IJWR

An Improved K-Means Clustering Feature Selection and Biogeography Based Optimization for Intrusion Detection

Aliakbar Tajari Siahmarzkooh*
Department of Computer Sciences, Golestan University, Gorgan, Iran, a.tajari@gu.ac.ir

ABSTRACT

In order to resolve the issues with Intrusion Detection Systems (IDS), a preprocessing step known as feature selection is utilized. The main objectives of this step are to enhance the accuracy of classification, improve the clustering operation on imbalance dataset and reduce the storage space required. During feature selection, a subset of pertinent and non-duplicative features is chosen from the original set. In this paper, a novel approach for feature selection in intrusion detection is introduced, leveraging an enhanced k-means clustering algorithm. The clustering operation is further improved using the combination of Gravity Search Algorithm (GSA) and Particle Swarm Optimization (PSO) techniques. Additionally, Biogeography Based Optimization (BBO) technique known for its successful performance in addressing classification problems is also employed. To evaluate the proposed approach, it is tested on the UNSW-NB15 intrusion detection dataset. Finally, a comparative analysis is conducted, and the results demonstrate the effectiveness of the proposed approach, in such a way that the value of the detection accuracy parameter in the proposed method was 99.8% and in other methods it was a maximum of 99.2%.

Keywords*— Intrusion detection, Gravity Search Algorithm (GSA), Biogeography Based Optimization (BBO), K-means clustering, Particle Swarm Optimization (PSO).*

1. Introduction

The advancements in computer technology have outpaced the effectiveness of conventional security measures and secure network protocols. Firewalls and antivirus software are no longer sufficient in safeguarding against the ever-growing sophistication of attacks [1, 2]. As a result, Intrusion Detection Systems (IDS) serve as an additional layer of protection by analyzing network and system activities to distinguish between normal operations and potential threats [3, 4]. Developing an IDS involves collecting and preprocessing data, detecting intrusions, and responding to them. Due to its interconnected nature, the network's vulnerability to attacks escalates. Numerous assaults and malevolent events can compromise diverse strata within networks, engendering apprehensions regarding security.

An effective measure to safeguard the integrity of communication across network layers is through an IDS. Various intrusion detection solutions have been introduced specifically for securing communication over the Internet. Their primary function involves continuous monitoring of network activity to identify any malicious incidents, which are promptly reported to the system administrator through alert messages [5, 6]. The advantages of network devices lie in their small and easily deployable nature, particularly in remote regions. Despite their compact size and limited battery capacity, it is crucial to develop lightweight protocols and algorithms for efficient communication. This becomes particularly relevant when addressing attack detection, as the algorithms should be optimized for low energy consumption alongside being lightweight [7, 8].

In hopes of uncovering harmful behaviors within computer systems, IDSs are designed for the purpose of detection. These systems, whether network or host-based, strive to address security challenges faced by network devices [9]. By diligently observing the flow of inbound and outbound traffic stemming from various devices, IDS plays a vital role in the identification of potential cyberattacks. Acknowledging the necessity to combat such threats, IDS operates through two distinctive approaches: one

http://dx.doi.org/10.22133/ijwr.2024.423344.1192
Citation A. Tajari Siahmarzkooh, " An Improved K-Means Clustering Feature Selection and Biogeography Based Optimization for Intrusion Detection ", *International Journal of Web Research*, vol.6, no.2,pp.57-66, 2023, doi: http://dx.doi.org/10.22133/ijwr.2024.423344.1192.

*Coressponding Author
Article History: Received: 1 August 2023 ; Revised: 12 November 2023; Accepted: 22 November 2023.

being signature-based detection, while the other stems from anomaly-based analysis [10].

The signature-based approach in intrusion detection systems involves pre-programming a list of recognized threats and their associated indicators, such as patterns of behavior commonly observed during network attacks. By comparing incoming packets to this database of known attack signatures, an IDS employing signature-based methods can detect any suspicious activity passing through the network. In contrast, anomaly-based IDS systems utilize Machine Learning (ML) training to establish a baseline of normal behavior for network systems. Once this baseline is established, each network activity is assessed against it, enabling the anomaly-based IDS to trigger alerts whenever it detects any deviations from this expected behavior [11].

The protection provided by intrusion detection systems ensures the safety of communication across various devices while identifying intrusions within the Internet of Thing (IoT) layers [12]. Multiple IDS solutions have been developed to establish secure communication over the web. These systems continuously observe and track malicious activities occurring within the network, promptly notifying system administrators upon the detection of any attacks. The compact size of IoT devices allows for effortless deployment even in remote regions; however, their computational capabilities are limited due to their small dimensions and relatively low battery capacities. Additionally, lightweight protocols are employed for efficient communication purposes [13]. Thus, it is crucial that the algorithms devised for detecting attacks possess low energy consumption and possess a lightweight nature to accommodate these constraints.

Current methods of intrusion detection systems typically gather and categorize data based on specific attributes. However, in real network scenarios where devices exchange information, data variability becomes more likely. Furthermore, most existing approaches solely employ a singular classification phase. When training models on imbalanced dataset, a bias towards the class with larger samples can occur. To address these challenges, we have proposed a two-tiered approach combining feature selection and classification strategy. This paper advances the field through the following contributions:

1. A novel and enhanced k-means clustering algorithm for generating rich-feature sets;
2. The utilization of Biogeography-Based Optimization (BBO) for data classification;
3. Thorough evaluation and comparison of our proposed methodology against alternative approaches based on accuracy parameters.

