A Hybrid Metaheuristic Approach for Intrusion Detection System Using SVM and ACO Technique.

Ms. Utkarsha Jiwane
Ph.D. Scholar, SSES Amt's, Science College, Department of Computer Science, Congress Nagar, Nagpur, India.

Abstract:

Intrusion detection systems (IDS) are essential for safeguarding computer networks against unauthorized access and malicious activities. This paper proposes an approach for intrusion detection that focuses on the strengths of Support Vector Machines (SVM) and Ant Colony Optimization (ACO) to identify an optimal subset of features that maximize classification accuracy by simulating the foraging behavior of ants. The proposed hybrid approach combines the high classification accuracy of SVM with the optimization capabilities of ACO to enhance the detection of various types of intrusions in network traffic. In this approach, SVM is employed to classify network activities into normal and suspicious categories based on selected features. However, the performance of SVM heavily relies on the optimal selection of its parameters which directly impacts its detection accuracy. To address this, ACO is introduced as a metaheuristic optimization technique to ensure the highest possible classification performance. The proposed method is validated through experiments on cybersecurity attacks dataset. The results demonstrate that the SVM-ACO approach outperforms traditional methods in terms of efficient processing, reduced execution time and improved accuracy with 99.91%. These findings show the effectiveness of the SVM-ACO approach in providing efficient solution for complex intrusion detection challenges.

Keywords: Intrusion detection systems, Cybersecurity, Feature Selection, Support Vector Machines, Ant Colony Optimization, network traffic, Metaheuristic Optimization, Network Security, Efficient Processing, Execution Time Optimization, accuracy.

1. Introduction

Security plays a crucial role by employing a range of measures and practices designed to protect systems, data and networks from unauthorized access, misuse or damage. Current security technologies such as encryption, firewalls and access control are commonly used but still do not provide complete security. System security can be further enhanced by implementing Intrusion Detection Systems (IDS) [1]. IDS are security mechanisms designed to monitor and analyze network or system activities for signs of malicious behavior, policy violations or other forms of intrusion [2]. These systems play a crucial role in identifying and responding to potential threats before they can cause significant damage. The ability of an IDS to classify a wide variety of intrusions in real time with high accuracy is crucial. An IDS dynamically monitors system events and determines whether these events indicate an intrusion attack or constitute legitimate use of the system [3]. Based on the data analysis approach an IDS may belong to two main groups misuse detection (or signature-based detection) or anomaly detection [4]. The misuse detection approach is the most widely used to detects only known attacks that have their signatures included in a database [5]. In contrast, the anomaly detection approach creates a normal behavior profile and detects intrusions based on significant deviations from this profile. Several challenges must be considered when building an intrusion

detection model such as achieving a high attack detection rate. Since the advent of IDS multiple techniques have been proposed to improve system performance [6]. SVM have become a popular machine learning approach in intrusion detection due to their efficient performance, absence of local minima and fast execution time. This paper proposes a hybrid approach for intrusion detection that focuses the strengths of SVM and ACO to identify an optimal subset of features that maximize classification accuracy by simulating the foraging behavior of ants [7]. SVM helps in distinguishing normal and malicious activities, efficient in handling high dimensional data while ACO is efficient in identifying the most relevant subset of features from high-dimensional network traffic data, reducing computational overhead, efficient in optimize feature subsets and parameter tuning for SVM for better performance by enabling it to find global optima by avoiding local minima.

2. Related Work

Ali et al. [18] proposed a Suspicious Pattern Detection (SPD) algorithm for identifying suspected cyber threats in instant chat messengers available on Social Networking Websites and Instant Messengers. The framework incorporates an Ontology-Based Information Extraction (OBIE) technique combined with a pre-defined knowledge base and a data mining approach using Association Rule Mining (ARM). The proposed framework consists of three main steps word extraction from unstructured text, monitoring system program implementation of SPD algorithm. The approach was tested using the Global Terrorist Database (GTD) and compared against other instant messengers, mobile phone applications and social networking sites based on their ability to detect suspicious information during online chats. The evaluation demonstrated that the proposed framework outperforms alternatives showing greater efficiency and effectiveness in detecting suspicious patterns based on the parameters considered.

