www.ijecs.in

International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 2 February , 2014 Page No. 3873-3876

Network Intrusion Detection Using Hybrid Simplified Swarm Optimization and Random Forest Algorithm on Nsl-Kdd Dataset

S. Revathi [1] *Dr. A. Malathi*[2]

[1]*Ph.D. Research Scholar, PG and Research, Department of Computer Science,*
Government Arts College, Coimbatore-18
revathisujendran86@gmail.com

[2]*Assistant Professor, PG and Research, Department of Computer Science*
Government Arts College, Coimbatore-18
malathi.arunachalam@yahoo.com

Abstract: *During the last decade the analysis of intrusion detection has become very significant, the researcher focuses on various dataset to improve system accuracy and to reduce false positive rate based on DAPRA 98 and later the updated version as KDD cup 99 dataset which shows some statistical issues, it degrades the evaluation of anomaly detection that affects the performance of the security analysis which leads to the replacement of KDD cup 99 to NSL-KDD dataset. This paper focus on detailed analysis on NSL- KDD dataset and proposed a new technique of combining swarm intelligence (Simplified Swarm Optimization) and data mining algorithm (Random Forest) for feature selection and reduction. SSO is used to find more appropriate set of attributes for classifying network intrusions, and Random Forest is used as a classifier. In the preprocessing step, we optimize the dimension of the dataset by the proposed SSO-RF approach and finds an optimal set of features. SSO is an optimization method that has a strong global search capability and is used here for dimension optimization. The experimental results shows that the proposed approach performs better than the other approaches for the detection of all kinds of attacks present in the dataset.*

Keywords :*NSL-KDD, Simplified Swarm Optimization, PSO, Random Forest*

I. Introduction

An Intrusion Detection System is an important part of the Security Management system for computers and networks that tries to detect break-in attempts. There is no disputing fact that the number of hacking and intrusion incidents is increasing year to year as technology rolls out, unfortunately in todays interconnected Ecommerce world there is no hiding place [1]. The impetus could also be a gain, intellectual challenge, espionage, political, or just trouble-making and it exposed to a variety of intruder threats.

The first important deficiency in the KDD [8] data set is the huge number of redundant record for about 78% and 75% are duplicated in the train and test set, respectively. Which makes the learning algorithm biased, that makes U2R more harmful to network. To solve these issues a new version of KDD dataset, NSL-KDD is publicly available for researchers through our website. Although, the data set still suffers from some of the problems discussed by McHugh [2] and may not be a perfect representative of existing real networks, because of the lack of open data sets for network-based IDSs, we believe it still can be applied as an effective benchmark data set to help researchers compare different intrusion detection methods.

The NSL- KDD dataset used for intrusion detection is a raw data which highly susceptible to noise, missing values and inconsistency [3]. To improve data efficiency feature reduction and filtering technique is needed, As a result the paper proposed a novel simplified swarm optimization incorporates with Random forest classifier for pre-processing, to mine raw data. Data mining provide decision support for intrusion management, and also help IDS for detecting new vulnerabilities and intrusions by discovering unknown patterns of attacks or intrusions.

The rest of the paper is structured as follows: section II present some related work based on intrusion detection research. Section III explains detailed description of the attacks present in NSL-KDD dataset. Section IV summarize in detail about proposed work of feature selection with classification algorithm. The result and analysis shown in section V and the conclusion is summarized in section VI.

II. Related Work

Intrusion Detection Systems gross raw network information or audit records as input that ends up in a large network traffic data size and the invisibility of intrusive patterns which are normally hidden among the irrelevant and redundant features to identify it as normal or attack. A new collaborating filtering technique for pre-processing the probe type of attacks is proposed by G. Sunil Kumar, [4] based on hybrid classifiers on binary particle swarm optimization and random forests algorithm for the classification of probe attacks in a network. Fernando [5] Used n-gram theory to identify redundant subsequence and proposed Hidden Markov Model for service selection to reduce audit data significantly. Wei-Chang yeh et.al [6] proposed new method by combining SSO with weighted exchange local search method for intrusion detection.

The inherent problem of KDD dataset leads to new version of NSL KDD dataset that are mentioned in [7, 8]. It is very difficult to signify existing original networks, but still it can be applied as an effective benchmark data set for researchers to compare different intrusion detection methods [2]. In [8] they have conducted a statistical analysis on this data set and found two important issues which highly affect the performance of evaluated system, and results in very poor evaluation of anomaly detection approaches. To solve these issues, they proposed a new dataset, NSL-KDD, which consists of only selected records form the complete KDD dataset and does not suffer from any of the mentioned shortcomings.