The subsequent segments of this document adopt this structure: an examination of current intrusion detection methodologies and associated studies in Section 2, followed by an extensive explanation of the suggested approach and its key algorithms, such as refined k-means clustering and BBO, in Section 3. The 4th section showcases experiments conducted to display the efficacy of the proposed techniques. Ultimately, in Section 5, this paper is brought to a close.

2. Related works

The expansion in data transmission and communication protocols has given rise to an upsurge in security concerns, necessitating the development of effective and customized intrusion detection systems. Researchers have extensively explored various methodologies for IDS. This segment aims to assess a range of diverse research studies conducted in recent years towards proposing an intrusion detection solution.

According to the findings of the researchers in [14], a novel approach was proposed for handling the dataset prior to training a Convolutional Neural Network (CNN). This involved transforming the data into a matrix structure that bore similarity to an image. Subsequently, this transformed matrix was employed as input to the CNN, achieving remarkable accuracy with an average of 98.9% evaluated based on the cross-validation. In a recent publication [15], researchers proposed a system utilizing Artificial Neural Networks (ANNs) for the purpose of detecting intrusions in a multi-class classification scenario. Their study involved extensive exploration of various hyperparameters in order to identify the most effective setup. The authors emphasized the crucial role of hyperparameter selection in producing accurate classification results, highlighting the significant impact that even slight adjustments can have. Notably, the activation function, optimizer, number of epochs, number of neurons, and batch size were among the hyperparameters fine-tuned during their investigation. The performance evaluation of the model was conducted using two well-known datasets, namely NSL-KDD and CICIDS2017.

In the realm of network security, authors in [16] devised an innovative approach known as the SLFN method, which proved to be effective in identifying instances of malicious activity. Their technique consisted of employing data reduction strategies such as clustering, along with the implementation of the SMOTE oversampling technique. To assess the model's performance, the authors employed key metrics like accuracy, precision, recall, and G-mean. Additionally, in [17], a distinct two-stage hybrid method specifically has been designed to detect malicious attacks within IoT networks. This technique integrated a Genetic Algorithm, known as

GA, to intelligently select relevant features, in conjunction with popular machine learning techniques such as Support Vector Machines (SVM), ensemble classifiers, and Decision Trees (DT).

In a recent study [18], a novel approach was presented for intrusion detection using the CISIDS2017 dataset. The researchers combined CNN and LSTM, integrating them seamlessly to enhance the accuracy of the detection system. To optimize feature reduction, an innovative optimization algorithm called NSGA was utilized. Experimental evaluations were conducted on the latest CISIDS2017 datasets, specifically focusing on DDoS attacks. The results were impressive, achieving an accuracy rate of 99.03% while significantly reducing the training time by five-fold, utilizing a High-Performance Computer (HPC).

In another paper [19], authors presented a novel approach to feature selection in classification models. Their model disregards irrelevant features and focuses solely on the significant ones, resulting in improved performance and reduced processing time. By combining association rule mining with the central point of attribute values, their proposal yielded promising results in terms of accuracy and false alarm reduction. Building upon this concept, another group of researchers [20] introduced a new model based on a customized generic algorithm and least squares support vector machine. Their evaluation demonstrated high rates of anomaly classification accuracy, with low false-positive rates. Additionally, a Reduced Error Pruning Tree algorithm was introduced in a separate study [21]. This model incorporates various layers, including feature selection based on user requirements, grouping network flows by protocol, anomaly detection, and inspection of detected abnormalities.

The model proposed in [22] was assessed by the authors using the Bot-IoT dataset, resulting in an impressive accuracy rate of 99.998% for multiclass detection. To enhance the development of deep learning-based intrusion detection systems, the authors highlight five essential design principles. Drawing from these principles, a novel approach called Temporal Convolution Neural Network (TCNN) was implemented, integrating Convolution Neural Network (CNN) with causal convolution. To address the issue of imbalanced datasets, TCNN was further combined with Synthetic Minority Oversampling Technique-Nominal Continuous (SMOTE-NC). Additionally, efficient feature engineering techniques were employed, encompassing feature space reduction and feature transformation.

In [23], researchers introduced an innovative approach to intrusion detection by leveraging deep learning techniques. Their novel schema effectively categorizes traffic for both binary and multi-class classification tasks, yielding exceptional performance across all evaluation metrics such as Accuracy, Recall, Precision, and F1-score. The achieved results were remarkably high, reaching close to 99% accuracy in binary classification. The IoT network security solution presented in [24] utilizes a hybrid DL approach, combining RNN and CNN models. This model exhibits an accuracy rate of 96.32%, an F1 score of 95.74%, and precision and recall values of 95.43% and 96.32% respectively.

The researchers in [25] devised a method for detecting network intrusions by employing convolutional neural networks (CNN-IDS). Instead of the conventional traffic vector format, they transformed it into an image format to minimize computational expense. To evaluate the effectiveness of their CNN model, they utilized the KDDCUP99 dataset. The experimental results demonstrated a significant reduction in classification time and an improvement in classification performance brought about by the proposed model.

