Cyber-security attack prediction using cognitive spectral clustering technique based on simulated annealing search

N.R. Rajalakshmi[1], Sathishkumar V. E.[2], C.Kannika Parameshwari[2], Maheshwari V.[3] and Prasanna M.[3]

[1]Department of Computer Science and Engineering, Vel Tech Ragarajan Dr.Sagunthala R& D Institute of Science and Technology, Avadi-600062, TN, India.
[1]Department of Industrial Engineering, Hanyang University, Seoul, Republic of Korea
[2]Department of Electronics and Communication Engineering, NPR College of Engineering and Technology, Natham - 624401, TN, India.
[3]School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India.

maheshwarivenkat24@gmail.com

Abstract. Data protection and security is a big challenging portion in a modern technical world against the cyber-attacks like; ransomware, man-in-the-middle, DDoS, etc. In order to overcome this scenario, there are lot of artificial intelligence framework have been introduced to detect and classify the cyber-attacks. In particular, neural networks, with their solid speculation execution ability, are capable to address an extensive variety of cyber-attacks. This article frames the training and testing of a neural network group such a way to deal with detection of cyber-attack using cognitive spectral clustering technique based on simulated annealing search method. The optimization of individual networks can be made by using adaptive memetic algorithm with simulated annealing search. It is used to enhance the neural network weights and hidden neurons respectively. This algorithm is a combination of both local and global search enhancement method and used to get rid of the premature convergence, and used to achieve the adaptive search output. The testing outcome of the proposed framework shows a better result 99.5% of overall accuracy, and effectively adaptive in terms of detecting the cyber-attacks.

Keywords: cyber-attack, cognitive spectral clustering, artificial intelligence, simulated annealing search, neural network cluster.

1. Introduction

A new cyber-attack is anticipated to begin every 40 seconds. Despite the speed and complexity with which threat actors have evolved in recent years, there is still reason to be positive about protection. For the first time, we have all of the components in place to forecast cyber incidents. After all, the predictors have always existed [1]. They're simply concealing themselves in plain sight. The use of artificial intelligence and machine learning allows cyber defenders like Looking-Glass to sift through massive amounts of data to uncover the insight that truly counts [2]. Cyber security experts place a higher emphasis on approaches to risk assessment and mitigation Creating effective techniques was a method defined across the subject of cyber security. Despite its success in cyber defense, machine learning is becoming a growing concern in data security [3]. Cloud computing, networking, and evolutionary computation have grown rapidly as a result of remarkable advances in computing, storage, and computational technologies [4]. As the world becomes more digitalized, there is a growing requirement for comprehensive and advanced privacy and security issues, as well as tactics to combat security threats that are getting more complex [5]. Cyberwarfare is expanding around the world by exploiting various computer flaws. Machine learning methods were employed to combat worldwide computer security threats such as malware detection, ransomware identification, fraud detection, and spoofing detection. Its study examines how cyber training is utilized in both defense and offence, offering information on cyber risks and machine learning approaches [6]. The managed entities may engage in both passively (eavesdropping, not participating) and active (directly attacking) attacks. The term "intrusion detection" refers to the practice of continuously keeping tabs on what's happening inside a network or system, parsing out any suspicious activity, and then taking measures to block it [7].

2. Related work

The number malware attacks are increased rapidly and it mainly focusing the data transaction and Internet of things gadgets. These attacks are very smart in the flow, and it would induce the user in order to gain the attention. Cyber-attacks do not focus or targeting the machine, where as they may look like hidden information, and they invite the user to inject themselves, and they do not realize the attack [8]. Cybercrime has evolved as one of the world's most critical challenges. They cause substantial financial damage to individuals and governments every day. The surge in cyber-attacks is paralleled by an increase in cybercrime. Identifying cybercrime offenders and knowing attack strategies are essential components in the fight against crime and criminals [9]. Cyber-attacks are difficult to identify and avoid. Researchers, on the other hand, have recently addressed these issues by developing security models and making predictions using artificial intelligence technology. Several crime prediction techniques are available in the literature [10]. Meanwhile, they can't anticipate new forms of cybercrime or cyberattacks. If an assault could be traced back to its source using real data, it could be possible to resolve this problem. The details include the kind of crime, the gender of the perpetrator, the extent of the damage, and the methods of assault [11]. Users who have fallen victim to a cyber-attack may file an application to forensic units in order to restore their data. In this research, we use machine learning techniques to compare two models of cybercrime and to measure the effectiveness of the supplied attributes in identifying both the cyberattack method and the offender. We used eight different machine-learning strategies and discovered that the relative accuracy was quite close. The Support Vector Machine Linear was found to be the most effective cyber-attack method with an accuracy rate of 96.02% [12].

