Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

An Optimized Genetic Algorithm For Intrusion Detection System In Data Mining

Nilu Majeed
M.Phil. Research Scholar,
Department of Computer Science
Sree Narayana Guru College
Chavadi, Coimbatore, India

Dr. R. Priya
Associate Professor and HOD of Computer Science
Sree Narayana Guru College
Chavadi, Coimbatore, India

***Abstract*— The Intrusion Detection System (IDS) plays a significant part in security schemes. Network IDS (NIDS) seems to have a position in the detection of harmful and unauthorized networks and systems among its various forms. The identification of Denial of Service (DoS) and Probe-based threats in most NIDS investigations was reasonably reliable in the literary works. Consequently, in the current Big-Data based Hierarchic-Deep-Learning System (BDHDLS) for multiple datasets, the detection rate of many other threat segments remains weak. Machine Learning has the capabilities to solve such an inaccuracy problem. In this research, an Optimized Genetic Algorithm (OGA) architecture was proposed for generating strongly optimized results for security analysts both in minor and major attack categories. The evolutionary design is developed utilizing the standard Genetic algorithm with optimization of Shift based Reverse-Logic Crossover. In the training process where the best particle in the GA interacts with the poor particle in the internal GA to produce new solutions which improve the detection of mutant threats. This means that the optimization method develops the right guidelines for important groups of attacks. Various tests were carried out over multiple datasets with varying settings. The results indicate that the proposed method is more accurate than many established methods and more effective.**

Keywords*— *Intrusion Detection System, Deep Learning, Genetic Algorithm, Optimization

I. INTRODUCTION

Despite the technical advances, the majority of real-life activities in the cyber environment are made available. In this modern world, there would also be heavy usage of a wide range of activities such as trading, purchasing, internet research, mobile business and correspondence. Through using mobile in general, people can communicate with this international platform and carry out purchases from wherever also whenever.

While digitization enables people to operate every day, leading to network vulnerability and rapidly developing intrusions, communications are mostly invaded by intruders that use the anonymity of the Internet not just to steal any data or assets, and sometimes to halt communication services [1]. In parallel to network defense antivirus, security managers typically favor secure passwords, authentication and security protocols. Such strategies are nevertheless not enough to secure the device. As this illustrates in Figure 1, several managers tend to use IDSs to track network traffic to block anomalous attacks.

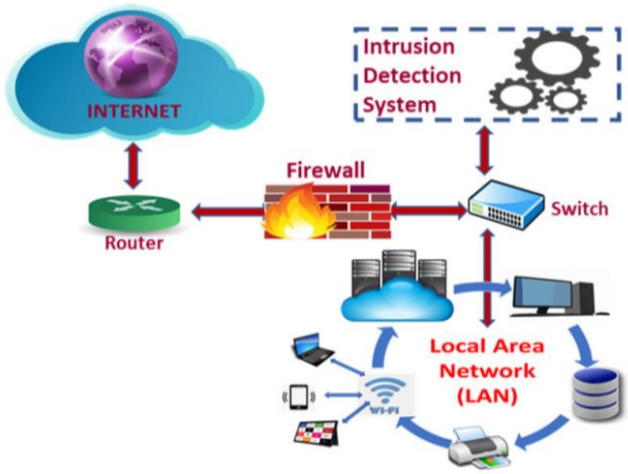

FIGURE 1: IDS AND LAN

Any unauthorized behavior which always creates harm to the security, functionality or credibility of the information in a database model is defined as an intrusion. One such form of operation is strongly prevented by IDSs. Signature-based IDS (SIDS), AIDS (AIDS) and Hybrid Systems are the three categories of IDS.

SIDS maintains suspicious signatures in an information base and attempts to identify intruders using design matching methods. In the meantime, those of AIDS make an effort to learn the usual behavior, which is suspicious with some. In such a method, a SIDS is not needed, and the framework may detect zero-day threats not recently experienced. In an attempt to optimize the identification rate of suspected harmful activity by decreasing the false positive rate of zero-day assaults, hybrid systems are composed of the use of SIDS and AIDS. Because of the benefits of AIDSs, most new IDSs utilize even support from AIDS individually. Such IDSs must be learned by analyzing the dataset using the machine-learning model.

