Intrusion Detection System by Combined Feature Selection Algorithm with Artificial Neural Network Classification

[1]N. Devijeyakkalyani , [2]Dr. D. Ravindran
[1]M.Phil. Scholar, [2]Associate Professor
[1]Dept.of Computer Science
[1]St.Joseph's College, Tiruchirappalli, Tamil Nadu, India.

***Abstract*:** The emerging trend of ubiquitous and pervasive computing aims at embedding everyday devices such as wristwatches, smart phones etc. with microprocessors and imparts them with wireless communication capability. Due to the scattering of system accessibility, the enthusiasm for system security and protection against advanced ambushes is consistently growing. Intrusion Detection System (IDS) play out a central occupation in the present system security. This paper proposes an IDS dependent on highlight determination and bunching algorithm utilizing filter and wrapper techniques. Filter and wrapper methodologies are named feature grouping based on linear correlation coefficient (FGLCC) algorithm and cuttlefish algorithm (CFA), separately. Artificial Neural Network is used as the classifier in the proposed technique. For execution confirmation, the proposed system was associated on KDD Cup 99 enormous informational indexes. The outcomes established a high accurate and detection rate with a low false positive rate variance with the current techniques in the writing.

IndexTerms - **Intrusion Detection, KDD, Artificial Neural Network, CuttleFish Algorithm (CFA), Linear Correlation Coefficient.**

I. INTRODUCTION

With the fast extension of the PC network during the previous couple of years, the data security issue turns out to be increasingly significant. There are many research subjects for network security [1] [2]. Like as, data encryption, weakness database, intrusion detection, and so on. Intrusion detection is one of the real data security issues. Intrusion Detection System help the framework in opposing outside assaults. Existing IDS can be separated into two classifications as indicated by the detection draws near: peculiarity detection and abuse detection or mark detection [2][3]. Data mining systems can be utilized for abuse and peculiarity intrusion detection. Abuse alludes to known assaults and destructive exercises that endeavour the known sensitivities of the framework. In abuse detection, each occasion in a data set is named as ―normal‖ or‖ intrusion‖ and a learning algorithm is prepared over the marked data. As there are many numbers of ID procedures utilizing data mining strategies, the obscure method and framework could be thought of as a standard for future prospect. Therefore, the reason for this paper is to survey related papers of utilizing data mining for intrusion detection. The commitment of this exploration paper is to give a correlation of IDS regarding data mining IDS procedures utilized for future research bearings [4][5].

An Intrusion Detection System is use to differentiate a wide range of malicious network trading and PC utilization that can't be identified by a standard firewall [6] [7]. It incorporates network assaults against delicate administrations, data driven assaults on PC applications, have based assaults, for example, benefit and authorizations acceleration, unapproved logins and access to touchy records, and malware (virus, Trojan steeds, and worms). IDS are the best fine grain filter put inside the ensured PC network, searching for known or amazing dangers in network traffic as well as review data recorded by hosts [8][9].

II. RELATED WORKS

Khraisat, Ansam, Iqbal Gondal, and Peter Vamplew [10] This paper inspects various data mining systems that could limit both the quantity of incorrect negatives and false positives. C5 classifier's viability is inspected and contrasted and different classifiers. Results should that false negatives are decrease and intrusion detection has been recover basically. Aljawarneh, Shadi, Monther Aldwairi, and Muneer Bani Yassein [11] the authors built up another composite model that can be deployed to assess the intrusion range threshold degree dependent on the network exchange data's ideal highlights that were made accessible for preparing. The test results uncovered that the crossover approach significantly affected the minimisation of the computational and time unpredictability included when determining the element affiliation effect scale. Sharma, Ruby, and Sandeep Chaurasia [12] n this methodology, bunch effectiveness is improved through a membership matrix generation (MMG) algorithm. Dissimilarity Distance Function (DDF) has been utilized to process the distance metric while making a group in proposing an IDS. The proposed upgraded fluffy c-implies algorithm has been tried upon ADFA Dataset and the model performs exceedingly considerable regarding exactness, accuracy, detection rates, and false alerts. Saxena, Akash, Khushboo Saxena, and Jayanti Goyal [13] proposed work, we at first apply KDD cup'99 dataset which is most comprehensively utilized strategy for identifying intrusion. DBSCAN is the most used strategy which is utilized to kill commotion from the data. At that point, we create the most importance contributions by breaking down and handling entire data which is finished by the choice of highlight strategy. K-means bunching performs gathering of data which is trailed by SMO classifier. So we proposed a mixture structure which improves the taken in general exactness. MATLAB and WEKA instruments are utilized to execute the entire procedure.Gupta, Amara SALG Gopal, G. Syam Prasad, and Soumya Ranjan Nayak [14] With the fast development of different advancements the level for the security has even turned out to be very testing and for the acknowledgment frameworks in peculiarity, a few strategies and philosophy and activities area unit made to pursue novel attacks on the frameworks or systems. Detection frameworks in oddity maintained predefined set of guidelines and conventions. All through this algorithmic program, gatecrasher just prepared to recover or improve key by speaking with the Intrusion Detection System and point of view the