Data mining [14] and machine learning technology has been extensively applied in network intrusion detection and prevention system by discovering user behavior patterns from the network traffic data.

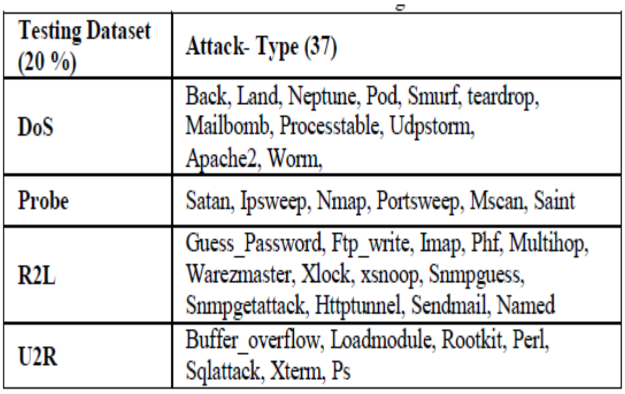
III. Dataset Description

The statistical analysis showed that there are important issues in the KDD data set [19] which highly affects the performance of the systems, and results in a very poor estimation of anomaly detection approaches. To solve these issues, a new data set as, NSL-KDD [7] is proposed, which consists of selected records of the complete KDD data set. The advantage of NSL KDD dataset are

1. No redundant records in the train set, so the classifier will not produce any biased result
2. No duplicate record in the test set which have better reduction rates.
3. The number of selected records from each difficult level group is inversely proportional to the percentage of records in the original KDD data set.

The training dataset is made up of 21 different attacks out of the 37 present in the test dataset. The known attack types are those present in the training dataset while the novel attacks are the additional attacks in the test dataset i.e. not available in the training datasets. The attack types are grouped into four categories: DoS, Probe, U2R and R2L. Table 1 shows the major attacks in both training and testing dataset [9].

Table I Attacks in Testing Dataset



Testing Dataset (20 %)	Attack- Type (37)
DoS	Back, Land, Neptune, Pod, Smurf, teardrop, Mailbomb, Processtable, Udpstorm, Apache2, Worm,
Probe	Satan, Ipsweep, Nmap, Portsweep, Mscan, Saint
R2L	Guess_Password, Ftp_write, Imap, Phf, Multihop, Warezmaster, Xlock, xsnoop, Snmpguess, Snmpgetattack, Httptunnel, Sendmail, Named
U2R	Buffer_overflow, Loadmodule, Rootkit, Perl, Sqlattack, Xterm, Ps

Fig 1 and 2 explains about the analysis of NSL KDD dataset in detail and shows the number of individual records in four types of attacks for both training and testing.

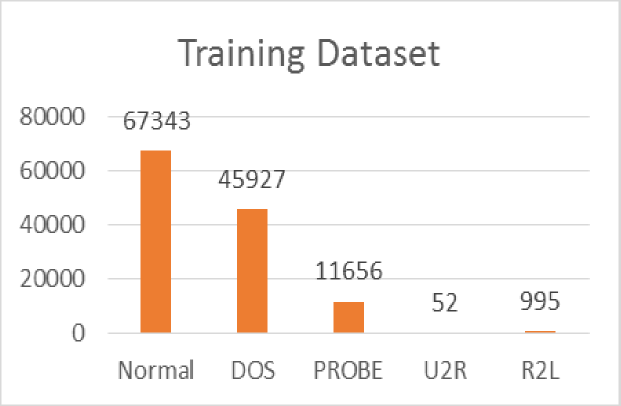


Figure 1. Number of Instance in Training Dataset

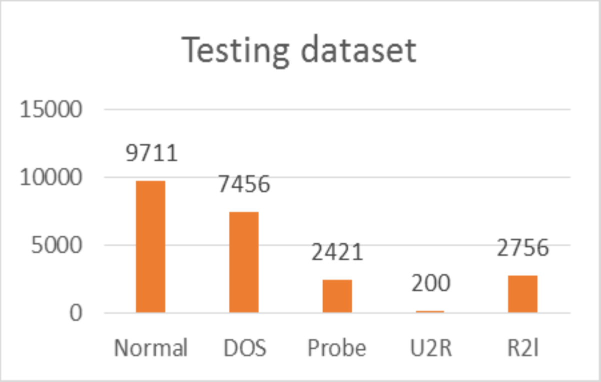


Figure 2. Number of Instance in Testing Dataset

III Proposed Work

3.1 Preprocessing Stage

In each connection there are 41 attributes describing different features of the connection [18] and a label assigned to each either as an attack type or as normal [20]. In this paper, a new proposed model namely simplified swarm optimization (SSO) is introduced. SSO is a simplified version of PSO and can be used to find the global minimum of nonlinear functions [10]. This approach is used to reduce dimensionality of dataset.

