A Novel Intrusion Detection System using Data Mining Techniques

[1] Dr. M.Manimekalai, [2] G. Anupriya

[1]Professor, Director and Head, [2]Research Scholar
[1,2]Department of Computer Science
[1,2]Shrimati Indira Gandhi College, Tiruchirappalli, Tamil Nadu, India-620002

Abstract: Intrusion detection systems (IDS) in MANET have to system blocks of packets with many features, which suspend the detecting of anomalies. Sampling and Feature Selection may be employed to depreciate computing time and hence diminishing the time of intrusion detection. A Novel Hybrid Feature Selection method mounts on Particle Swarm Optimization (PSO) and Information Gain analysis. The execution of the proposed Hybrid Feature Selection Method on KDD CUP dataset to decrease the volume of primary features and accurate by implementing better detection performance in the classification methods relating with other feature selectors. The relevant features and removing redundant features of KDD CUP dataset is Optimal Dataset. ANN with Multi-Layered Perceptron classification method was used to classify the nodes of MANET**.**

***IndexTerms* - Intrusion Detection System, Mobile Ad Hoc Network, Data Mining Technique, Artificial Neural Network, Naïve Bayes classification.**

I. INTRODUCTION

The world is becoming more interconnected with the advent of the Internet and new networking technology [1]. There is a large amount of personal, commercial, military, and government information on networking infrastructures worldwide. Network security is becoming of great importance because of intellectual property that can be easily acquired through the internet. Network security starts with authorization, commonly with a username and a password. Network security consists of the provisions and policies adopted by a network administrator to prevent and monitor unauthorized access, modification in system, misuse, or denial of a computer network and network-accessible resources [2]. Basically network security involves the authorization of access to data in a network, which is controlled by the network admin. It has become more important to personal computer users, and organizations. If this authorized, a firewall forces to access policies such as what services are allowed to be accessed for network users. So that to prevent unauthorized access to system, this component may fail to check potentially harmful content such as computer worms or Trojans being transmitted over the network [3]. Anti-virus software or an intrusion detection system (IDS) helps detect the malware. Today anomaly may also monitor the network like wire shark traffic and may be logged for audit purposes and for later on high-level analysis in system. Communication between two hosts using a network may be uses encryption to maintain privacy policy [4].

II. IMPORTANCE OF NETWORK SECURITY

System and network technology are a key technology for a wide variety of applications. Security is crucial to networks and applications. Although, network security is a critical requirement in emerging networks, there is a significant lack of security methods that can be easily implemented.

There exists a “communication gap” between the developers of security technology and developers of networks. Network design is a well- developed process that is based on the Open Systems Interface (OSI) model [5]. The OSI model has several advantages when designing networks. It offers modularity, flexibility, ease- of- use, and standardization of protocols. The protocols of different layers can be easily combined to create stacks which allow modular development [6]. The implementation of individual layers can be changed later without making other adjustments, allowing flexibility in development. In contrast to network design, secure network design is not a well-developed process. There isn’t a methodology to manage the complexity of security requirements. Secure network design does not contain the same advantages as network design [7] [19] [20]. When considering network security, it must be emphasized that the whole network is secure. Network security does not only concern the security in the computers at each end of the communication chain. When transmitting data the communication channel should not be vulnerable to attack. A possible hacker could target the communication channel, obtain the data, decrypt it and reinsert a false message. Securing the network is just as important as securing the computers and encrypting the message. When developing a secure network, the following need to be considered:

1. *Access* – authorized users are provided the means to communicate to and from a particular network [8].
2. *Confidentiality* – Information in the network remains private.
3. *Authentication* – Ensure the users of the network are who they say they are.
4. *Integrity* – Ensure the message has not been modified in transit.
5. *Non-repudiation* – Ensure the user does not refute that he used the network.

An effective network security plan is developed with the understanding of security issues, potential attackers, needed level of security, and factors that make a network vulnerable to attack [9][10] [21][22][23].

III. LITERATURE REVIEW

Dutta, Sharmishtha, Tanjila Mawla, and Md Forhad Rabbi [11] Network intrusion means any attempt to compromise the confidentiality or availability of a computer network. The growing speed of data transmission poses a challenge in the detection of such intrusions. Most of the existing systems for detecting network intrusion ignore a significant feature associated with every sort of data, time. On the other hand, some systems have taken the temporal aspect of data into account. These systems show better accuracy, low false alarm rates, higher bandwidth with full coverage rate, and system availability in case of data flow rate close to 1 GB/s. The concept of temporal data mining in network intrusion detection is being more popular as it is providing more promising results than traditional mining techniques.