tip result after it and by maltreatment this topic can't set up to meet security standards. Along these lines upheld learning we'd as of late like the theme that can help us with giving additional security on Data Storage.

From the related works, coming up next are the issue proclamation Anomaly-put together detection systems depend with respect to the suspicion that the gatecrasher's conduct is not quite the same as should be expected network practices. These procedures study the ordinary traffic of the net-work and distinguish every degenerate conduct as malignant conduct. This framework gives the likelihood of recognizing both unknown and known attacks. The fundamental impediment of this framework is its high false positive rate.

III. METHODOLOGY

3.1 System Architecture

The General Process of Proposed System fig 3.1 depict the proposed work aims to build an enhanced intrusion detection system using combined feature selection with ANN classifier. The imported dataset from the KDD '99 CUP is used for analytical purpose in WEKA environment. After importing the dataset, data pre-processing is done for selecting the features. The Feature selection is done using the combined feature selection mechanism involving linear correlation coefficient and CuttleFish algorithm. The resultant pre-processed data is classified using Artificial Neural Network.

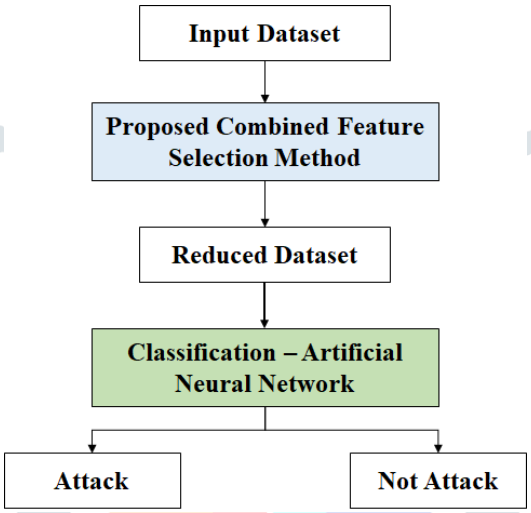

Figure 3.1: Proposed Framework for the Intrusion Detection System

3.2 Linear Correlation Coefficient

In statistics, the correlation coefficient shows the quality and heading of a connection between two random factors. The commonest use alludes to a linear relationship. 'All in all, factual use, correlation or correlation refers to the flight of two random factors from autonomy. Condition (1) demonstrates the figuring of the correlation co-efficient connecting two factors x and y. There are absolutely n perceptions.

$$r_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\ \sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}}$$

3.3 CuttleFish Algorithm

Another meta-heuristic bio-motivated optimization algorithm, called Cuttlefish Algorithm (CFA) is launched. The algorithm impersonates the system of shading changing conduct utilized by the cuttlefish to take care of numerical global optimization problem. The proposed algorithm thinks about two principle forms: visibility and reflection.

3.4 Proposed Combined Feature Selection Method

In this proposed combined feature selection method, the filter-based correlation feature selection method with CuttleFish algorithm. The following step by step procedure represents the combined feature selection method.

Step 1: Generate the population randomly.

Step 2: Initialize the population size as 100, Mutation rate as 0.02 and crossover rate as 0.9

Step 3: Initialize the starting number of generation and maximum generation. Consider the starting number of generations as 0 and maximum number of generation as 200.

Step 4: Consider the feature A in the total number of populations. And calculate the objective function for each individual in A.