In [26], researchers introduced a novel architecture for the detection of multiple classes using artificial neural networks, which they referred to as CANIDS. To evaluate the performance of this model, they conducted experiments on two datasets: UNSW-NB15 and KDD99. The results indicated an overall accuracy of 86.40% and 94.96% for each dataset, respectively. Building on this work, Al-Zewairi et al. [27] further explored the potential of deep learning classifiers, specifically focusing on a multilayer feedforward artificial neural network optimized through backpropagation and stochastic gradient descent. Their findings shed light on the effectiveness of these techniques. The authors in [28] aimed to enhance the performance of ANN models by reducing computational costs. They achieved this by applying an efficient feature selection algorithm. The outcomes of their evaluation on the UNSW-NB15 and NSL-KDD datasets demonstrated that their proposed ANN model achieved a remarkable accuracy score of 95.45%, surpassing the benchmark methods. In an effort to safeguard the integrity of neural network models, Nguyen et al. [29] proposed a multiclass cascade of ANNs for attack detection. Their experiments on the UNSW-NB15 dataset showcased an accuracy of roughly 95.84%, alongside 83.40% precision and 79.19% recall rates. To address the issue of web application attacks, Moustafa et al. [30] developed an anomaly-based detection system. This system consists of a network data collector, a module for dynamic feature selection based on association rule mining, and an optimized Outlier Gaussian Mixture classification module. In line with these advancements, researchers in [31] proposed an intrusion detection system tailored for cyber-physical systems. By combining the hidden Markov model and beta mixture model, they achieved outstanding performance in detecting security threats, as

evidenced by a detection rate of 95.89%, accuracy of 96.32%, and a false-positive rate as low as 3.82%.

The primary objective of above mentioned research studies has been to tackle security and privacy challenges in Internet networks through the development of Intrusion Detection Systems, employing a range of Machine Learning (ML) and Deep Learning (DL) techniques. Scholars have put forward their solutions using two main approaches: framework and model. In terms of dataset selection, authors often rely on popular options like UNSW-NB15 and NSL-KDD datasets. The performance of these proposed approaches varies depending on the chosen datasets and their input characteristics. This paper aims to achieve improved results for a diverse set of potential attack classes by combining optimized clustering algorithm with BBO methodology.

3. Proposed approach

In this section, a proposed solution for intrusion detection is described, which is a combination of selecting the best features using a modified k-means clustering algorithm [32] and biogeography-based optimization [33] for data classification.

The proposed approach consists of four parts: 1) dataset preprocessing, 2) selection of the best features, 3) data classification and 4) performance evaluation. First, the dataset is preprocessed to remove any redundant data and missing values. After that, the feature selection strategy based on an optimized k-means clustering algorithm is performed in order to find the minimum set of features that can be effectively used for classification. Then the data is sent to the biogeography optimization to determine the data class (normal or abnormal). In the final stage, the evaluation results of the proposed method for both normal and malicious traffic are shown on the standard dataset. The following sections describe the phases of attack detection.

3.1. Preprocessing phase

The clustering algorithm and so other techniques necessitate data preprocessing, which entails transforming data into a compatible structure for more processes. This phase encompasses various tasks, such as data cleansing, encoding labels and normalizing values.

Data cleansing

Prior to model training, it is imperative to ensure that the dataset is devoid of any empty or undefined instances. In this particular study, the dataset validation process was carried out using the Panda library, a Python tool. To rectify this, all instances with missing values were eliminated, resulting in a cleaned dataset.

Encoding labels

One popular method to handle categorical values is known as label encoding. This technique associates a distinct numerical value with each categorical value. Prior to implementing clustering algorithms, the dataset's input and output values must be in integer form. The dataset employed in this study possesses categorical features, each encompassing numerous categories. While one-hot encoding demands excess memory and time in these instances, the label encoder method was adopted to transform the categorical features into numeric representations.

Normalizing values

The process of normalization is frequently employed in the initial treatment of data for clustering and classification algorithms. Its main aim is to standardize the numeric column values within a dataset, ensuring that variations in value ranges are preserved. Diverse features within the dataset possess distinct values, ranging from positive hundreds to negative values that adversely affect the model's performance. To address this issue, the data undergo a normalization process utilizing the min-max method, where values are scaled down to a range between 0 and 1, as denoted by Equ (1).

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

3.2. Feature selection

Due to the high dimensions of the dataset, the improved clustering method helps to reduce the dimensions of the dataset to the desired level. Therefore, the minimum number of features are selected based on the level of changes they have. Because clustering is an unsupervised method, it does not use class labels to find principal components, which maximizes the variance of the data. Here we describe the performance of the optimal clustering algorithm of k-means.

Improved k-means clustering algorithm

The general process of clustering algorithm based on canopy density is as follows:

1. Receiving the input dataset and calculating the average distance between samples using Equ (2) [34].

$$Mean(D) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} d(x_i, x_j) \quad (2)$$

2. The environmental density of all dataset samples with the help of Equ (3).

$$p(i) = \sum_{j=1}^{n} f\left[d_{ij} - Mean(D)\right], f(x) = \begin{cases} 1, x < 0 \\ 0, x \geq 0 \end{cases} \quad (3)$$

3. Starting the clustering process by selecting the first cluster head among the samples with

the highest environmental density (max p(i)) [35].