Aljawarneh, Shadi, Monther Aldwairi, and Muneer Bani Yassein [12] this paper develops a new hybrid model that can be used to estimate the intrusion scope threshold degree based on the network transaction data's optimal features that were made available for training. The experimental results revealed that the hybrid approach had a significant effect on the minimisation of the computational and time complexity involved when determining the feature association impact scale. The accuracy of the proposed model was measured as 99.81% and 98.56% for the binary class and multiclass NSL-KDD data sets, respectively.

Sharma, Ruby, and Sandeep Chaurasia [13] In this approach, cluster efficiency is improved through a membership matrix generation (MMG) algorithm. Dissimilarity Distance Function (DDF) has been used to compute the distance metric while creating a cluster in proposing an IDS. The proposed enhanced fuzzy c-means algorithm has been tested upon ADFA Dataset and the model performs highly appreciable in terms of accuracy, precision, detection rates, and false alarms.

Saxena, Akash, Khushboo Saxena, and Jayanti Goyal [14] In our proposed work, we initially apply KDD cup'99 dataset which is most broadly used method for detecting intrusion. DBSCAN is the most utilized method which is used to eliminate noise from the data. Then, we generate the most meaning inputs by analyzing and processing whole data which is done by the selection of feature method. K-means clustering performs grouping of data which is followed by SMO classifier. So we proposed a hybrid structure which improves the taken as a whole accuracy. MATLAB and WEKA tools are used to execute the whole process.

Gupta, Amara SALG Gopal, G. Syam Prasad, and Soumya Ranjan Nayak [15] In this way supported learning we'd quite recently like the topic that can assist us with providing extra security on Data Storage. To reduce the attack risk, a dynamic key theory is bestowed and analyzed we've an inclination to face live about to planned theme for extra security that is ready to be secure delicate information of fluctuated domains like in consideration area enduring associated information like contact points of interest and antiquity.

IV. FEATURE SELECTION IN PRE-PROCESSING

Feature Selection (FS) is a commonly used step in machine learning, especially when dealing with a high dimensional space of features. The objective of FS is to simplify a data set by reducing its dimensionality and identifying relevant underlying features without sacrificing predictive accuracy. By doing that, it also reduces redundancy in the information provided by the selected features. In real world problems FS is a must due to the abundance of noisy, irrelevant or misleading features. FS is extensive and it spreads throughout many fields, including text categorization, data mining, pattern recognition, signal processing and intrusion detection [1] [2] [16][17][18].

Feature selection Definition: A "feature" or "attribute" refers to an aspect of the data. Usually before collecting data, features are specified or chosen. Features can be discrete, continuous, or nominal. Generally, the features are characterized as:

1. Relevant: features which have an influence on the output and their role cannot be assumed by the rest.

2. Irrelevant: Irrelevant features are defined as those features not having any influence on the output, and whose values are generated at random.

3. Redundant: A redundancy exists whenever a feature can take the role of another. Problem of selecting some subset of a learning algorithms input variables upon which it should focus attention, while ignoring the rest. Feature selection [3] is the process of selecting the best feature among all the features. Because all the features are not useful in constructing the clusters; some features may be redundant or irrelevant thus not contributing to the learning process. An important stage of pre-processing and is one of two ways of avoiding the curse of dimensionality (the other is feature extraction). The main aim of feature selection is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features. In many real-world problems, Feature selection is a must due to the abundance of noisy, irrelevant or misleading features. For instance, by removing these factors, learning from data techniques can benefit. To be completely sure of the attribute election, we would ideally have to test all the enumerations of attribute subsets, which is infeasible in most cases as it will result in 2n subsets of n attributes [3].

4.1Population and Sample

- It reduces the dimensionality of the feature space, to limit storage requirements and increase algorithm speed.
- It removes the redundant, irrelevant or noisy data.
- It speeds up the running time of the learning algorithms.
- Reduce computational complexity of running algorithm.
- Improving the data quality.
- Increasing the accuracy of the resulting model.
- Feature set reduction, to save resources in the next round of data collection or during utilization;
- Performance improvement, to gain in predictive accuracy;
- Data understanding, to gain knowledge about the process that generated the data or simply visualize the data.