The proposed SSO-RF algorithm is presented below.

Step 1: Initialize the swarm size (m), the maximum generation (maxGen), the maximum fitness Value (maxFit), Cw, Cp and Cg.

Step 2: In every iteration, a random number R that is in the range of 0 and 1 will be randomly generated for each dimension.

Step 3: Perform the comparison approach as:

If (0 ≤ R < Cw), then {xid = xid};

Else if (Cw ≤ R < Cp), then {xid = pid};

Else if (Cp ≤ R < Cg), then {xid = gid};

Else if (Cg ≤ R ≤ 1), then {xid = new (xid)};

Algorithm	No.of Features Reduced	Records Filtered (approx. in %)	Accuracy
PSO-RF	32	10	91.5
SSO	29	15	94.2
SSO-RF	13	40	98.72

Class Names	Test Accuracy (%) with 41Features	Test Accuracy (%) with 13 Features
Normal	99.1	99.8
DOS	98.8	99.5
Probe	96.1	98.6
U2R	95.6	97.6
R2l	95.1	98.1

Step 4: Choose m variables which used to split each node. m<<M, where M is the number of input variables.

Step 5: In a growing tree at each and every node select m variables at random from M and bust them out to have the best split.

Step 6: This process will be repeated until the termination condition is satisfied.

The proposed SSO-RF method filter raw data which reduce dimensionality problem for both discrete and continuous variables in dataset [11]. This approach is significantly different from other research work which had combine only data mining and PSO. The proposed method yield high accuracy and achieves near optimal solution for pre-processing phase.

3.2 Classification Algorithm

3.2.1 Random Forest

Random Forest for each Decision Tree can be built by randomly sampling a feature subset. By injecting randomness at each node of the grown tree [12], it has improved accuracy. The correlation between trees is reduces by randomly selecting the features which improves the prediction power and results in higher efficiency. As such the advantages of Random Forest are [11]:

- Overcoming the problem of over fitting
- In training data, they are less sensitive to outlier data
- Parameters can be set easily and therefore, eliminates the need for pruning the trees variable importance and accuracy is generated automatically

Random Forest not only keeps the benefits achieved by the Decision Trees but through the use of bagging on samples, its voting scheme through which decision is made and a random subsets of variables, it most of the time achieves better results than Decision Trees [13]. It can easily handle high dimensional data modelling such as missing values and can handle continuous, categorical and binary data [14]. The bootstrapping and ensemble scheme makes Random Forest strong enough to overcome the problems of over fitting and hence there is no need to prune the trees.

IV. Experimental Result

This section describes the experimental results and performance evaluation of the proposed system. For experimental simulation NSL KDD data, which is widely used for evaluating intrusion detection system is used. The proposed system can easily filters and reduce large scale dataset.

PSO [15, 17], SSO [16] and proposed SSO-RF algorithms were applied to all 41 features as input for IDS to reduce the dimension of the dataset and later for filtering of records to improve detection accuracy and to classify a network traffic as normal or attack behavior. The results for the three optimization methods are presented in Table II. The parameters compared are number of features, Records filtered, accuracy.

Table II. Overall comparison of three methods

Table III Test Accuracy of SSO-RF with different type of attacks

Table III shows the test accuracy that achieved by SSO-RF for different types of attacks that compared with 41 features and with the reduced set of features. The below figure 3 explains test accuracy of SSO-RF method

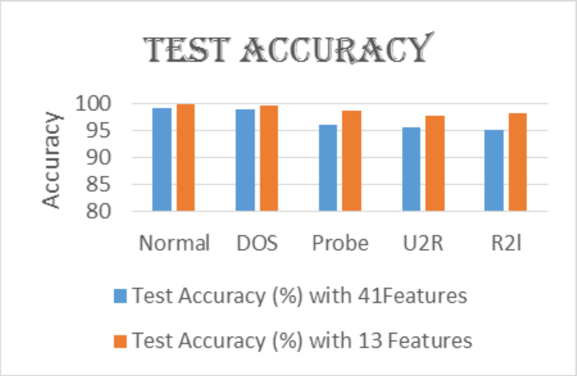


Fig 3. Test Accuracy of SSO-RF Method

Hence the proposed SO-RF algorithm shows the highest accuracy compared with other two method with and without feature reduction.

