International Journal of Scientific Research and Engineering Development-– Volume 7 Issue 2, Mar-Apr 2024
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

A Trailblazing Intrusion Detection System Powered by Bee-Inspired Optimization and Ensemble Methods

Priyanka Sahu*, Dr.Abha Tamrakar**
(Department of Computer Science &Engineering,ISBM University, Nawapara, C.G.
Email: roshsahu81@gmail.com)
(Department of Computer Science &Engineering,ISBM University, Nawapara, C.G.
Email: roshsahu81@gmail.com)

----------**********----------

Abstract:
In this study, we present a comprehensive framework aimed at enhancing the performance of machine learning systems through innovative feature selection and ensemble classification methods. Firstly, we introduce novel filter feature selection techniques that amalgamate Mutual Information (MI), ANOVA, and chi-square methodologies utilizing the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) algorithm. Our proposed approach exhibits superior performance compared to state-of-the-art methods in terms of feature selection accuracy and efficacy.This paper introduces a pioneering IDS framework that combines the power of Bee-Inspired Optimization (BIO) with ensemble methods to achieve superior detection accuracy and efficiency. This research heralds a new era in intrusion detection, offering a robust and adaptable solution that promises to significantly advance cybersecurity defense mechanisms.

Keywords —Numbers of Security related keywords.

----------**********----------

I.INTRODUCTION

Intrusion detection is the process of classifying and responding to malevolent activities targeted at computing and network resources". An intrusion detection, also known as attack, mentions to a sequence of actions by use of which an intruder endeavours to gain control over a system. Network security is of vital significance in the present data communication. Programmers and interlopers can make numerous effective endeavours to cause the crash of the systems and web benefits by unapproved interruption.Network intrusion detection system (NIDS)" monitors traffic on a network looking for doubtful activity, which could be an attack or illegal activity[1]. The intrusion detection techniques based upon data mining are generally plummet into one of two categories: misuse detection and anomaly detection.This paper introduces a groundbreaking IDS that harnesses the principles of Bee-Inspired Optimization (BIO) coupled with the robustness of ensemble learning methods. BIO algorithms, inspired by the foraging behaviour of bees, optimize the search for solutions in complex spaces,making them ideal for the dynamic and intricate nature of intrusion detection. When integrated with ensemble methods, which effectively improve prediction accuracy by aggregating multiple models, the result is a system designed for enhanced detection capabilities with reduced false alarms[2]. This paper introduces a groundbreaking IDS that harnesses the principles of Bee-Inspired Optimization (BIO) coupled with the robustness of ensemble learning methods.

Categorization of IoT Threats

IoT Threats Categorization by Layers: IoT systems are structured with multiple layers, each serving a distinct purpose. Figure 1 illustrates a standard representation of IoT architecture with three main layers: the perception (physical) layer, the network (transport) layer, and the application layer [KU14]. Each layer introduces its own vulnerabilities, as depicted in Figure

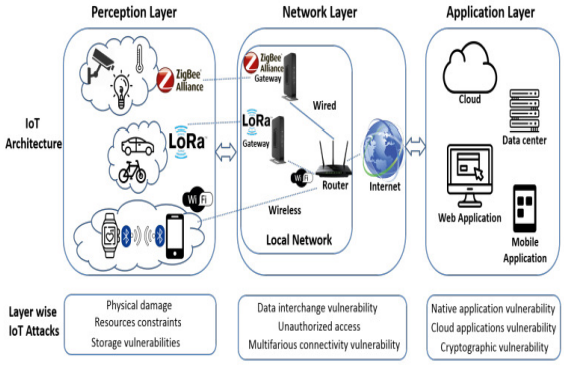

Fig:-1IoT Architecture & Layer-Wise Attacks

Perception (Physical) Layer: This hardware layer comprises sensors and actuators utilizing various communication standards like Bluetooth, RFID, and 6LowPAN. Vulnerabilities include exposure to environmental hazards and risks of physical attacks or unintentional damage.

Network (Transport) Layer: Responsible for effective data transmission, this layer employs communication protocols like Wi-Fi, 3G, 4G, 5G, GSM, IPv6, etc. Vulnerabilities include network-related issues such as data interchange vulnerabilities, unauthorized access, and connectivity vulnerabilities.