Many studies on this subject have been utilizing old databases comprising obsolete and unbalanced data form volumes [2]. Although several latest databases containing up-to-date information may be found, researchers also find the unbalanced dimensions of data sort a problem. IDS performance is closely correlated to the chosen training model and data set quality [3]. A high-quality data set are being described as a data set to enhance the efficiency of transactions in the real world. A dataset collection is reported to be imbalanced if the groups are not evenly distributed.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

This is a general concern related to the data sets used by many of the classification issues. Imbalanced dataset sets contribute to the classification prejudices used by the main class, but in others, the goal is to identify the minority class.

This causes a significant category mistake for the samples of the minority community which may be missed [4]. It can be balanced as per data forms to enhance the accuracy of the data collection. Consequently, the optimized genetic algorithm is used in this study to classify the dataset either malicious intent or not.

II. RELATED WORKS

The researchers in [5] carried out a comparative analysis to address accuracy issues and related factors utilizing algorithms such as Extreme-Learning, Support Vector Machine (SVM) and Random-Forest (RF). The NSL-KDD dataset has been utilized and is a reference point for the estimation of IDSs. The findings reveal all of this despite accuracy, specificity and consistency accordingly the Extreme-Learning Method is greater than most other methodologies. The researchers in [6] utilized the dataset CIDDS-001 to manage imbalanced dataset sets for the creation of an effective IDS utilizing a multitude of technologies. Those who systematically study the CIDDS-001 sampling methods and test the data through Voting and Stacking models using Deep Neural-Networks and Auto-Variation Encoder. When utilizing an imbalanced dataset, this device observed attacks with 99.99% precision.

The researchers in [7] regulated a machine learning method to distinguish network transmission. For research and preparation, they utilized the NSL-KDD dataset, since they needed to figure out if traffic was suspicious or ordinary. They utilized architectures and function selection methods from Artificial Neural-Network (ANN) and Support Vector Machine (SVM). The ANN with the range of features was more effective than the SVM.

The researchers in [8] developed a hybrid model for a dataset KDDCup99. To decrease the function aspect of the dataset collection, they employed Filter-bases attribute selection. For the identification of attacks in a dataset, K-Means and Minimal Sequential Optimization algorithms have been used. Their approach proposed increases the precision rate considerably.

The researchers in [9] utilize the dataset NSL-KDD with an evolutionary ensemble training algorithm to validate and build an IDS. The multiple techniques they utilized: Deep Neural-Networks, Decision-Tree, K-Nearest Neighbor and Random-Forest. They have also developed a group algorithm for adaptive voting. To validate their method, they were using an NSL-KDD database. The algorithm for Decision-Tree is 84.2% accurate and the adaptive algorithm's final accuracy is 85.2%. They finally compare the corresponding research papers and noticed that their adaptive ensemble paradigm increases the precision of detection.

III METHODOLOGIES

A. *EXISTING MODEL*

The Big-Data integrated Hierarchical Deep-Learning System (BDHDLS) based on was used to arrange multifaceted, hierarchical tree structures for deep learning models was an existing system [10]. The hierarchical tree framework is built to separate samples in multiple level clusters, and the single distribution of data is adjusted to every deep learning model learned on similar samples in one single cluster.

BDHDLS is designed in five stages:

- The first stage is to isolate and pick behavioral attributes and material functionality utilizing big data technology.
- In the second stage, a parallel enhanced K-mean method divides the whole dataset into one-level clusters. Samples reveal identical trends from every cluster of the tree.
- The hierarchical clustering method takes place in the third stage with a poor consistency for each cluster at the same time as producing the cluster sub-tree. Such sub-trees are then merged to create hierarchical multi-level cluster trees. Through particular terms, the whole dataset is broken into various clusters on several layers.
- Over every cluster in the hierarchy tree, the fourth stage builds a deep learning model to learn the distinctive patterns of data propagation for the cluster.
- Throughout the final stage, the judgment factors of deep learning methods are integrated to determine whether or not the research sample is invasive.

A simple fusion approach was used as the main disadvantage in this existing model to integrate the performance of the various deep learning methods in the cluster that cannot be used to provide heterogeneous data.

Furthermore, the specialized fusion-based decision Methods that integrate the outcome from numerous deep learning methods are required to maximize performance. This methodology cannot be the ideal solution.

B. *PROPOSED MODEL*

a) GENETIC ALGORITHM (GA)

Goldberg's GA was built based on Darwin's theory of evolution, which claims that organism survival has an impact on the "most surviving organisms". Darwin even said the reproduction, crossover and mutation mechanism would ensure the viability of an organism. Darwin's development principle is then generalized to the computer technique to obtain a natural approach to an issue named objective function.