4. It is calculated by the distance of the samples to the cluster head and the membership of those whose Euclidean distance to the cluster head is less than the average distance of the samples.

5. Calculation of intra-cluster distance for samples in the dataset (Equ(4)).

$$a(i) = \frac{2}{p(i)p(i-1)} \sum_{i=1}^{p(i)} \sum_{j=i+1}^{p(i)} d(x_i, x_j) \quad (4)$$

6. Calculate the supercluster distance for each sample (in the first steps of the algorithm, each member can potentially be a cluster head) (Equ(5)).

$$S(i) = \begin{cases} mindist\{i,j\}, \exists j, p(i) < p(j) \\ maxdist\{i,j\}, \nexists j, p(i) < p(j) \end{cases} \quad (5)$$

7. If a sample has a low density value and a large intra-cluster distance, it means that it is suspicious of an anomaly and should be removed from the input dataset to prevent the size of the cluster.

8. In this step, the weighting factor of the remaining samples in the dataset is calculated based on the parameters of density, compactness and intra-cluster distance. If a sample receives a higher weighting factor, it means that it is located in the center and can be selected as a cluster head (See Equ(6)). Therefore, the next cluster is formed and the current sample is added to the collection. After that; all samples that satisfy the minimum distance condition to the cluster head are considered as its members. After clustering; again, the samples assigned to the k-th cluster are removed from the input dataset, and this process uses a weight parameter instead of choosing the threshold value of the cluster boundaries manually, as in the classical canopy algorithm. It can significantly increase the classification accuracy.

$$w(i) = p(i * \frac{1}{a(i)} * S(i)) \quad (6)$$

9. In the last step, the weight coefficient of the remaining samples in the dataset was calculated repeatedly using Equ(4), and until the dataset is empty, the algorithm was repeated from the fifth step. It is repeated in this way, k cluster heads are selected from the dataset and initial clustering is done as mentioned. The main goal of the canopy algorithm is to choose the correct value of k and also reduce the sensitivity to noise. This issue is presented in steps 7 and 8 of the above mentioned algorithm, respectively. In the next section, the final clustering should be done.

Final clustering

The process of selecting the cluster heads using the hybrid algorithm based on the gravity and Particle Swarm Optimization (PSO) is given below [36].

To improve the clustering algorithm, the most important factors that worsen the performance of the k-means algorithm are identified. Therefore, by using the technique of determining the initial cluster heads and the best cluster heads in successive iterations of the algorithm, the deterioration of clustering in successive iterations is avoided. Our main finding is that when the clusters overlap, k-means can be significantly improved using the combination of the two mentioned algorithms. With our clustering measure, the number of false clusters is expected to be greatly reduced, which is discussed in the results section. When the data have well-separated clusters, the performance of k-means depends on initializing the clusters and determining the cluster heads at each step. Therefore, if high clustering accuracy is required, a better algorithm should be used instead.

1. Generating the initial population in the problem space.

2. Calculating the fitness function: Here, each factor refers to a member of the cluster head. The minimization of inter-cluster distance is considered as a fitness function. Therefore, the cluster heads that can reduce the inter-cluster distance increase the similarity of the categories, and as a result, the data classification is done more accurately.

3. Calculate the strength of the reaction of the particles to each other at the current time by using Equ(7).

$$F_{ij}^d = G(t) \frac{M_{pi}(t) * M_{aj}(t)}{R_{ij}(t) + \varepsilon} (X_j^d(t) - X_i^d(t)) \quad (7)$$

4. Determining the total reactive power of each particle using k optimal particles and a random value between 0 and 1 (Equ(8)).

$$F_i^d(t) = \sum_{j \in kbest, i \neq j} rand F_{ij}^d(t) \quad (8)$$

5. Calculation of resistance against changing the speed or direction of movement of the object (inertial mass using Equ(9).

$$a_i^d(t) = \frac{F_i^d(t)}{M_i^d(t)} \quad (9)$$

6. Calculating the speed and next location of the particle based on the inertial mass (Equ(10-11). In Equ(10), in order to overcome the local optimality problem and also the wide scanning of the particle problem space by using PSO algorithm, the particles, in addition to relying on the inertial mass, move towards the best particle with the greatest gravity or G_{best} and increase the power of global search or exploration.

$$v_i^d(t+1) = C1 * r1 * \left[v_i^d(t) - Gbest\right] + C2 * r2 * \left[v_i^d(t) - Lbest\right] + a_i^d(t) \quad (10)$$

$$x_i^d(t+1) = v_i^d(t+1) + x_i^d(t) \quad (11)$$

7. Updating the gravitational mass and inertia of each particle based on the best and worst value obtained (Equ(12 -13).

$$q_i(t) = \frac{fit_i(t) - worst(t)}{best(t) - worst(t)} \quad (12)$$

$$M_i(t) = \frac{q_i(t)}{\sum_{j=1}^{N} q_j(t)} \quad (13)$$

8. Returning to the step 2 of the algorithm until we have reached convergence or maximum execution.

3.3. Data classification

Biogeography Based Optimization (BBO) is an algorithm created by Dan Simon [32] that incorporates principles from the natural world. BBO utilizes a population of different habitats, each with distinct features such as rainfall and land area. These habitats, which are randomly generated, represent potential solutions. The fitness of each habitat is evaluated using the Habitat Suitability Index (HSI), which serves as BBO's fitness function. BBO involves a combination of dependent and independent variables, with Suitability Index Variables (SIVs) serving as independent variables and HSI as a dependent variable. Over time, the likelihood of habitats with high HSI values migrating to habitats with low HSI values is greater. Conversely, habitats with low HSI values tend to attract new habitats from outside with more force than those with a strong health information system. Despite their HSI values, habitats may experience unforeseen changes in their surroundings. The elitism solutions will be preserved for future generations.