Application Layer: Also known as the software layer, it provides business logic and user interfaces. Vulnerabilities in this layer include software-related problems like account enumeration, insecure account credentials, and cloud application attacks.

II. REVIEW OF LITERATURE

A comprehensive comparison between feature reduction methods of intrusion detection in terms of various performance metrics, namely, precision rate, recall rate, detection accuracy, as well as runtime complexity, was provided in the presence of the modern UNSW-NB15 dataset and both binary and multiclass classification [3].

Mina Eshak et al. proposed a combination of feature selection and adoptive voting was used to find network intrusions using the NSL-KDD dataset, a high-dimensional dataset that has been widely used for network intrusion detection [4].

Wang et al. as discussed by the authors proposed an edge network intrusion detection method based on feature selection and TextCNN to enhance the correlation between features and eigenvalues through feature selection, and reduce the intrusion detection cost of nodes [5].

Xiao Zheng et al. said that with the exponential growth of internet data, traditional intrusion detection systems face challenges like reduced efficiency and diminishing utility of single machine learning models. To address this, we propose a novel model based on chi-square feature selection and a stacking approach. Chi-square reduces dataset dimensionality, while ensemble learning with SVM, BPNN, K-Means, and XGBoost enhances accuracy. K-Fold cross-validation mitigates overfitting. Tested on NSL-KDD dataset, our model achieves 97.9% accuracy, 95.2% precision, and 98.7% recall, outperforming single machine learning models. This underscores its effectiveness in intrusion detection, crucial for ensuring network security [6].

Houssam Zouhri et al. said that high dimensionality poses a challenge in Intrusion Detection Systems (IDS), leading to overfitting and increased false positive and false negative rates. This study evaluates the impact of five univariate and three multivariate feature selection techniques on the performance of four classifiers across various intrusion detection datasets. Results indicate that XGBoost and Random Forest trained with multivariate methods effectively reduce feature dimensionality without compromising classification performance or detection rate, outperforming other filtering techniques [7]

III. FEATURE SELECTION APPROACH FOR IDS

Problem in the real world mostly consists of huge amount of data, getting useful information, storage of data becomes very multifaceted. Collected data from various sources are very high dimensional

which contains substantial number of features or attributes.

Filter Method: Filter methods are commonly used as a pretreatment step. The current selection of features is independent of any machine learning algorithms. On the other hand, features chosen based on their scores in the various statistical tests for their correlation with the result variable.[8]

Wrapper Method: On the wrapper methods, the feature selection procedure is based on a particular machine learning algorithm that we are trying to fit on a given dataset. [9] This follows a greedy look up approach by assessing every possible combination of features contrary to the assessment criterion.

Embedded Method: Embedded method combine the best qualities of filter and wrapper techniques [10].In this approach, the features are selected during the learning process.It is less computationally expensivePotentialof overfitting problem/phenomena is very less.