The general principle of the GA are as follows:

- A GA solution is considered as a chromosome, whereas a population is called a chromosome array.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

- A chromosome consists of genes, and its meaning will depend on the issue, either be numeric, conditional, sign or text.
- Such chromosomes take on a fitness-function mechanism to determine the appropriateness of the solution with the problem created by GA.
- Few chromosomes in the population join through a crossover mechanism that develops various chromosomes known as descendants, the genes becoming their parent's mix.
- Several chromosomes mutate in their genes over a generation as well.
- A crossover rate and mutation value are used to monitor the number of chromosomes undergoing crossover and mutation.
- Chromosomes would be chosen from the population maintained for the next generation based on the Darwinian development theorem. The chromosome with a higher fitness value is most likely to be recalled in the next generation.
- The chromosome significance corresponds over a range of generations to a definite, the strongest solution for the issue.

The following segment includes the pseudo-code for GA for IDS.

Pseudo code for GA for IDS

1. Initialize the population
2. N = total number of records in the Dataset
3. For each chromosome in the population
4. A = 0, AB = 0
5. For each record in the set
6. If the record matches the chromosome
7. AB = AB + 1
8. End if
9. If the record matches only the “condition” part
10. A = A + 1
11. End if
12. End for
13. Fitness = 1 / (1+F_obj),
14. Where F_obj = f (x) = (a + 2b + 3c + 4d) - 41)
15. If Fitness of chromosome > among all Fitness values
16. Select the chromosome into new population
17. End if
18. End for
19. For each chromosome in the new population
20. Apply crossover operator to the chromosome
21. Apply mutation operator to the chromosome
22. End for
23. If number of generations is not reached, goto line no. 4

Processing phases of GA for IDS

Phase 1: Initializing Phase

- During the first phase, a zero matrix is developed or simply a matrix that in all places contains "0" implies initially that it can suggest, that it is randomly under the initialization point.
- Currently, if it is using some element from the dataset, it would take a single function column out of fixed functions, and it only fills in the row or column as row by column of zero matrix, such that the randomly initialized phase is done.
- Thus, the dataset column can be randomly chosen means from defined functions from the initial input data collection.

Phase 2: For every chromosome, the calculation of fitness is done

The fitness function for every chromosome or gene must be determined by the following equation in this stage.

Fitness = 1 / (1+F_obj)

Where F_obj = f (x) = (a + 2b + 3c + 4d) - fixed features)

Phase 3: Selecting the best

- That's the initial fitness measure iteration. In this, two of the chromosome top fitness values are picked, most genes with a fitness value above 93.7% and 96% chromosome fitness value are chosen for the other measures.
- Currently, this was necessary to repeat this process in iterations 2 and 3, by adjusting the matrix value and calculating fitness values for each freshly formed chromosome, and selecting the maximum value of row or genes or chromosomes for further measures.

Phase 4: Solutions

If the fitness attribute is to be more precise, therefore two kinds of operators are mainly present: crossover and mutation.

Crossover Operation:

Crossover is the first stage in the reproduction process. It is used to make a whole new chromosome with parents' genes. The traditional recombination for the GA is an action that requires two parents, but which may include arrangements for more parents. Two of the most popular algorithms used are scattered conventional crossover and intermediate Blending Crossover. In this region, two top fitness value genes are taken and the crossover procedure on them is applied. New chromosomes are then developed and the fitness of these newly generated genes must now be calculated.

Since crossover, two genes or chromosomes have been newly generated, now the fitness of these newly produced genes must be calculated and compared to two fitness chromosomes which are the maximum in history. It must now select the two genes with the greatest fitness. In this scenario, it would pick the two most valuable genes for

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

additional estimation. Mutation procedure on it must be used for further fitness estimation.