The aforementioned principles achieve equilibrium between exploitation and exploration. BBO effectively employs these principles to enhance HIS for diverse habitats. As a result, habitats with elevated HSI serve as optimal solutions, while habitats with low HSI values are deemed ineffective. Thus, habitats with the lowest HSI assimilate SIV characteristics from neighboring habitats with high HSI, while the most robust habitats transfer their defining traits to nearby habitats. BBO plays a vital role in the current habitats by facilitating migration, which involves exchanging information between habitats, and mutation, which promotes population diversity enhancements. These deterministic factors engender novel solutions, or habitats, that ultimately lead to the attainment of the optimal solution over several generations.

The concept of migration in BBO can be paralleled to how various relatives contribute to the development of a single offspring through the evolutionary phase. Linear migration model of BBO allows for the liberation of the current solution and island, facilitating the process of adaptation for habitat Hi. The modification of likelihood Hi is predetermined by the influx rate of immigration, λ_i, while the released likelihood originates from Hi and aligns with the emigration rate μ_i. Individual i within the BBO system possesses unique λ_i and μ_i values, dependent on the number of spices present in the habitat. These values are expressed in Equ(14) to (15) respectively.

$$\lambda_i = I\left(1 - \frac{R_i}{n}\right) \quad (14)$$

$$\mu_i = E\left(\frac{R_i}{n}\right) \quad (15)$$

$$R_i = i - n \quad (16)$$

The mutation operator randomly modifies the environment utilizing the preliminary probability of habitat viability. Solutions with exceptionally low HSI and extremely high HSI are improbable to materialize. Nevertheless, solutions with medium HSI are comparatively improbable. The calculation for the mutation rate, mi, of the given solution, can be derived as Equ(17):

$$m = m_{max}\left(\frac{1 - p_i}{p_{max}}\right) \quad (17)$$

Within the workings of BBO, it is evident that the migration and mutation procedures effectively ensure the absence of duplicated features within a given habitat. At each execution, the option to manually define the quantity of classification is available. Consequently, each solution consistently retains a constant count of chosen features throughout the process.

The representation of habitats in BBO for the classification problem involves an array that holds information about SIVs. The values for SIVs can be 0 or 1, where 1 indicates the normal data in the habitat and 0 means abnormal data is present.

BBO is employed in classification with the aim of enhancing the effectiveness of the model while alleviating the issues associated with imbalanced data. Consequently, the HSI serves as a measure akin to the evaluation function in BBO.

Initially, the BBO solution is formed randomly with a subset of features. The HIS (Health Indicator Score), representing the fitness value, is primarily dependent on the number of features found in the island. On the other hand, the performance of HSI relies on the SIVs. To assess fitness, the study utilizes the accuracy rate of the classifier, following the recommendations of several related works. A 10-fold cross validation is employed to determine the accuracy rate of each solution. In HSI, a quality solution is characterized by a high accuracy rate, and vice versa. A good HSI solution shares its SIVs with a poor solution. This sharing occurs during the migration phase and is controlled by emigration and immigration rates of the habitats. The parameter values of BBO including population size, maximum mutation coefficient rate, maximum habitat probability and generations are considered 100, 0.005, 1 and 20, respectively.

4. Evaluation

The work done in this paper has been done using UNSW-NB15 dataset to detect intrusion in the network.

Network packets of the UNSW-NB15 dataset were generated by the IXIA PerfectStorm tool at the Australian Cyber Lab to create a mix of normal activity and artificial attack behaviors. This dataset has 9 types of attacks which are Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms. Twelve algorithms have been developed to generate 49 class-labeled features.

The total number of records is two million and 540,044, which are stored in four CSV files. A partition of this dataset is configured as a training set and a test set. The number of records in the training set is 175,341 records and the test set is 82,332 records of different types of attack and normal.

Experimental observations have been made for the labels in this dataset. For classification, the dataset contains two labels, one representing normal traffic and the other representing abnormal attack traffic. This dataset has 10 classified labels for multi-class detection, which has 9 labeled classes. For each labeled class, the performance metrics of the proposed method with and without considering feature reduction have been evaluated to show the effect of selecting the best features in the mentioned dataset.

4.1. Evaluation metrics

There are four possible outcomes that can be obtained from an entanglement matrix (for binary classification). Based on these four results from the desired matrix i.e. True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN), the overall results have been analyzed using 10 known evaluation criteria.

TP (True Positive): attack features that are correctly labeled by the model.

FP (False Positive): non-attack features that are incorrectly labeled by the model.

TN (True Negative): non-attack features that are correctly labeled by the model.

FN (False Negative): attack features that are incorrectly labeled by the model.