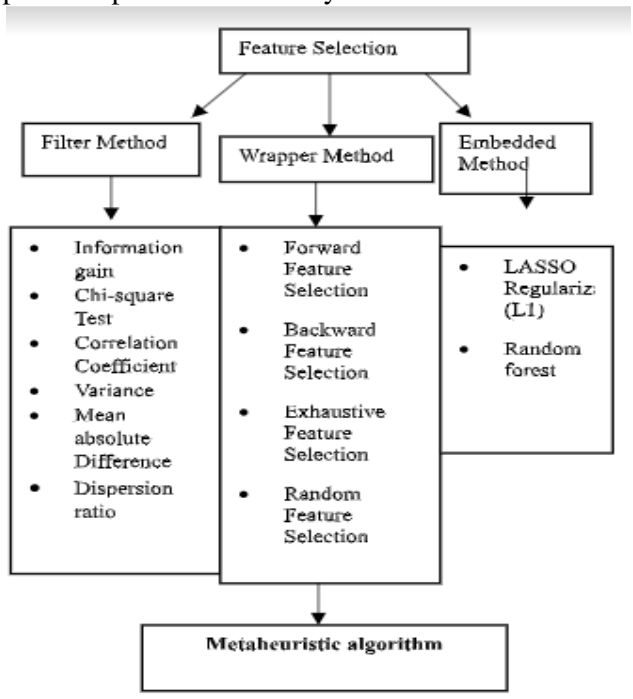

Fig. 2 Feature Selection Methods

IV.FILTER-BASED FEATURE SELECTIONUSING TOPSIS

Intrusion Detection Systems (IDS) play a pivotal role in safeguarding against malicious activities and unauthorized access. As the complexity and sophistication of cyber threats continue to evolve, the need for effective IDS solutions becomes increasingly paramount[11].

feature selection techniques like ANOVA, Chi-Square, and Mutual Information have been employed to rank the importance of features in intrusion detection datasets. ANOVA assesses the variance between groups, Chi-Square measures the dependence between variables, and Mutual Information quantifies the amount of information shared between variables.

The motivation behind the proposed approach lies in the desire to develop a novel methodology that integrates the strengths of ANOVA, Chi-Square, and Mutual Information feature selection techniques using the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS). By leveraging the discriminative power of these methods and aggregating their rankings, we aim to enhance the effectiveness and reliability of intrusion detection systems in accurately identifying and mitigating network intrusions.

One-Hot Encoding

One-hot encoding is a popular technique used in machine learning and data preprocessing to handle categorical variables. When dealing with categorical data, such as color, gender, or country, machine learning algorithms often require numerical input. One-hot encoding converts categorical variables into a binary format that can be fed into machine learning algorithms more effectively.

For example, consider a dataset with a categorical variable "Color" containing three categories: Red, Blue, and Green. With one-hot encoding, this variable would be transformed into three binary columns: "Red", "Blue", and "Green". Each row in the dataset would have a 1 in the column

corresponding to its color and 0s in the other columns.
One-hot encoding is essential for machine learning algorithms because many algorithms, such as linear regression or support vector machines, require numerical input data. Categorical variables cannot be directly used in these algorithms, as they rely on mathematical operations that assume numerical values.

ANOVA

ANOVA, or Analysis of Variance, is a statistical method used to analyze the differences between group means in a dataset.

The total variability (SST) observed in the data can be expressed as the sum of the between-group variability and the within-group variability:

$$SST = SSB + SSW$$

This is done by calculating the F-statistic, which is the ratio of the mean square between groups (MSB) to the mean square within groups (MSW):

$$F = \frac{MSW}{MSB}$$

$$MSB = \frac{SSB}{dfB}$$

MSB, is the mean square between groups, calculated by dividing the between-group sum of squares (SSB) by the degrees of freedom between groups (dfB).

CHI-SQUARE

The Chi-Square (χ^2) test is a statistical method used to determine whether there is a significant association between categorical variables in a contingency table.
Mathematically, the Chi-Square statistic is calculated as follows:

$$\chi^2 = \sum \frac{(o_i - E_i)^2}{E_i}$$

Where:
χ^2, is the Chi-Square statistic,
o_i is the observed frequency for each cell in the contingency table,
E_i is the expected frequency for each cell in the contingency table,
$\sum$ denotes the sum over all cells in the contingency table.

MUTUAL INFORMATION

Mutual Information is a measure of the amount of information shared between two random variables

Mutual Information (MI) between two discrete random variables X and Y is defined as:

$$I(X;Y) = \sum x \in X \sum y \in Y p(x,y) \log(p(x)p(y)p(x,y))$$

Where:
$I(X;Y)$ is the mutual information between variables X and Y,
$p(x,y)$ is the joint probability mass function of X and Y,
$p(x)$ and $p(y)$ are the marginal probability mass functions of X and Y, respectively,
$\sum x \in X \sum y \in Y$ denote the sum over all possible values of X and Y, respectively.