3.2 Approaches in Feature Selection

There are two approaches used in Feature selection:

1. Forward Selection: Start with no variables and add them one by one, at each step adding the one that decreases the error the most, until any further addition does not significantly decrease the error.

2. Backward Selection: Start with all the variables and remove them one by one, at each step removing the one that decreases the error the most (or increases it only slightly), until any further removal increases the error significantly. To reduce over fitting, the error referred to above is the error on a validation set that is distinct from the training set.

V. METHODOLOGY

1.5.1 Classification Step: Artificial Neural Network

ANN is an effective calculating system whose principal theme has acquired from the resemblance of biological NN. ANNs has also described as "Parallel Distributed Processing Systems." ANN obtains a massive number of units that are interrelated in some pattern to enable connection among the units. Those units also mentioned to as neurons or nodes are mere CPUs which operate in parallel. In this work, the ANN has adopted for classification of Intrusion Detection in the system. The following method depicts the steps of the MLP-NN training algorithm.

Step 1: *Initialize Bias, Learning rate* α*, weights, to begin the training of Multi-Layered Perceptron Neural Network. For simplicity and calculation, need to set weight =0 and bias* $\alpha = 1$.

Step 2: *Proceed step 3-8 at the terminating condition is true.*

Step 3: *Proceed step 4-6 for all training vector a.*

Step 4: *Initiate each input as follows:*

$$r_j = s_j \ (j = 1 \ to \ m)$$

Step 5: *Get the net input with the next relations*

$$s_{jn} = b + \sum_{j}^{m} r_j w_{jk}$$

Here bias is given as b, and the whole amount of input neuron is given by 'n'.

Step 6: *Apply the activation function to obtain the final output for each input unit k=1 to n*

$$f(s_{jm}) = \begin{cases} 1 \quad if \ s_{jmk} > \theta \\ 0 \ if -\theta \le s_{jmk} \le \theta \\ -1 \ if \ s_{jmk} < -\theta \end{cases}$$

Step 7: *Adjust the weight and bias for r=1 to m and k=1 to n as follows:*

Step 7.1: Case 1: if $s_k \neq t_k$ *the m*

$$w_{jk}(new) = w_{jk}(old) + \propto t_k \, r_j$$

$$s_k(new) = s_k(old) + \propto t_k$$

Step 7.2: Case 2: if $s_k = t_k$ *then*

$$w_{jk}(new) = w_{jk}(old)$$

$$s_k(new) = s_k(old)$$

Here 's' is the exact output, and 't' is the desired/target output.

Step 8: *Testing for terminating condition, which will occur while there is no variation in weight.*

VI. RESULTS AND DISCUSSION

6.1 Number of Features obtained by Proposed Feature Selection Method

Following table 8.1 provides the outcome attained by the proposed Hybrid Feature Selection method and current filter-based feature selection techniques like Information Gain and Particle Swarm Optimization. From table 8.1, Particle Swarm Optimization filters 31 features, Information Gain screens only 27 features, and the proposed Hybrid Feature Selection gives only 20 features. To assess the competence of the proposed Hybrid Feature Selection and other approaches by consuming classification techniques like Artificial Neural Network (ANN). The assessment of metrics is like Accuracy, Error rates, True Positive Rate, False Positive Rate, Precision, Recall and ROC curves.

Table 1: Number of Features obtained by using Information Gain, Particle Swarm Optimization and Proposed Hybrid Feature Selection Method

Sl.No	**Information Gain**	**Particle Swarm Optimization**	**Proposed Hybrid Feature Selection Method**
1	num_failed_32s	urgent	Protocol_type
2	srv_diff_host_rate	Wrong_fragment	diff_srv_rate
3	hot	num_compromised	rerror_rate
4	srv_serror_rate	same_srv_rate	srv_serror_rate
5	dst_host_srv_diff_host_rate	diff_srv_rate	srv_rerror_rate
6	same_srv_rate	count	Service
7	rerror_rate	dst_host_srv_diff_host_rate	dst_host_diff_srv_rate
8	logged_in	srv_count	dst_host_count
9	dst_host_srv_serror_rate	dst_host_same_src_port_rate	dst_host_srv_rerror_rate
10	count	dst_host_diff_srv_rate	dst_host_serror_rate
11	dst_host_srv_rerror_rate	dst_host_count	Src_bytes
12	Service	dst_host_rerror_rate	dst_host_srv_count
13	Dst_bytes	dst_host_srv_count	srv_diff_host_rate
14	Src_bytes	dst_host_serror_rate	srv_diff_host_rate
15	dst_host_same_srv_rate	dst_host_srv_serror_rate	dst_host_srv_diff_host_rate
16	dst_host_same_srv_rate	logged_in	serror_rate
17	dst_host_diff_srv_rate	is_guest_32	dst_host_same_src_port_rate
18	srv_Count	Dst_bytes	srv_Count
19	dst_host_serror_rate	Src_bytes	dst_host_srv_serror_rate
20	dst_host_same_src_port_rate	dst_host_same_srv_rate	Dst_bytes
21	dst_host_count	dst_host_srv_rerror_rate	
22	srv_rerror_rate	Service	
23	diff_srv_rate	hot	
24	serror_rate	srv_rerror_rate	
25	is_guest_32	Flag	
26	Protocol_type	rerror_count	
27	num_compromised	srv_serror_rate	
28		serror_rate	
29		Protocol_type	
30		srv_diff_host_rate	
31		num_failed_32s	
32		land	