Mutation Operation:

In plain terms, the mutation is just by changing the fitness value of every location. The freshly formed population may be further added to mutations by selection and crossover. Mutation indicates certain rows or gene components are altered. These variations can be triggered by errors when the parent genes are copied. For GA, mutation implies a spontaneous shift in the population's significance of the gene. The chromosome, the gene, and the gene alterations are all selected at random. Now that the mutation procedure has been performed, it would choose from 4 genes, just 2 genes to choose from, both chromosomes with the highest fitness. For various matrix configurations of a given dataset collection, the same operations must be performed and the fitness value measured for greater precision.

b) Optimization in GA

The shift reverse logic crossover operation to be used in this research for optimizing GA as following:

- The colony's primary running parameters are encoded using a binary scheme, which contains the size of the colony 'M', the Probability of crossover "Pc", Probability of variation "Pm" and maximum generation quantity 'T'.
- Make a primary colony of 'M' individuals at random.
- Evaluate the fitness of the colony's individuals.
- Determine if the algorithm's final criteria are fulfilled, and if so, avoid it or proceed to the next step.
- Pick a method for conducting a roulette collection operation depending on the fitness standard.
- Probability "Pc" is used to implement the gene reconfiguration crossover process and Probability "Pm" is used to implement the variation operation.
- Retain the fitness of the individuals that will make up the colony's next generation.

IV RESULTS AND DISCUSSION

The KDD CUP 1999 Datasets were used to assess methods efficiency over a wide range of research samples. The procedure for the methods of the given dataset is given as follows:

- The complete data set was split into ten batches, with every batch of test sets having its experiment.
- The data is tabulated and the experimental findings are provided for misuse identification, anomaly detection using existing (BDHDLS) and proposed models (OGA).
- The True Positive Rate is the percentage of attack packets correctly found by the IDS out of the total packets.
- The proportion of regular packets mistakenly marked as attack packets by the IDS to the total packets is known as the False Positive Rate.
- The training process of the experiment begins with the selection of a subset of tuples from its data collection.
- Through the testing process, this subset is never observed again. This subset's tuples are all labeled by corresponding attack category or a normality indicator.
- The testing process continues after the training phase is finished.
- The tuples are not labeled during the testing process.
- The Misuse Identification Method processes the tuples first.
- The outcome has been processed if the program could categorize the packet data with adequate trust.
- The tuple was being moved to Anomaly Detection Process if the categorization is not confident.
- Both suspicious tuples were compiled and combined with an established dataset with a strong occurrence of regular packets while anomaly detection.
- This is accomplished after every session. The OGA and BDHDLS methods were assigned to the merged data.
- When a suspicious tuple is classified as an anomaly, the method suspects it is related to a new threat.
- Or else, the tuple is identical to a regular packet.

TABLE 1: PERFORMANCE OF TP RATE

TOTAL SAMPLES	BDHDLS	OGA
2000	96	99
4000	91	98
6000	85	96
8000	79	94
10000	71	91

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

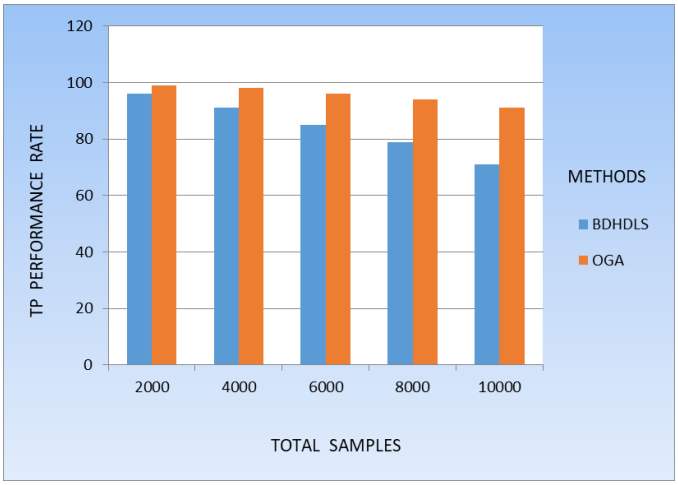

FIGURE 2: GRAPHICAL COMPARISON OF TP RATE

Table 1 and Figure 2 show the performance rate of the TP in which the rate gets affected depending on the variation in the number of samples ranging from 2000 to 10000. The higher TP rate will prove better accuracy. Thus for the given dataset, the performance of the TP rate is better for the proposed OGA when compared with the existing BDHDLS. Hence it proves for a larger dataset the OGA will give better accuracy for identifying the intruder dataset.