Step 5: Calculate the non-dominated fronts for each features in A.

Step 6: While A and Fronts are not equal to null, then do the following steps.

Step 7: Calculate the drowning distance of each features in A.

Step 8: Initialize the null set as B. Check for the size of B with the population size.

Step 9: Apply event selection on the fittest individual feature a1.

Step 10: Apply the event selection on the fittest individual feature a2.

Step 11: Apply random uniform operation on the two features a1 and a2. And stored on separate variable as nc.

Step 12: Check the crossover rate with the value of nc

Step 13: If the value of nc is greater than crossover then do the crossover of a1 and a2 and stored in A.

Step 14: Repeat the above steps for each features in the individual do

Step 15: Apply the random uniform function on each features in the individuals and stored as nm.

Step 16: If the value of nm is greater than the mutation rate then

Step 17: Then flip the feature values. End the for function.

Step 18: Append the off proceed to the B.

3.5 Artificial Neural Network

Artificial Neural Network (ANN) is an effective figuring framework whose focal topic is acquired from the similarity of natural neural networks. ANN secures an enormous accumulation of units that are interconnected in some example to permit correspondence between the units. So as to frame a feed-forward multi-layer in MLP, the accumulation of non-linear

neurons is associated with each other. This strategy is known to be valuable for forecast and characterization issues. Cross-approval is utilized to decide the 'optimal' number of shrouded layers and neurons which were depended on the test plan of the Intrusion Detection order framework. These units, likewise refers to as hubs or neurons, are uncomplicated processors which work in equivalent.

IV. RESULT AND DISCUSSION

4.1 Dataset Description

With the far reaching utilization of PC networks, the quantity of attacks has developed widely, and numerous new hacking devices and nosy strategies have showed up. Utilizing an intrusion detection framework (IDS) is one method for managing suspicious exercises inside a network. Data mining-based intrusion detection systems can be arranged by their detection technique.

This section talk about the KDD'99 cup dataset used for analyse the proposed system. The dataset accommodate five million records and each connection record is described by 41 features. It has 22 categorise of attacks from the following four classes Table 4. 1

Table 4.1: **Attacks falling into four major categories**

Category	Number of Attacks
Denial of Service Attacks	Back, land, neptune, pod, smurf, teardrop
User to Root Attacks	Buffer_overflow, loadmodule, perl, rootkit,
Remote to Local Attacks	Ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster
Probes	Satan, ipsweep, nmap, portsweep

4.2 Experimental Results

Mostly the classification methods are utilized to evaluate the effectiveness of the results obtained from the feature selection techniques. In this research work, the classification methods like ANN and NB have used. The performance analysis of the proposed combined feature selection method processed dataset using ANN and NB classification methods. From the table 4.2, it is clear that the ANN gives better performance than the Naïve Bayes classification method. The performance metrics like Accuracy, Kappa Statistics, Error rates like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Root Absolute Error (RAE), Root Relative Squared Error (RRSE), Precision, Recall, F-Measure, False Positive Rate are considered for evaluating the proposed feature selection with existing methods.

Table 4.2: Performance analysis of the proposed combined feature selection method processed dataset with using ANN and NB Classification Methods

Performance Metrics	Classification Methods	
	Artificial Neural Network	**Naïve Bayes**
Accuracy	98%	92.667%
Kappa Statistics	0.7594	0.7432
MAE	0.0088	0.0128
RMSE	0.1077	0.1311
RAE	20.0334%	73.7135%
RRSE	60.0193%	64.0312%
TP Rate	0.98	0.927
FP Rate	0.155	0.251
Precision	0.98	0.865
F-Measure	0.978	0.894