3. Proposed system model

The proposed system in this study is divided into three parts. The intrusion detection model is built with simulated annealing search for data clustering and classification, cognitive spectral clustering and neural networks for training, and a support vector machine for learning and detection.

3.1. Dataset availability

The suggested method employs experiments on the KDD99 data set, that contains a wide spectrum of interference and normal operations replicated in a defense network. Simulated threats are classified as one of four types: DoS, R2L, U2R, and Probing. Each instance in the collection comprises the extracted characteristics of a connection record. Test results are generated utilizing 312,029 data, about 31,000 of which are utilized for training and the remainder for testing.

3.2. Preprocessing

This is the first stage of our model, and it is regarded an early step toward fulfilling our model's main aim by organizing data and translating it into a suitable input format for subsequent phases. The Codification module codes features into an appropriate style after capturing network data, while Clustering removes input data features to reduce space feature dimensions. The simulated annealing module is used to group the various assault types. Figure 1 depicts the framework of the preprocessing phase. This phase is made up of the modules listed below. Capturing is the model that gathers network data from the environment, whereas Codification is the process of coding characteristics into a certain style. Feature extraction is a technique for transforming input data into a set of features that can subsequently be stored in a database.

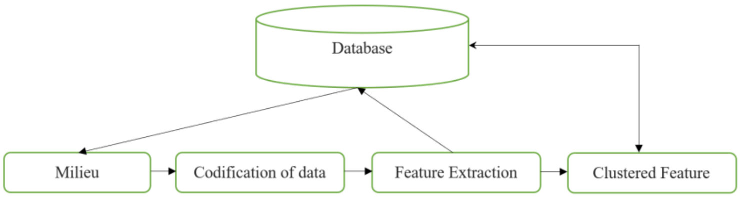

Figure 1. Preprocessing stage.

Clustering: Using the SA method, the dataset is divided into 22 groups based on different working features, with the major five categories depending on incursion form.

3.3. Training mode

This second stage comprises the training unit, which is constructed using cognitive spectral clustering as supervised learning with fixed weight and the SA algorithm to increase its performance. Cognitive spectral clustering functions as a classifier, but only for attack-type behaviors. SA may also function as a classifier, but only for behaviors that have not yet been classified as attacks or not, and if yes, it identifies the sort of attack. As seen in Figure 2, training operates as follows.

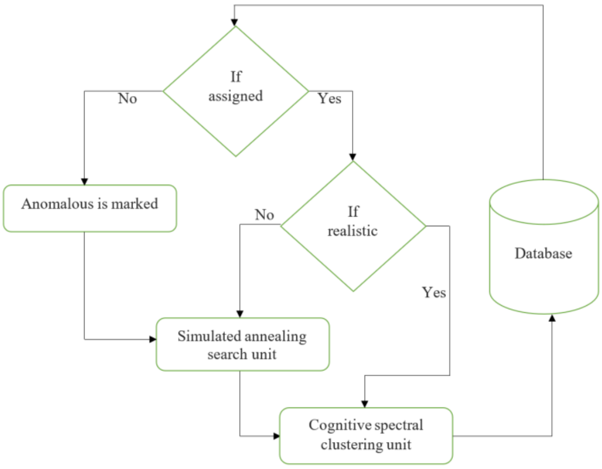
If assigned
No
Yes
Anomalous is marked
No
If realistic
Yes
Database
Simulated annealing search unit
Cognitive spectral clustering unit