Hosseinkhani et al. [19] presented a comprehensive review of existing crime data mining techniques for detecting suspicious information on the web. One of the key challenges highlighted in their work is the rapid growth in the volume of cyber data and increasing network traffic coupled with the diverse formats of available data such as audio, video and text. The study provides a theoretical overview of various mining concepts including data mining, web mining and crime data mining emphasizing their relevance to crime detection. Specifically, the review focuses on textual data mining techniques for identifying criminal activities. Key techniques discussed include sequential pattern mining, classification, association rule mining and clustering among others. These methods provide valuable insights into the effective detection of suspicious activities within complex and voluminous datasets.

Alami et al. [20] proposed a similarity distance evaluation-based approach to differentiate suspicious content from authentic content, posts or blogs. The analysis utilized a dataset comprising Twitter text corpus. The proposed method evaluates a similarity index by comparing social database entries. The framework involves three main steps text corpus collection, corpus processing and classification using the similarity-based approach. The performance of the proposed method was assessed using the similarity distance index. However, the results indicated longer execution times and lower precision rates highlighting the need for improvements in both precision and processing efficiency to enhance its overall effectiveness.

Jain et al. [21] proposed an approach to differentiate between actual information and rumors in Twitter data. The authors extracted Twitter data related to specific topics using hashtags. For concept validation they considered data from well-known news channels and evaluated the results through semantic and sentiment analysis of the tweets. As part of the proposed concept the authors also developed a prototype called “The Twitter Grapevine,” specifically designed to target rumors in Indian domains. The overall results were evaluated based on accuracy

analysis for initial experiments on rumors related to Digital India and facebook.org followed by experiments on Kerala House and Beef Rumors. In these experiments both favorable and unfavorable predictions were evaluated.

Murugesan et al. [22] employed a statistical corpus-based data mining approach for detecting suspicious activities on online forums. The authors focused on the textual data of online forums and provided a detailed explanation of the process for extracting suspicious information. After performing pre-processing steps such as stop word removal and stemming using the Brute Force algorithm the authors applied a matching algorithm for recognizing suspicious keywords. Finally they utilized keyword spotting techniques a learning-based method and a hybrid approach of the defined techniques to comprehensively detect suspicious human activity.

Harsh Arora et al. [23] online forums, primarily text-based, have become key sources for user opinions and information, but they are also prone to misuse by spammers, fraudsters and other malicious actors. Existing tools and methods aim to detect suspicious activities in online forums, but they have certain limitations. To address this research proposes an integrated framework combining Support Vector Machine (SVM) and Particle Swarm Optimization (PSO) where SVM is used for data classification and PSO optimizes its parameters to improve detection accuracy and efficiency.

Several studies have evaluated on intrusion detection. An IDS dynamically monitors the events taking place in a system and decides whether these events are symptomatic of an attack (intrusion) or constitute a legitimate use of the system [11][12]. Many challenges need to be considered when building an intrusion detection model such as obtaining a high attack Detection Rate without generating many false alarms (low False Alarm Rate). Since the appearance of IDS multiple techniques have been proposed in order to improve the performances of these systems [13]. Recently, several machine learning techniques have been applied Kruegel and Toth [14] proposed an approach based on decision trees for classifying the detection rules. Xiao et. al. [10] present a hybrid model based on information theory and genetic algorithm to detect network attacks. Their approach considers only discrete features. Chen, Abraham and Yang [15] proposed a flexible neural tree (FNT) model for intrusion detection based on neural tree for attribute selection and particle swarm optimization for parameter optimization. In 2012 Pu, Xiao and Dong [3] use an improved ant colony algorithm in order to determine the best parameters for the SVM classifier. Support Vector Machines is a novel machine learning approach that has become a popular research method in intrusion detection [16] [17]. Our findings were more similar to those obtained by P. Shinde et al. [1] who have reported that survey on proposed an IDS model that combines IG feature selection. With the SVM classifier model can obtain better results in terms of attack detection rate, false alarm rate and accuracy.

In our study, ACO-SVM a hybrid approach method is used to achieve an optimal balance between feature selection and classification performance by focusing on a subset of critical features the model avoids overfitting and achieves better efficiency.