6.2 Result and Discussion on the ANN Classification

Mostly the classification methods are utilized to evaluate the effectiveness of the results obtained from the feature selection techniques. In this chapter, the classification technique called Artificial Neural Network has utilized to classify the node as Malicious Node and Legitimate node. The following metrics like classification accuracy, Kappa Statistic value, reduced error rates like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Root Relative Squared Error (RRSE) and Relative Root Absolute Error (RRAE), True Positive Rate (TPR), False Positive Rate (FPR), Precision, Recall, F-Measure and ROC Area are used to evaluate the proposed HRLR-FS method and original dataset.

Table 2: Performance analysis of original dataset and Proposed Hybrid HRLR-FS method by using Artificial Neural Network

Performance Metrics	Original Dataset	Proposed HRLR-FS Method
Classification Accuracy	69.3333 %	98 %
Kappa Statistic	0.5539	0.7594
Mean Absolute Error (MAE)	0.0411	0.0788
Root Mean Squared Error (RMSE)	0.1317	0.1311
Root Relative Squared Error (RRSE)	91.7907 %	73.7135 %
Relative Root Absolute Error (RRAE)	89.9791 %	60.0193 %
TP Rate	0.693	0.98
FP Rate	0.15	0.355
Precision	0.553	0.98
Recall	0.693	0.98
F-Measure	0.601	0.978
ROC Area	0.847	1

Table 2 gives the Performance analysis of original dataset and Proposed Hybrid HRLR-FS method by using Artificial Neural Network. Figure 1 depicts the Accuracy, Root Relative Squared Error (RRSE) and Relative Root Absolute Error (RRAE) values for original dataset and proposed HRLR-FS processed dataset. From the above table 2, ANN gives better result in terms of accuracy, Kappa statistic, error rates, TPR, FPR, Precision, Recall, F-Measure and ROC area for proposed HRLR-FS processed dataset than the original dataset.

VII. CONCLUSION

In this research work, the number of features in the dataset has reduced in the pre-processing step. With the reduced dataset, the classification accuracy has increased by using the ANN classification method. The error rates are reduced and true positive, ROC values have increased. Association Rule Mining method has used to generate the rule structure. PSO gives the perfect rule structure than by using ARM. In the future, the risk severity of the malicious node will be predicted by using Decision-making methods.

REFERENCES

[1] Yao, Xuanxia, *et al.*, “A lightweight multicast authentication mechanism for small scale IoT applications”, *IEEE Sensors Journal*, Vol.13, No. 10, pp. 3693-3701, 2013.

[2] S. A. Joshi and Varsha S. Pimprale, "Network Intrusion Detection System (NIDS) based on data mining," *International Journal of Engineering Science and Innovative Technology (IJESIT)*, Vol. 2, No. 1, pp. 95-98, 2013.

[3] Elhag, Salma, et al, “On the combination of genetic fuzzy systems and pairwise learning for improving detection rates on intrusion detection systems,” *Expert Systems with Applications*, Vol.42, No.1, pp.193-202, 2015.

[4] Eesa, Adel Sabry, Zeynep Orman and Adnan Mohsin Abdulazeez Brifcani, “A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems,” *Expert Systems with Applications* ,Vol.42, No.5, pp.2670-2679.2015.

[5] Yanfang Ye, *et al., “*DeepAM: a heterogeneous deep learning framework for intelligent malware detection,” *KnowledgeandInformationSystems*, Vol. 54, No.2, pp.265-285, 2018.