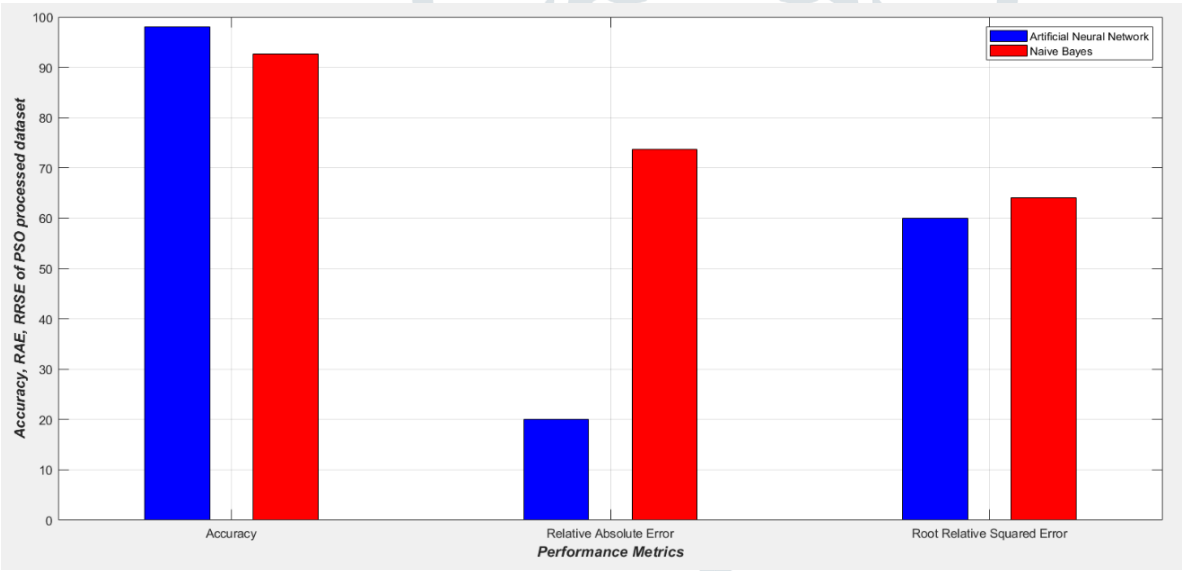

Figure 4.1: Graphical representation of the performance analysis of the proposed combined feature selection processed dataset using ANN and NB classification methods

V. Conclusion

The pre-processing method has implemented to eject the unnecessary and irrelevant features from the dataset. This proposed technique has used to improve the prognostication accuracy. In this work, a Linear Correlation Co-efficient feature selection method has introduced by comparing the PSO and ACO analysis. This method was presented to eliminate the unnecessary feature for the classification in the IDS dataset. From the obtained results it has intent on that the proposed technique worked better than the present feature selection method in the IDS. Also, it enhances the prognostication accuracy and diminishes the error rates. This diminishing of error rates results in the excellent classification accuracy.

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. By using this methodology, the node can be categorise into a malicious and legitimate category

REFERENCES

[1] R. Akbani, Korkmaz, T., Raju, G. V. S: Mobile ad hoc network security. In: Lecture Notes in Electrical Engineering, Springer, vol. 127 (2012).

[2] T. Anantvalee and Wu, J.: A survey on intrusion detection in mobile ad hoc networks. In: Wireless/Mobile Security. New York: Springer (2008).

[3] Elhadi, M., Shakshuki, EAACK.: A secure intrusion-detection system for MANETs. In: IEEE Transactions on Industrial Electronics, vol. 60(3) (2013).

[4] Gungor, V.C., Hancke, G.P.: Industrial wireless sensor networks: challenges, design principles, and technical approach. IEEE Trans. Ind. Electron. 56(10), 4258–4265 (2009).

[5] Haldar, N.A.H.: An activity pattern based wireless intrusion detection system. In: Information Technology: pp. 846–847 (2012).

[6] Shen, J.: Network intrusion detection by artificial immune system, IECON, pp. 716–720 (2011).

[7] Khattab, S., Gabriel, S., Melhem, R., Mosse, D.: Live baiting for service-level DoS attackers. Proceeding of the IEEE INFOCOM (2008).

[8] Macia´-Pe´rez, F.: Network intrusion detection system embedded on a smart sensor, industrial electronics. IEEE Trans. 58(3), 722– 732 (2012).

[9] Benjie Chen, Kyle Jamieson, Hari Balakrishnan And Robert Morris" An Energy-Efficient Coordination Algorithm for Topology Maintenance in Ad Hoc Wireless Networks," in Proceedings of the wireless network, 2002.

[10] Khraisat, Ansam, Iqbal Gondal, and Peter Vamplew. "An Anomaly Intrusion Detection System Using C5 Decision Tree Classifier." *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, Cham, 2018.