TOPSIS

TOPSIS provides efficient outcome with simple and easy way, and is capable of evaluating the relative performance of various decisions [12]. Following are the steps for TOPSIS

Step 1: Create a Decision Matrix (DM)

$$DM = \begin{bmatrix} X_{01} & \cdots & X_{0n} \\ \vdots & \ddots & \vdots \\ X_{m1} & \cdots & X_{mn} \end{bmatrix};\ i = \overline{0,m};\ j = \overline{1,n} \quad (1)$$

Step 2: Normalize the DM

$$N_{ij} = \frac{X_{ij}}{\sqrt{\sum_{i=1}^{m} X_{ij}^2}} \quad (2)$$

Step 3: Weight Normalization

$$w_j = (w_1, w_2, \ldots, w_n) \quad (3)$$

Where w_j is j^{th} criteria's weight and $\sum w_j = 1$. The normalized weight matrix (V) is obtained by following equation

$$V_{ij} = w_j N_{ij} \quad (4)$$

Step 4: Choosing the best course of action Using this approach, a matrix of positive and negative ideal solutions

$$A^+ = \left(maxV_{ij} | j \in J\right), \left(minV_{ij} | j \in J'\right), i = 1,2, \dots, m \quad (5)$$

$$A^- = \left(minV_{ij} | j \in J\right), \left(maxV_{ij} | j \in J'\right), i = 1,2, \dots, m \quad (6)$$

Step 5: Calculate Separation

(a) Alternative distance S^+ from the ideal positive solution is

$$S_i^+ = \sqrt{\sum_{j=1}^{n} \left(V_{ij} - V_j^+\right)^2}, i = 1,2, \dots, m \quad (7)$$

(b) Alternative distance S^- from the ideal negative solution

$$S_i^- = \sqrt{\sum_{j=1}^{n} \left(V_{ij} - V_j^-\right)^2}, i = 1,2, \dots, m \quad (8)$$

Step 6: Calculate Positive Ideal Solution

$$C_i^+ = \frac{S_i}{S_i^- + S_i^+} \quad (9)$$

Step 7: Alternative Rank Computation

Alternative C^+ ranked from higher to lower value. The option with the uppermost value of C^+ is the most ideal choice.

Proposed Filter Feature Selection

In this section, we outline the proposed methodology for feature selection and classification in the context of intrusion detection. The methodology consists of the following steps:
Feature Selection using ANOVA, Chi-Square, and Mutual Information:

1)Anova: Conduct Analysis of Variance to assess the significance of differences between group means for each feature in the intrusion detection dataset.
2)Chi-Square: Compute the Chi-Square statistic to measure the association between each feature and the target variable (intrusion vs. non-intrusion).
3)Mutual Information: Calculate the Mutual Information score between each feature and the target variable to quantify their dependency.
Performance Evaluation:

1)The performance of each classifier is assessed for each selected feature subset size (5, 10, 15, 20, 25) % of total features.

2)Comparative analysis is conducted to determine the effectiveness of feature subsets in classification accuracy and the efficiency of different classifiers.

V.RESULT AND DISCUSSION

The proposed algorithm was implemented and tested on two benchmark datasets: KDD and ToN_IoT. These datasets are commonly used in intrusion detection research to evaluate the effectiveness of algorithms in identifying malicious network activity.

Table 1Accuracy Comparison

Dataset	% of features selected	ANOVA	Chi-square	MI	Proposed	[17]	[31]
ToN_IoT 2017	5	34.9953	54.0354	34.9953	66.2489	42.0347	33.8550
	10	65.0047	74.4117	65.0047	59.5099	43.6049	37.8180
	15	65.0047	79.0584	65.0047	59.0417	49.4098	34.3094
	20	34.9953	65.0228	34.9960	67.3683	32.6572	44.3442
	25	65.0047	65.0228	34.9974	68.4834	44.0778	36.0311
KDD 2009	5	86.1880	92.1320	91.6662	93.3372	78.5558	78.5120
	10	86.6488	90.7346	91.0435	94.6676	72.5214	78.4270
	15	83.0996	90.3397	89.3929	94.2007	77.6755	83.5829
	20	83.1452	71.3382	89.0588	97.7738	81.9157	86.4619
	25	93.0688	86.9171	86.9424	99.1716	81.2230	71.5019

The ToN_IoT dataset, sourced from the IEEE Transactions on Networking, focuses specifically on IoT network traffic, providing insights into the unique challenges of securing IoT environments.Implemented in a Python environment on a Windows, the algorithm selected feature subsets comprising 5%, 10%, 20%, 30%, 40%, and 50% of the total features from each dataset. Subsequently, the performance of the selected feature subsets was compared with state-of-the-art methods. For both datasets, the proposed algorithm demonstrated promising results in feature selection and classification. The performance metrics, including accuracy, precision was evaluated for each selected feature subset size.