Below is a short definition of each of these 6 criteria along with their related equation.

Accuracy: This criterion is actually a fraction consisting of dividing the correctly classified samples by all samples of the test set (Equ(18)).

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (18)$$

F1 Score: This measure interprets the harmonic mean of recall and precision for a specific model (See Equ(19)).

$$F1-score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (19)$$

Specificity: This measure is also known as true negative rate (TNR) and represents the portion of normal data that is correctly labeled by the model (See Equ(20)).

$$Specificity = \frac{TN}{TN+FP} \quad (20)$$

Sensitivity: This measure shows the probability of a positive test result (See Equ(21)).

$$Sensitivity = \frac{TP}{TP+FN} \quad (21)$$

Precision: This measure is the fraction of relevant instances among the retrieved instances (See Equ(22)).

$$Precision = \frac{TP}{TP+FP} \quad (22)$$

Matthew's Correlation Coefficient (MCC): This measure is used to evaluate the difference between predicted and actual values (See Equ(23)).

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FN)*(TN+FP)*(TP+FP)*(TN+FN)}} \quad (23)$$

4.2. Evaluation results

In this paper, all the simulations are done using MATLAB simulation software. All operational tasks of this work have been implemented in the following hardware environment: Intel core i7-3000U CPU (3.70 GHz), 8 GB RAM and Windows 10 operating system.

The effectiveness of a model in intrusion detection depends on the values obtained from its evaluation criteria. In order to better observe and also to show the effect of choosing the best features in improving the results, the experiments were done once with the entire dataset and again with a part of the dataset that includes the selected features.

The main purpose of using optimal clustering in the proposed method is to reduce the complexity of the dataset and at the same time preserve the original data patterns. In this section, we have tried to detect the influence and calculate the evaluation parameters with and without feature reduction. From Table 1, it can be seen that for the UNSWNB15 dataset, all of the parameter values for the sets have increased somewhat. It should be noted that out of a total of 49 features in the dataset, 13 more important and effective features have been selected by applying the improved clustering method (All features with an impact higher than the average impact are selected as important features.). This number of features when applying normal clustering is 28 features, which has led to worse results (as shown in Table 1).

Since the benefits of clustering are in finding unrelated data in the dataset, it has been concluded that its use in the dataset is more effective. However, in addition to the reduction of complexity, due to the reduction of information, there is a risk of reduced performance when using reduced features, but on the other hand, the amount of computing time is also reduced.

In data mining, the ROC curve is a graphical way to show the balance between the true positive rate and the false positive rate of the model using AUC parameter. The optimal ROC curve has a higher true positive rate with a lower false positive rate. In fact, the area under this curve shows the usefulness of the model. Figure 1 shows the ROC curve for UNSW-NB15. Using the obtained results and examining Figure 1, it can be concluded that the use of enhanced clustering is useful in improving the performance of the model, that is, we will have more levels under the graph.

4.3. Comparison

In Table 2, a comparison has been made between the accuracy of the proposed method and some other works done in the past. The results show the superiority of the proposed method in most of the parameters for evaluating the effectiveness of the proposed intrusion detection system.

Table 1. Results with/ without clustering/ improved clustering

Parameter values	*Without clustering*	*With simple clustering*	*With improved clustering*
Accuracy	0.916	0.973	0.998
F1-score	0.843	0.935	0.991
Specificity	0.905	0.926	0.996
Sensitivity	0.883	0.941	0.998
Precision	0.816	0.953	0.993
MCC	0.832	0.931	0.982
AUC	0.783	0.944	0.981

Table 2. Comparison of proposed approach with some works

Approach	*Accuracy*
RF+LR [37]	0.981
ALO [38]	0.968
SSA [39]	0.953
K-Means+RF [40]	0.944
CNN + LSTM [41]	0.992
Our approach	0.998

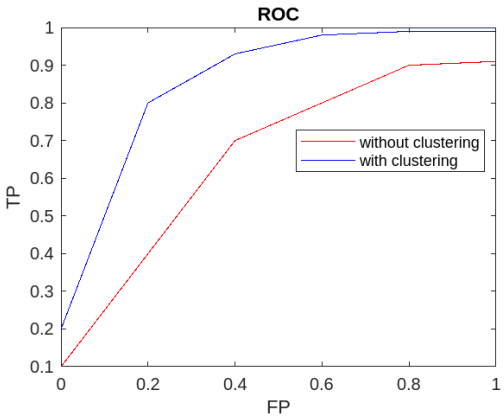

Figure. 1. ROC area of proposed method

5. Conclusion

In order to deal with the growing risks against data privacy through breaches in the network architecture, intrusion detection systems have become the main priority over the years. Although many studies have been conducted in this field, most of these studies do not use network traffic information properly. Considering this problem, in this paper, we presented a comprehensive study of network intrusion detection using a hybrid model. The UNSW-NB15 dataset has been investigated to address network traffic anomalies. The explosive nature of high-dimensional data can often compromise the computational process during real-world pattern analysis. As a suitable measure against

such a problem, the enhanced k-means clustering method was used, which has led to the reduction of dimensions while maintaining the useful features of the data. Also, BBO was used to classify dataset to normal and anomaly data. The results of the observations showed that by adopting the right policies in the data pre-processing stage, the complexity of the model is significantly reduced and along with using the proposed algorithm. The results are obtained with higher accuracy than the previous methods.