TABLE 2: PERFORMANCE OF FP RATE

TOTAL SAMPLES	BDHDLS	OGA
2000	**0.9**	**0.1**
4000	**1.3**	**0.1**
6000	**1.8**	**0.2**
8000	**2.3**	**0.3**
10000	**2.9**	**0.5**

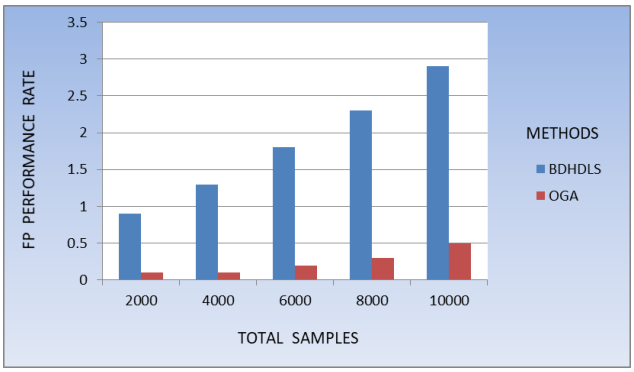

FIGURE 3: GRAPHICAL COMPARISON OF FP RATE

Table 2 and Figure 3 show the performance rate of the FP in which the rate gets affected depending on the variation in the number of samples ranging from 2000 to 10000. The lower FP rate will prove better accuracy. Thus for the given dataset, the performance of the FP rate is better for the proposed OGA when compared with the existing BDHDLS. Hence it proves for a larger dataset the OGA will give better accuracy for identifying the normal dataset.

V. CONCLUSION

Because of the widespread usage of the Internet in recent decades, computer machines may now link to an international platform at whatever time and in any place. Even so, the anonymous nature of the Internet leads to several security flaws in the network, resulting in intrusions. Furthermore, today's hackers are much more advanced and could innovate real malware with the aid of automated development software, relying on IDS' limited detection capabilities. In certain instances, pre-collected datasets are used to train IDS. Most of these databases, though, are unbalanced, with varying mismatch ratios. Imbalanced datasets lead to bias in favor of the dominant party, although in certain rare cases, minority groups are completely overlooked. Such minority groups, on the other hand, are mostly optimistic. As a consequence, the imbalance ratio can be lowered to boost the system's performance while not lowering its average accuracy. In this article, OGA is introduced as an IDS that can effectively identify different forms of network intrusions and harmful behaviors as compared to the existing BDHDLS framework. In the future the research move towards many real-time data with a good accuracy rate.

REFERENCES

[1] E. K. Viegas, A. O. Santin, and L. S. Oliveira, ``Toward a reliable anomaly based intrusion detection in real-world environments," Comput. Netw., vol. 127, pp. 200-216, Nov. 2017.

[2] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, ``Classification with class imbalance problem: A review," Int. J. Adv. Soft Comput. Appl., vol. 7, no. 3, pp. 176-204, 2015.

[3] J. M. Johnson and T. M. Khoshgoftaar, ``Survey on deep learning with class imbalance," J. Big Data, vol. 6, no. 1, p. 27, 2019.

[4] P. Mishra, V. Varadharajan, U. Tupakula, and E. S. Pilli, ``A detailed investigation and analysis of using machine learning techniques for intrusion detection," IEEE Commun. Surveys Tuts., vol. 21, no. 1, pp. 686-728, 1st Quart., 2019.

[5] I. Ahmad, M. Basheri, M. J. Iqbal, and A. Rahim, ``Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection," IEEE Access, vol. 6, pp. 33789-33795, 2018.

[6] R. Abdulhammed, M. Faezipour, A. Abuzneid, and A. Abumallouh, ``Deep and machine learning approaches for anomaly-based intrusion detection of imbalanced network traffic," IEEE Sens. Lett., vol. 3, no. 1, pp. 1-4, Jan. 2019.

[7] K. A. Taher, B. Mohammed Yasin Jisan, and M. M. Rahman, ``Network intrusion detection using supervised machine learning technique with feature selection," in Proc. Int. Conf. Robot., Electr. Signal Process.Techn. (ICREST), Jan. 2019, pp. 643-646.

[8] A. Chandra, S. K. Khatri, and R. Simon, ``Filter-based attribute selection approach for intrusion detection using k-means clustering and sequential minimal optimization technique," in Proc. Amity Int. Conf. Artif. Intell. (AICAI), Feb. 2019, pp. 740-745.

[9] X. Gao, C. Shan, C. Hu, Z. Niu, and Z. Liu, ``An adaptive ensemble machine learning model for intrusion detection," IEEE Access, vol. 7, pp. 82512-82521, 2019.

[10] W. Zhong, N. Yu and C. Ai, "Applying big data based deep learning system to intrusion detection," in Big Data Mining and Analytics, vol. 3, no. 3, pp. 181-195, Sept. 2020, doi: 10.26599/BDMA.2020.9020003.