3. Methodology

3.1 Ant Colony Optimization

ACO draws inspiration from the social behavior of ant colonies. It is observed that a group of ants can collectively determine the shortest route between their food source and their nest [24]. In nature ants deposit a chemical substance called pheromone on their paths which helps other ants find the shortest route between their nest and a food source. The pheromone trails left by ants influence the decision-making process of other ants guiding them towards the most efficient routes. Ant Colony Optimization method minimize redundancy by selecting a subset

of features each ant in relation to the previously selected features selects the lowest similarity features. Therefore, if a feature is selected by most ants indicates that the features have the lowest similarity with the other features. The features receive the largest amount of pheromone and the chances of its selection by other ants will be increased in subsequent iterations. Finally, by considering the similarity between the features the selected main features will have high pheromone values [25].

3.2 Support Vector Machine

SVM are often regarded as classifiers that achieve high accuracy across various tasks. They work by constructing a hyperplane with the maximum Euclidean distance or margin from the nearest training examples. Simply put SVM represent instances as points in space which are mapped to a high-dimensional plane where the instances of different classes are separated by the largest possible margin from the hyperplane [26]. New instances are mapped into the same space and based on which side of the hyperplane they fall predicted to belong to a particular class. The SVM hyperplane is determined by a relatively small subset of the training data known as support vectors while the rest of the training data does not influence the final classifier.

3.2 ACO-SVM Hybrid Algorithm

In this method of hybridization**,** ACO is used to identify the optimal subset of features that maximizes the classifier's accuracy. The fitness function for ACO is defined to evaluate the accuracy of an SVM classifier (SVC) on the selected feature subset. ACO is applied to the subset of features to further refine or confirm the feature selection. After feature selection using ACO an SVM classifier is trained on the selected features. Accuracy, precision, F1 score and other performance metrics are calculated for ACO selected feature subsets. The SVM classifier's performance is evaluated on the test set during iteration. Execution time is recorded for ACO to compare the computational efficiency of the model's performance [8].

4. Result and Discussion

The proposed experiment was conducted on a laptop equipped with an Intel i5 8th Gen processor and 8 GB of RAM running the Windows 10 operating system. We utilized a dataset containing 100 records obtained from Kaggle titled "Cybersecurity Attacks" and the coding was performed in Python. The dataset was divided into two parts: 80% for training and 20% for testing. The dataset comprises a total of 25 features as shown in Table 1.

Attributes
Timestamp, Source IP Address, Source Port, Destination Port, Protocol, Packet Length, Packet Type, Traffic Type, Payload Data, Malware Indicators, Anomaly Scores, Alerts/Warnings, Attack Type, Attack Signature, Action Taken, Severity Level, User Information, Network Segment, Geo-location Data, Proxy Information, Firewall Logs, IDS/IPS Alerts, Log Source

Table 1. Dataset attributes

This experiment demonstrates the efficiency of combining ACO and SVM classifier for feature selection in supervised machine learning. Specifically, it highlights how metaheuristic optimization techniques can improve model performance by identifying the most relevant features from a dataset.

- **Performance Metrics**: Accuracy, precision, F1 score and classification reports are generated for ACO feature selection methods and SVM classifier.
- **Feature Selection**: The dataset contains a number of features but not all may be equally relevant for predicting the target variable. Selecting a subset of relevant features can improve the classifier's accuracy and reduce computational complexity.
- **Hybrid Approach**: The hybrid approach by combining the global search capability of metaheuristics ACO algorithm with the classification power of SVM to achieve effective feature selection and model performance.

Based on Figure 1, the ACO-SVM model is shown to be a better solution for achieving higher accuracy.

Firstly, preprocessing of handling the missing attributes was done among the available features, features that have more rows without value are removed from the set of features. The Second part is splitting of data. The third part is identifying the related attributes which we are focus for IDS by targeting the features [9]. The initial feature selection performed by ACO identify the most relevant features based on accuracy. The hybrid approach significantly improved performance metrics for evaluation on classification accuracy, precision, F1 scores and detailed classification reports. These metrics illustrated the effectiveness of the feature selection methods in enhancing SVM performance, confirming that the selected features contributed effectively to the model's predictive power. The computational time of implementation was recorded to demonstrate the efficiency of the hybrid approach. Although execution time may vary with complexity the optimizations employed in both algorithms helped enhance the feature selection process resulting in efficient and effective model training and evaluation.