V. CONCLUSION

The proposed algorithm for feature selection and classification in intrusion detection, evaluated on the KDD and ToN_IoT datasets, has yielded promising results. By selecting feature subsets comprising 5%, 10%, 15%, 20%, and 25% of the total features from each dataset and employing ensemble classifiers, superior performance has been achieved compared to state-of-the-art methods.This results in improved classification accuracy and robustness against diverse types of network intrusions. Furthermore, the evaluation of different feature subsetsizes (5%, 10%, 15%, 20%, and 25%) has provided valuable insights into the trade-offs between dimensionality reduction and classification performance.

ACKNOWLEDGMENT

We extend our heartfelt gratitude to all those who contributed to the realization of this trailblazing intrusion detection system powered by bee-inspired optimization and ensemble methods. Our sincerest appreciation goes to the researchers, developers, and organizations whose expertise, dedication, and support made this project possible. we express our gratitude to the participants of this study whose feedback and involvement were instrumental in refining our approach.

REFERENCES:

[1]M. Prasad, S. Tripathi, and K. Dahal, "An intelligent intrusion detection and performance reliability evaluation mechanism in mobile ad-hoc networks," Eng Appl ArtifIntell, vol. 119, p. 105760, 2023.

[2] A. Alotaibi and M. A. Rassam, "Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense," Future Internet, vol. 15, no. 2, p. 62, 2023.

[3] V.-D. Ngo, T.-C. Vuong, T. Van Luong, and H. Tran, "Machine learning-based intrusion detection: feature selection versus feature extraction," Cluster Comput, pp. 1–15, 2023.

[4] M. E. Magdy, A. M. Matter, S. Hussin, D. Hassan, and S. A. Elsaid, "Anomaly-based intrusion detection system based on Feature selection and Majority Voting," Indones. J. Electr. Eng. Comput. Sci., 2023.

[5] H. Wang, S. Pan, X. Ju, and Y. Feng, "Edge network intrusion detection method based on feature selection and TextCNN," in International Conference on Electronic Information Engineering and Data Processing (EIEDP 2023), SPIE, 2023, pp. 815–821.

[6] X. Zheng, Y. Wang, L. Jia, D. Xiong, and J. Qiang, "Network intrusion detection model based on Chi-square test and stacking approach," in 2020 7th International Conference on Information Science and Control Engineering (ICISCE), IEEE, 2020, pp. 894–899.

[7] H.Zouhri, A. Idri, and A. Ratnani, "Evaluating the impact of filter-based feature selection in intrusion detection systems," Int J Inf Secur, pp. 1–27, 2023.

[8] Ieracitano, C., Adeel, A., Morabito, F. C., & Hussain, A. (2020). A novel statistical analysis andautoencoder driven intelligent intrusion detection approach. Neurocomputing, 387, 51-62

[9]Kumar, P., Gupta, G.P. & Tripathi, R. Toward Design of an Intelligent Cyber Attack Detection System using Hybrid Feature Reduced Approach for IoT Networks. Arab J Sci Eng 46, 3749–3778 (2021).https://doi.org/10.1007/s13369-020-05181-3

[10] Mahendra P., Sachin T., Keshav D., An efficient feature selection-based Bayesian and Rough set approach for intrusion detection, Applied Soft Computing, Volume 87, 2020, 105980, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2019.105980

[11] M. H. L. Louk and B. A. Tama, "Dual-IDS: A bagging-based gradient boosting decision tree model for network anomaly intrusion detection system," Expert Syst Appl, vol. 213, p. 119030, 2023.

[12] R. Rahim et al., "TOPSIS method application for decision support System in internal control for selecting best employees," in Journal of Physics: Conference Series, IOP Publishing, 2018, p. 012052.