Figure 2. Training session.

```
Start;
    Step 1: Extract Data Pattern Characteristics
    Step 2: Evaluate to see if this data pattern has a label.
    Step 3: Analyze whether the indicated data pattern is acceptable or inappropriate.
        Step 4: Normal data patterns should be sent to the Cognitive spectral clustering unit,
          whereas aberrant data patterns should be sent to the SA entity.
    Step 5: The SA unit creates new vectors if the data pattern is designated as abnormal.
    Step 6: Set the data pattern to abnormal whether it stays unmarked.
    Step 7: Send the data pattern to the Cognitive spectral clustering unit once the SA unit
      has clustered and labelled it.
End;
```

The outcome of the training stage is often gathered in order to broaden the pool of incursion vectors compared in classification mode. This aids in the identification of additional intrusions.

3.4. Cognitive spectral clustering

Spectral clustering approaches in statistical models employ the spectrum of the data's similarity matrix to conduct dimensionality reduction before grouping in fewer dimensions. As an input, the similarity matrix is supplied, which is a quantitative estimate of the similar technology of every pair of values in the database. Given r data co-ordinates $a_1, a_2, a_3 \dots, a_n$ and the correlation function $f(\|a_i - a_j\|, \alpha)$, then the weight matrix is expressed by,

$$X_{i,j} = f(\|a_i - a_j\|, \alpha) \tag{1}$$

However, the correlation between the two data co-ordinates based on the range between the co-ordinates and a parameter scaling α, for the gaussian connection function is,

$$f(\|a_i - a_j\|, \alpha) = e^{\frac{-\|a_i - a_j\|^2}{2\alpha^2}} \tag{2}$$

The parameter of the scaling is computing the local framework of the link between the co-ordinates. The degree matrix is expressed by,

$$Deg_{i,j} = \sum_{j=1}^{r} X_{i,j} \tag{3}$$

$$L_m = Deg_{i,j} - X_{i,j} \tag{4}$$

The Laplacian matrix is denoted as L_m, and the eigen value and vector of L_m, is utilized to make data group. The normalization is done prior the Laplacian in order to calculate the decomposition value of the spectral. The percentage of each cluster that is made up by the most common category is shown in the second column, while the classification performance for each category is shown in the first column (assuming attacks are assigned to the least common category within each cluster).

3.5. Prediction phase

Instead of the cognitive spectral clustering, the support vector machine unit is used as a classifier in this phase to enhance intrusion detection. To optimize the AR rate and decrease FPR, the support vector machine technique is best suited for creating a new IDS. Vector samples from the environment can be collected and compared to the vector collection from the clustering and training phases. Figure 3 depicts the operation of a support vector machine, which is separated into the following phases. Examine the new vectors from the testing data to the group created by the intersection of the key cluster vectors and the simulated annealing Model. Identify the current target class using the main (recorded) attack kinds. Each of these phases delivers the essential data to the other phases, which all work together to achieve the main purpose, which is to detect various forms of infiltration.

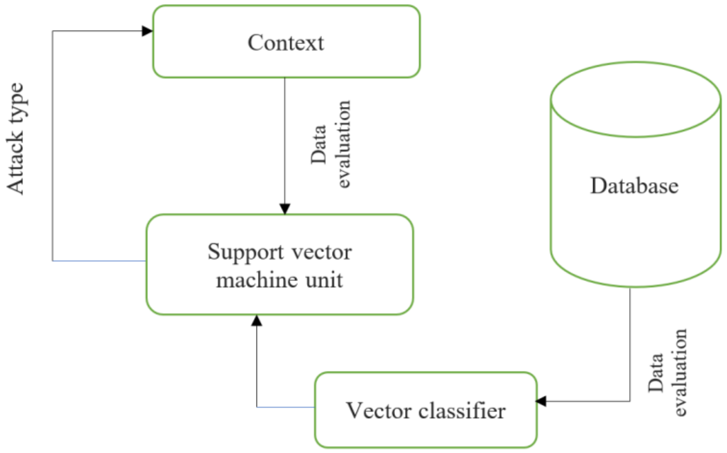

Figure 3. Prediction phase work flow.