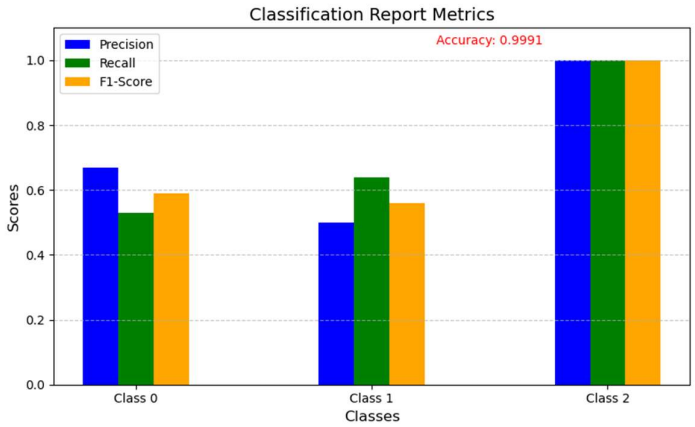

Figure 1. Model Performance

Best Feature Subset: ['Malware Indicators', 'Anomaly Scores', 'Attack Signature', 'Severity Level']
Best Accuracy: 0.9990833333333333

Classification Report:

	precision	recall	f1-score	support
0.0	0.67	0.53	0.59	15
1.0	0.50	0.64	0.56	11
2.0	1.00	1.00	1.00	11974
accuracy			0.9991	12000
macro avg	0.72	0.72	0.72	12000
weighted avg	0.99	0.99	0.99	12000

Accuracy: 0.9991
Execution Time: 126.69 seconds

5. Conclusion

The ACO algorithm efficiently navigates the high dimensional feature space to identify subsets of features that contribute most to the model's performance. The SVM classifier evaluates the predictive power of each selected subset with accuracy serving as the objective function to guide ACO. By hybrid approach these methods achieve an optimal balance between feature selection and classification performance. The hybrid ACO-SVM framework extended cybersecurity domain to identify key features and build robust, efficient models. In the context of cybersecurity attacks the selected features may provide valuable insights into the patterns or factors contributing to attacks or anomalies.

Overall, the ACO-SVM hybrid approach successfully demonstrates how optimization techniques can lead to superior feature selection and classification. By focusing on a subset of critical features the model avoids overfitting and achieves better efficiency, the classification accuracy of the ACO-SVM is maximized as reflected in the classification report, reduced feature dimensionality leads to faster training and conclusion making the approach suitable for real time cybersecurity applications.

6. References

[1] P. Shinde, T Parvat, "Analysis on: Intrusions Detection Based On Support Vector Machine Optimized with Swarm Intelligence", International Journal of Computer Science and Mobile Computing, Vol. 3 pg.559 – 566.

[2] X. R. Yang, J. Y. Shen and R. Wang, "Artificial immune theory based network intrusion detection system and the algorithms design", in Proc. of 2002 International Conference on Machine Learning and Cybernetics, Beijing, China, 2002, pp.73–77.

[3] M. Tavallaee, E. Bagheri, W. Lu and A. A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set", In Proc. of the 2009 IEEE symposium on computational Intelligence in security and defense application (CISDA), Ottawa, ON, Canada, 2009, pp.1–6.

[4] H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. 2nd edition, San Francisco: Morgan Kaufmann, 2005.

[5] M. A. Hall and G. Holmes. "Benchmarking Attribute Selection Techniques for Discrete Class Data Mining ", IEEE Transactions on knowledge and data engineering, vol. 15, no. 6, 2003, pp.1437–1447.

[6] D. Karaboga(2010), "Artificial bee colony algorithm", Scholarpedia [On-line], Vol. 5(3), pp.6915. Available: http://www.scholarpedia.org/article/Artificial_bee_colony_algorithm 2013.