VI. Conclusion and Future Work

In this paper, we have analyzed the NSL-KDD dataset that solves some of the snags of KDD99 dataset. Our analysis shows that NSL-KDD dataset is very ideal for comparing different intrusion detection models. Using all the 41 features in the network to evaluate the intrusive patterns may leads to time consuming detection and also the performance degradation of the system. Some of the features in this are redundant and irrelevant for the process. We have used the proposed SSO-RF technique for reduce the dimensionality of the data. Our experiment has been carried out with different optimization algorithms for the dataset with and without feature reduction and in that proposed SSO-RF shows a high test accuracy compared to all other algorithms in both the cases. So in the case of reduced feature set this analysis shows that SSO-RF is speeding up the training and the testing methods for intrusion detection that is very essential for the network application with a high speed and even providing utmost testing accuracy. In future we can try to use fuzzy logic with optimization technique to build an efficient intrusion detection system.

Reference

1. Abraham A, Grosan. C, Vide. C.M, (2007) "Evolutionary design of intrusion detection programs", International Journal of Network Security 4 (March (3)) 328–339.

2. J. McHugh, "Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory," ACM Transactions on Information and System Security, vol. 3, no. 4, pp. 262–294, 2000.

3. Deris tiawan, Abdul Hanan Abdullah, Mohd. Yazid dris, (January 2011), "Characterizing Network Intrusion Prevention System", International Journal of Computer Applications (0975 – 8887), Volume (14– No.1).

4. Sunil Kumar. G, Sirisha C.V.K, Kanaka Durga.R, Devi. A, (January 2012), "Robust Preprocessing and Random Forests Technique for Network Probe Anomaly Detection", International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, (Volume-1, Issue-6).

5. Fernando Godinez and Dieter Hutter and Rahul Monroy, (2005), "Service Discrimination and Audit File Reduction for Effective Intrusion Detection", Proceedings of the 5th international conference on Information Security Applications, Springer-Verlag Berlin, Pages 99-113.

6. Changseok bae, Wei-Chang yeh, Noorhaniza wahid,yuk ying chung and yao liu, "A new simplified swarm optimization (sso) using exchange local search scheme". ICIC International issn 1349-4198. Volume (8- No.6), June 2012.

7. "Nsl-kdd data set for network-based intrusion detection systems." Available on: http://nsl.cs.unb.ca/KDD/NSL-KDD.html, March 2009.

8. Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set", In the Proc. Of the IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009), pp. 1-6, 2009.

9. Vipin Kumar, Himadri Chauhan, Dheeraj Panwar, "K-Means Clustering Approach to Analyze NSL-KDD Intrusion Detection Dataset", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-3, Issue-4, September 2013.

10. Yeh W.C, Chang W.W, Chung Y.Y, (2009) ,"A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method", Expert System with Applications 36 (May (4)) 8204–8211.

11. S.Revathi, A.Malathi," Optimization of KDD Cup 99 Dataset for Intrusion Detection Using Hybrid Swarm Intelligence with Random Forest Classifier", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, July 2013.

12. L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

13. Bosh, A., Zisserman, A., Munoz, and X.: "Image classification using Random Forests and ferns". *In: IEEE ICCV* 2007.

14. Ned Horning, "Introduction to Decision Trees and Random Forests", American Museum of Natural History's.

15. Kennedy J, Eberhart R. "Particle swarm optimization". Proceedings of the IEEE international conference on neural networks (Perth, Australia), 1942–1948. Piscataway, NJ: IEEE Service Center; 1995.

16. Changseok bae, Wei-Chang yeh, Noorhaniza wahid,yuk ying chung and yao liu, "A new simplified swarm optimization (sso) using exchange local search scheme". *ICIC International @ 201*2 issn 1349-4198. Volume 8, number 6, June 2012.

17. Chen.G Chen.Q,Guo.W, "A PSO-based approach to rule learning in network intrusion detection", in: Advances in Soft Computing, (vol. 40), Springer, Berlin- Heidelberg, 2007, and pp. 666–673.

18. Sanoop Mallissery, Sucheta Kolekar, Raghavendra Ganiga, "Accuracy Analysis of Machine Learning Algorithms for Intrusion Detection System using NSL-KDD Dataset", Proc. of the Intl. Conf. on Future Trends in Computing and Communication -- FTCC 2013.

19. KDD Cup 1999. Available on http://kdd.ics.uci.edu/ Databases/kddcup 99/kddcup99.html, Ocotber 2007.

20. Dharmendra G. Bhatti, P. V. Virparia, Bankim Patel, (November 2012), "Data Preprocessing for Reducing False Positive Rate in Intrusion Detection", International Journal of Computer Applications (0975 – 8887) Volume 57– No.5.