4. Results and discussion

The spectral clustering findings are shown in Table 1. First, the percentage of each cluster consisting of the most frequent category is shown, and then the classification accuracy for each category is displayed in the next column (given that attacks are assigned to the least common category within each cluster).

Table 1. Classification accuracy of proposed and existing techniques.

Technique	Classification accuracy of attacks					Global accuracy (%)
	Normal	DoS	Probe	U2R	R2L	
Proposed system	99.12	99.26	99.49	99.77	99.84	99.5
Backpropagation neural network computation	96.78	99.56	99.34	-	97.31	98.2
Hybrid Neural network model	-	96.61	97.11	97.79	98.22	97.4

DoS and Normal traffic were classified with a high degree of accuracy, whereas Probe and U2R (Probe intrusions were often incorrectly labelled as U2R intrusions and U2R) attacks were classified with a lower degree of accuracy (99.5%). Classification accuracy is summarized in Table 1 for both new and current methods, and it is evident that the suggested framework achieved a higher worldwide classification accuracy rate for cyber-security forecasting. The worldwide classification accuracy of the proposed model is compared to that of other existing models. The suggested method works well with anomaly and misuse detection systems that work together, and threat clustering reduces the complexity of the problem and classifies threats into more manageable buckets, improving both accuracy and flexibility. To a best of knowledge, this is the first study to take a systematic look at online encroachment.

References

[1] W. K. AL-Rashdan, R. Naoum, and A. S. Wafa'S, "Novel network intrusion detection system using hybrid neural network (Hopfield and Kohonen SOM with conscience function)," IJCSNS, vol. 10, no. 11, p. 10, 2010.

[2] Alves LGA, Ribeiro HV, Rodrigues FA. 2018. Crime prediction through urban metrics and statistical learning. Physica A: Statistical Mechanics and its Applications 505:435–443.

[3] Arora T, Sharma M, Khatri SK. 2019. Detection of cyber-crime on social media using random forest algorithm. In: 2nd International Conference on Power Energy, Environment and Intelligent Control (PEEIC). Piscataway: IEEE, 47–51.

[4] Bharathi ST, Indrani B, Prabakar MA. 2017. A supervised learning approach for criminal identification using similarity measures and K-Medoids clustering. In: ICICICT. Piscataway: IEEE, 646–653.

[5] Kuzlu, M., C. Fair, and O. Guler, Role of artificial intelligence in the Internet of Things (IoT) cybersecurity. Discover Internet of things, 2021. 1 (1): p. 1-14.

[6] Truong, T. C., et al., Artificial intelligence and cybersecurity: Past, presence, and future, in Artificial intelligence and evolutionary computations in engineering systems. 2020, Springer. p. 351-363.

[7] Azim, A. W., Bazzi, A., Shubair, R., & Chafii, M. (2022). Dual-Mode Chirp Spread Spectrum Modulation. IEEE Wireless Communications Letters, 1-1. doi:10.1109/LWC.2022.3190564.

[8] Bazzi, A., & Meilhac, L. (2022). Method for decoding an rf signal bearing a sequence of symbols modulated by cpm and associated decoder: Google Patents.

[9] Bazzi, A., & Slock, D. (2020). Robust Music Estimation Under Array Response Uncertainty. Paper presented at the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[10] Njima, W., Bazzi, A., & Chafii, M. (2022). DNN-based Indoor Localization Under Limited Dataset using GANs and Semi-Supervised Learning. IEEE Access, 10, 69896-69909.

[11] Reddy, V. A., Bazzi, A., Stuber, G. L., Al-Dharrab, S., Mesbah, W., & Muqaibel, A. H. (2020). ¨ Fundamental Performance Limits of mm-Wave Cooperative Localization in Linear Topologies. IEEE Wireless Communications Letters, 9 (11), 1899-1903.

[12] Ghelani, D., & Hua, T. K. (2022). Conceptual Framework of Web 3.0 and Impact on Marketing, Artificial Intelligence, and Blockchain. International Journal of Information and Communication Sciences, 7 (1), 10.