbach Publications, 2016.

[5] Fossaceca, J.M., Mazzuchi, T.A. and Sarkani, S., 2015. MARK-ELM: Application of a novel Multiple Kernel Learning framework for improving the robustness of Network Intrusion Detection. Expert Systems with Applications, 42(8), pp.4062-4080.

[6] R. Venkata Rao Jaya: A simple and new optimization algorithm for solving constrained and unconstrained optimization problems. International Journal of Industrial Engineering Computations 7 (2016) 19–34

[7] Liu H ,Setiono R, Motoda H, Zhao Z, Feature Selection: An Ever Evolving Frontier in Data Mining, JMLR: Workshop and Conference Proceedings 10, 2010, pp. 4-13.

[8] Hee-su Chae, Byung-oh Jo, Sang-Hyun Choi, Twae-kyung Park: Feature selection for intrusion detection using NSL-KDD. In:Proceedings of the 12th WSEAS International Conference on Information Security and Privacy, pp. 184–187, Nov- 2013

Thupakula Bhaskar is currently a PhD student in Sri Satya Sai University of Technology & Medical Sciences,Bhopal,M.P. ,India.He received his M.Tech (CSE) from JNTU Hyderabad , India in 2011. His current research interests include Machine Learning , Deep Learning and intelligent optimization techniques.

Dr.Trayambak Hiwarkar was born in Maharastra , Induia in 1965. He received the B.Tech(1994) ,M.tech(1996) & Ph.D. Degree(2003) in CSE from Bundelkhand University Jhansi.He was published many papers in National / International level Journals.He has life memberships in IEEE,ACM, Institution of Engineers(India),MCSI,IETE and many more etc.

Dr.K.Ramanjaneyulu received the Ph.D. degree from CoE, Andhra University, Visakhapatnam, Andhra Pradesh, India , in 2012. He is currently working as professor in PVP Siddhartha Institute of Technology,Vijayawada,AndhraPradesh,India.He was published 18 papers in National and International Journals and 22 papers presented in various conferences National and International level.