[6] Cabaj, Krzysztof, Marcin Gregorczyk and Wojciech Mazurczyk, “Software-defined networking-based crypto ransomware detection using HTTP traffic characteristics,” *Computers & Electrical Engineering*, Vol. 66, pp.353-368, 2018.

[7] Demertzis, Konstantinos and Lazaros Iliadis, “A hybrid network anomaly and intrusion detection approach based on evolving spiking neural network classification,” *International* Conference on e-Democracy, *Springer, Cham*, 2013.

[8] M. Elbasiony, Reda, *et al*, “A hybrid network intrusion detection framework based on random forests and weighted k-means,” *Ain Shams Engineering Journal*, Vol. 4, No. 4, pp.753-762, 2013.

[9] Iftikhar Ahmad, *et al., “*Enhancing SVM performance in intrusion detection using optimal feature subset selection based on genetic principal components,” *Neural computing and applications*, Vol.24, No.7-8, pp.1671-1682, 2014.

[10] Eesa, Adel Sabry, Zeynep Orman and Adnan Mohsin Abdulazeez Brifcani, “A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems,” *Expert Systems with Applications* ,Vol.42, No.5, pp.2670-2679.2015.

[11] Dutta, Sharmishtha, Tanjila Mawla, and Md Forhad Rabbi. "A Comparison Study of Temporal Signature Mining Over Traditional Data Mining Techniques to Detect Network Intrusion." *Emerging Technologies in Data Mining and Information Security*. Springer, Singapore, 2019. 757-763.

[12] Aljawarneh, Shadi, Monther Aldwairi, and Muneer Bani Yassein. "Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model." *Journal of Computational Science* 25 (2018): 152-160.

[13] Sharma, Ruby, and Sandeep Chaurasia. "An enhanced approach to fuzzy C-means clustering for anomaly detection." *Proceedings of First International Conference on Smart System, Innovations and Computing*. Springer, Singapore, 2018.

[14] Saxena, Akash, Khushboo Saxena, and Jayanti Goyal. "Hybrid Technique Based on DBSCAN for Selection of Improved Features for Intrusion Detection System." *Emerging Trends in Expert Applications and Security*. Springer, Singapore, 2019. 365-377.

[15] Gupta, Amara SALG Gopal, G. Syam Prasad, and Soumya Ranjan Nayak. "A New and Secure Intrusion Detecting System for Detection of Anomalies Within the Big Data." *Cloud Computing for Geospatial Big Data Analytics*. Springer, Cham, 2019. 177-190.

[16] Poornappriya, T. S., and M. Durairaj. "High relevancy low redundancy vague set based feature selection method for telecom dataset." *Journal of Intelligent & Fuzzy Systems*, Preprint: 1-18.

[17] M. Durairaj, T S Poornappriya, "Choosing a spectacular Feature Selection technique for telecommunication industry using fuzzy TOPSIS MCDM.", *International Journal of Engineering & Technology*, 7 (4) (2018) 5856-5861.

[18] M. Durairaj, T. S. Poornappriya, "Importance of MapReduce for Big Data Applications: A Survey", *Asian Journal of Computer Science and Technology*, Vol.7 No.1, 2018, pp. 112-118.

[19] M. Lalli, V.Palanisamy,(2016), "Filtering Framework for Intrusion Detection Rule Schema in Mobile Ad Hoc Networks", International Journal of Control Theory and Applications –(IJCTA),9(27), pp. 195-201, ISSN: 0974-5572

[20] M. Lalli, V.Palanisamy,(2017), "Detection of Intruding Nodes in Manet Using Hybrid Feature Selection and Classification Techniques", Kasmera Journal, ISSN: 0075-5222, 45(1) (SCIE)(Impact Factor:0.071).

[21] M. Lalli, V.Palanisamy, (Sep 2014), "A Novel Intrusion Detection Model for Mobile Adhoc Networks using CP-KNN", International Journal of Computer Networks & Communications- (IJCNC), Vol.6, No.5, ISSN:0974-9322.

[22] M. Lalli, "Statistical Analysis on the KDD CUP Dataset for Detecting Intruding Nodes in MANET", *Journal of Applied Science and Computations,* Volume VI, Issue VI, JUNE/2019, 1795-1813.

[23] M. Lalli, "Intrusion Detection Rule Structure Generation Method for Mobile Ad Hoc Network", *Journal of Emerging Technologies and Innovative Research*, June 2019, Volume 6, Issue 6, 835-843.