The use of other methods based on artificial intelligence as well as changes in the amount of parameters related to them can be considered as future solutions and future research.

Declarations

Funding

This research did not receive any grant from funding agencies in the public, commercial or non-profit sectors.

Authors' contributions

ATS: Study design, acquisition of data, interpretation of the results, drafting the manuscript, revision of the manuscript.

Conflict of interest

The author declares that no conflicts of exist.

References

[1] M. Capellupo, J. Liranzo, M.Z.A. Bhuiyan, T. Hayajneh, and G. Wang, “Security and Attack Vector Analysis of IoT Devices,” International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage, Guangzhou, 2017, pp. 593-606.

[2] K. Xu, F. Wang, R. Egli, A. Fives, R. Howell, \and O. Mcintyre, “Object-Oriented Big Data Security Analytics: A Case Study on Home Network Traffic.” International Conference on Wireless Algorithms, Systems, and Applications, Harbin, 2014, pp. 313-323.

[3] Y. K. Saheed, A. I. Abiodun, S. Misra, M. K. Holone, and R. Colomo-Palacios, “A machine learning-based intrusion detection for detecting internet of things network attacks,” Alexandria Engineering Journal, vol. 61, pp. 9395-9409, 2022.

[4] M. Bagaa, T. Taleb, J. B. Bernabe, and A. Skarmeta, “A machine learning security framework for iot systems,” IEEE Access, vol. 8, pp. 114066- 114077, 2022.

[5] V. Kumar, H. Chauhan, and D. Panwar, “K-means clustering approach to analyze NSL-KDD intrusion detection dataset,” International Journal of Soft Computing and Engineering (IJSCE), vol. 3, pp. 1-4, 2013.

[6] Y.N. Soe, P. I. Santosa, and R. Hartanto, “Ddos attack detection based on simple ann with smote for iot environment,” fourth international conference on informatics and computing (ICIC), 2019, pp. 1-5.

[7] B. Subba, S. Biswas, and S. Karmakar, “A neural network based system for intrusion detection and attack classification,” twenty second national conference on communication (NCC), 2016, pp. 1–6.

[8] K. A. Taher, B. Mohammed Yasin Jisan, and M.M Rahman, “Network intrusion detection using supervised machine learning technique with feature selection,” international conference on robotics, electrical and signal processing techniques (ICREST), 2019, pp. 643–646.

[9] Y. Yang, L. Wu, G. Yin, L. Li, and H. Zhao, “A survey on security and privacy issues in internet-of-things,” IEEE Internet of Things Journal, vol. 4, pp. 1250–1258, 2017.

[10] P. Maniriho, E. Niyigaba, Z. Bizimana, V. Twiringiyimana, L.J. Mahoro, and T. Ahmad, “Anomaly-based intrusion detection approach for iot networks using machine learning,” international conference on computer engineering, network, and intelligent multimedia (CENIM), 2020, pp. 303–308.

[11] M. Mamdouh, M.A. Elrukhsi, and A. Khattab, “Securing the Internet of Things and Wireless Sensor Networks via Machine Learning: A Survey,” International Conference on Computer and Applications, 2018, pp. 215-218.

[12] D. Geneiatakis, I. Kounelis, R. Neisse, I. Nai-Fovino, G. Steri, and G. Baldini, “Security and Privacy Issues for an IoT Based Smart Home,” 40th International Convention on Information and Communication Technology, Electronics and Microelectronics, 2017, pp. 1292-1297.

[13] S. Gurung, M.K. Ghose, and A. Subedi, “Deep learning approach on network intrusion detection system using NSL-KDD dataset,” International Journal of Computer Network and Information Security,” vol. 11, pp. 8-14, 2019.

[14] P. Van Huong, and D.V. Hung, “Intrusion detection in IoT systems based on deep learning using convolutional neural network,” 6th NAFOSTED Conference on Information and Computer Science (NICS), 2019, pp. 448-453.

[15] M. Choras, and M. Pawlicki, “Intrusion detection approach based on optimised artificial neural network,” Neurocomputing, vol. 452, pp. 705-715, 2021.

[16] R. Qaddoura, A. Al-Zoubi, I. Almomani, and H. Faris, “A multi-stage classification approach for iot intrusion detection based on clustering with oversampling,” Applied Sciences, vol. 11, 2021.

[17] T. Saba, T. Sadad, A. Rehman, Z. Mehmood, and Q. Javaid, “Intrusion detection system through advance machine learning for the internet of things networks,” IT Professional, vol. 23, pp.58–64, 2021.

[18] M. Roopak, G.Y. Tian, and J. Chambers, “An intrusion detection system against ddos attacks in iot networks,” 10th annual computing and communication workshop and conference (CCWC), 2020, pp. 562-567.

[19] N. Moustafa, and J. Slay, “A hybrid feature selection for network intrusion detection systems: Central points,” arXiv 2017, arXiv:1707.05505.

[20] H. Gharaee, and H.A. Hosseinvand, “A new feature selection IDS based on genetic algorithm and SVM,” Proceedings of the 2016 8th International Symposium on Telecommunications (IST), 2016, pp. 139-144.