[7] D. Karaboga and C. Ozturk, “A novel clustering approach: Artificial Bee Colony (ABC) algorithm”, Applied Soft Computing, Elsevier Science Publishers B. V. Amsterdam, The Netherlands, Vol. 11(1), pp. 652–657, Jan. 2011.
[8] B. Dehghan , M.Reza Sabri and A. Ahmadi1 et.al., “ Identifying the Factors Affecting the Incidence of Congenital Heart Disease Using Support Vector Machine and Particle Swarm Optimization”, May. 2023.
[9] Mastrogiannis N, Boutsinas B, Giannikos I. A method for improving the accuracy of data mining classification algorithms. Comput Oper Res 2009;36:2829-39.
[10] Cortez P, Embrechts MJ. Using sensitivity analysis and visualization techniques to open black box data mining models. Inf Sci 2013;225:1-17.
[11] T. Xiao, G. Qu, S. Hariri, and M. Yousif, “An Efficient Network Intrusion Detection Method Based on Information Theory and Genetic Algorithm ”, in Proc. of the 24th IEEE International Performance Computing and Communications Conference (IPCCC 2005), Phoenix, AZ, USA, 2005,pp.11–17.
[12] D. Karaboga and C. Ozturk, “A novel clustering approach: Artificial Bee Colony (ABC) algorithm”, Applied Soft Computing, Elsevier Science Publishers B. V. Amsterdam, The Netherlands, Vol. 11(1), pp. 652–657, Jan. 2011.
[13] X. R. Yang, J. Y. Shen and R. Wang, “Artificial immune theory based network intrusion detection system and the algorithms design”, in Proc. of 2002 International Conference on Machine Learning and Cybernetics, Beijing, China, 2002, pp.73–77.
[14] M. A. Hall and G. Holmes. “Benchmarking Attribute Selection Techniques for Discrete Class Data Mining ”, IEEE Transactions on knowledge and data engineering, vol. 15, no. 6, 2003, pp.1437–1447.
[15] J. Wang ,X. Hong and R. Ren, T. Li, “A real-time intrusion detection system based on PSO-SVM”, in Proc. of the International Workshop on Information Security and Application 2009 (IWISA 2009), Qingdao, China, 2009,pp. 319–321
[16] H. G. Jung, P. J. Yoon and J. Kim, “Genetic algorithm-based optimization of SVM-based pedestrian classifier”, In The 22nd international technical conference on circuits/systems, computers and communications(ITC-CSCC2007), Busan, Korea, July 2007, pp. 783–784.
[17] A. Elngar, D. El A. Mohamed and F. M. Ghaleb, “A Real-Time Anomaly Network Intrusion Detection System with High Accuracy “, Information Sciences Letters, Vol. 2, No. 2, pp.49–56, May 2013.
[18] Ali, Mohammed Mahmood, Khaja Moizuddin Mohammed, and Lakshmi Rajamani. "Framework for surveillance of instant messages in instant messengers and social neworking sites using data mining and ontology." InStudents' Technology Symposium (TechSym), 2014 IEEE, pp. 297-302. IEEE, 2014.
[19] Hosseinkhani, Javad, Mohammad Koochakzaei, Solmaz Keikhaee, and Javid Hosseinkhani Naniz. "Detecting suspicion information on the Web using crime data mining techniques."International Journal of Advanced Computer Science and Information Technology 3, no. 1 (2014): 32-41.
[20] Alami, Salim, and Omar EL Beqqali. "Detecting Suspicious Profiles Using Text Analysis Within Social Media. "Journal of Theoretical & Applied Information Technology 73, no. 3, 2015.
[21] Jain, Suchita, Vanya Sharma, and Rishabh Kaushal. "Towards automated real-time detection of misinformation on Twitter." InAdvances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on, pp. 2015-2020. IEEE, 2016.
[22] Murugesan, M. Suruthi, R. Pavitha Devi, S. Deepthi, V. Sri Lavanya, and Annie Princy. "Automated Monitoring Suspicious Discussions on Online Forums Using Data Mining Statistical Corpus Based Approach."Imperial Journal of Interdisciplinary Research 2, no. 5 2016.

[23] Harsh Arora, Govind Upadhyay. "A Framework for the Detection of Suspicious Discussion on Online Forums using Integrated approach of Support Vector Machine and Particle Swarm Optimization ". Volume 8, No. 5, May – June 2017, ISSN No. 0976-5697
[24] H. Ahmed, J. Glasgow, "Swarm Intelligence: Concepts, Models and Applications", 2012.
[25] Oren La'adan, Amnon Barak, "Inter Process Communication Optimization In A Scalable Computing Cluster", Annual Review of Scalable Computing. September 2000, 121-173.
[26] Basari S.A, Hussin B, Ananta G., et.al. "Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization", doi: 10.1016/j.proeng.2013.02.059.