[21] M. Belouch, S. El Hadaj, and M. Idhammad, “A two-stage classifier approach using reptree algorithm for network intrusion detection,” International Journla of Advanced Computing Sciences and Applications, vol. 8, pp. 389–394, 2017.

[22] A. Derhab, A. Aldweesh, A.Z. Emam, and F.A. Khan, “Intrusion detection system for internet of things based on temporal convolution neural network and efficient feature engineering,” Wireless Communications and Mobile Computing, 2020.

[23] M. Ge, X. Fu, N. Syed, Z. Baig, G. Teo, and, A. Robles-Kelly, “learning-based intrusion detection for IoT networks,” IEEE 24th pacific rim international symposium on dependable computing (PRDC), 2019, pp. 256-25609.

[24] M. Lopez-Martin, B. Carro, A. Sanchez-Esguevillas, and J. Lloret, “Network traffic classifier with convolutional and recurrent neural networks for Internet of Things,” IEEE access, vol. 5, pp. 18042-18050, 2017.

[25] X. Yihan, X. Cheng, Z. Taining, and Z. Zhongkai, "An intrusion detection model based on feature reduction and convolutional neural networks," IEEE Access, vol. 7, pp. 42210-42219, 2019.

[26] M.M. Baig, M.M. Awais, E.S.M. El-Alfy, "A multiclass cascade of artificial neural network for network intrusion detection," Journal of Intelligent and Fuzzy Systems," vol. 32, pp. 2875–2883, 2017.

[27] M. Al-Zewairi, S. Almajali, and A. Awajan, "Experimental evaluation of a multi-layer feed-forward artificial neural network classifier for network intrusion detection system," Proceedings of the 2017 International Conference on New Trends in Computing Sciences (ICTCS), 2017, pp. 167-172.

[28] S. Guha, S.S. Yau, A.B. Buduru, "Attack detection in cloud infrastructures using artificial neural network with genetic feature selection," Proceedings of the 2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 2016, pp. 414-419.

[29] K.K. Nguyen, D.T. Hoang, D. Niyato, P. Wang, D. Nguyen, and E. Dutkiewicz, "Cyberattack detection in mobile cloud computing: A deep learning approach," Proceedings of the 2018 IEEE wireless communications and networking conference (WCNC), 2018, pp. 1–6.

[30] N. Moustafa, G. Misra, and J. Slay, "Generalized outlier gaussian mixture technique based on automated association features for simulating and detecting web application attacks," IEEE Transanctions on Sustainable Computing, vol. 6, pp. 245–256, 2018.

[31] N. Moustafa, E. Adi, B. Turnbull, J. Hu, "A new threat intelligence scheme for safeguarding industry 4.0 systems," IEEE Access, vol. 6, pp. 32910–32924, 2018.

[32] D. Simon, "Biogeography-based optimization," IEEE Transactions on Evolutionary Computation, vol. 12, pp. 702–713, 2008.

[33] D. Albashish, A.I. Hammouri, M. Braik, J. Atwan, and S. Sahran, "Binary biogeography-based optimization based SVM-RFE for feature selection," Applied Soft Computing, v. 101, pp. 1-19, 2021.

[34] R. Ananda, M.Z. Naf'an, A.B. Arifa, and A. Burhanuddin, "Sistem Rekomendasi Pemilihan Peminatan Menggunakan Density Canopy K-Means," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 4, pp. 172- 179, 2020.

[35] A.C. Benabdellah, A. Benghabrit, and I. Bouhaddou, "A survey of clustering algorithms for an industrial context," Procedia computer science, vol. 148, pp. 291-302, 2019.

[36] G. Zhang, C. Zhang, and H. Zhang, "Improved K-means Algorithm Based on Density Canopy," Knowledge-Based Systems, vol. 145, 2018.

[37] S. Waskle, L. Parashar, and U. Singh, "Intrusion detection system using PCA with random forest approach," International Conference on Electronics and Sustainable Communication Systems (ICESC), (ICESC), 2020, pp. 803-808.

[38] M. Mafarja, I. Aljarah, A.A. Heidari, A.I. Hammouri, H. Faris, A.-Z. Ala'M, S. Mirjalili, "Evolutionary population dynamics and grasshopper optimization approaches for feature selection problems," Knowledge-Based Systems, vol. 145, pp. 25–45, 2018.

[39] M. Mafarja, I. Aljarah, H. Faris, A.I. Hammouri, A.-Z. Ala'M, and S. Mirjalili, "Binary grasshopper optimisation algorithm approaches for feature selection problems," Expert Systems and Applications, vol. 117, pp. 267-286, 2019.

[40] K. Samunnisa, G. Kumar, and K. Madhavi, "Intrusion detection system in distributed cloud computing: hybrid clustering and classification methods," Measurement: Sensors, vol. 25, 100612, 2023.

[41] S. Sivamohan, S.S. Sridhar, and S. Krishnaveni, "An effective recurrent neural network (RNN) based intrusion detection via bi-directional long short-term memory," International Conference on Intelligent Technologies (CONIT), IEEE, 2021, pp. 1-5.

Aliakbar Tajari Siahmarzkooh is an Assistant Professor at Faculty of Computer Sciences, Golestan University. His research interests include artificial intelligence, data mining and data security. He graduated from University of Tabriz, majoring in intrusion detection systems.