Survey on Feature Selection Techniques for Intrusion Detection System

Krisha Khandhar[1] , Shreya Bachhav[2],Shrutika Badgujar[3]Neha Bagul[4], Tejaswini Pawar[5]

[1,2,3,4]Student, Department of Information Technology
[5] Assistant Professor,Department of Information Technology
MVPs KBTCOE, Savitribai Phule Pune University

Abstract—This paper presents the overview of Intrusion Detection System(IDS) and Collaborative Intrusion Detection System(CIDS). An IDS is a system that monitors network traffic for suspicious activity and also detects attack that occur in network. CIDS consists in a set of cooperating IDSs which use collective knowledge and experience to achieve improved intrusion detection accuracy. It has been developed to improve the detection capability of single IDS. All this is achieved using the Machine Learning phases. One of the important phase is of feature selection for which the optimization algorithm proves to be the efficient and accurate technique. Among the several optimization algorithms, the best accuracy and detection rate is given by combination of Ant Colony(ACO) and Cuttle Fish(CFA) optimization algorithms.

Keywords—Intrusion Detection, Optimization Algorithm, CIDS, ACO-CFA

1. Introduction

As technology evolves, cyber criminals improve their attacks through different techniques, tools and methods. These cyber attacks can disrupt the network of organization. Hence, avoiding these cyber attacks is the need of the hour. As all the data is stored on the network, its security cannot be compromised with. Intrusion Detection System(IDS) is major network security tool which helps to identify unusual and unauthorized access to secure network. Monitoring and examining the activities taking place on a computer system or network for indications of invasions is the process of intrusion detection. Intrusion detection systems (IDSs) are deployed to protect the computer infrastructures. Critical challenge for designing of this intrusion detection system is malware(malicious software) as they have become more sophisticated. IDS's foremost challenge is to identify obfuscated malware and hence we need to have a robust Intrusion detection system for the growing attacks. IDS is a kind of 'burglar alarm' for computer security. It can easily analyze the security problems and threats as it automates the process of monitoring these threats. In recent years, network attacks have become more frequent and severe, and IDS has become a crucial component of security infrastructure. An IDS gathers information from several computer or network sources, including system commands, system logs, system accounting, security logs, and network logs. The system administrator is notified of the intrusion once the data is analysed to look for any security violations. IDS can be divided into two categories: host-based and network-based. Network-based IDS are placed over the networks and monitors the

malicious activities of those networks. While Host-based IDS scans the host devices. Due to large dependency of networked computers, their data increases. And to maintain their security, there is need to make the existing IDS infrastructure more robust as well as resilient. Collaborative Intrusion Detection System is one approach (CIDS). It is for distributed system intrusion detection that is accurate and effective. Collaborative IDSs (CIDSs) have emerged since traditional IDSs are neither scalable to large enterprise networks and beyond, nor to attacks that are conducted in massively parallel. They are made up of a number of monitoring components for distributed network monitoring. The network, kernel, and application layers are where CIDS employs a number of specialised detectors. In essence, CIDS combines the alarms from these detectors to produce a single intruder alarm. Compared to separate detectors, this improves detection accuracy without noticeably degrading performance. One of the important advantage of CIDS over IDS is it's increased accuracy of detection. In order to help those detectors detect the attack, optimization algorithm is used which helps faster detection of attack and graph based detection is shown to detect the attack . Machine learning steps are performed for the same from Normalization , feature selection to classification and the attack is detected with the highest accuracy and lowest time stamp. This attack name is then forwarded to the other IDS and a combined alarm is send to the distributed network. All of the above mentioned Machine Learning steps is the Feature Selection. Feature Selection is the main and the most important step in detection of attack. The primary and most crucial phase in attack detection is feature selection. The process of trimming the dataset and removing the noisy features is known as feature selection. Filter, wrapper, and hybrid feature selection methods are categories of it. It generally extracts important features from dataset. The technique which can make this feature selection accurate is optimization technique. This paper presents the survey related to the feature selection and optimization techniques.

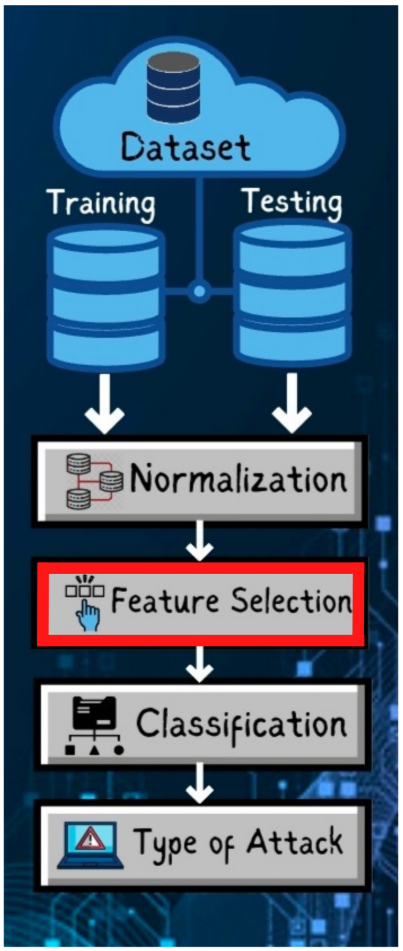

Figure 1. Machine Learning phases

2. Related Work

This section presents the research done on the feature selection method to detect the attacks and different optimization algorithms required for feature extraction.

2.1. Feature Selection

Feature Extraction is required for selecting the important features from the dataset which further helps in detection of attack. It eliminates the dataset's noisy features. Filter, wrapper, and hybrid feature extraction techniques are categories of it. It generally reduces the dimensions of dataset by giving the relevant features for the task to perform. These feature extracted are efficient and relevant features from all other features present. For extracting the features, it makes use of metrics like attribute ratio, separation, correlation coefficient, similarity, etc. Different optimization techniques can help in this feature extraction. There are numerous optimization algorithms available, including ACO, SSO, Bee, Cuttlefish (CFA), PSO, genetic, artificial bee colony, bat, and many more. which helps us in extracting more accurate features in less period of time. Ant Colony Optimization(ACO) and Cuttle Fish Algorithm(CFA) are the optimization algorithms that can do feature selection with better accuracy . The detection rate and accuracy are improved using ACO and decision trees. Features are given a weight, and then ACO is used to choose which features to use. To choose the pertinent features, CFA is utilised, and accuracy has increased.The proposed system uses. The combination of ACO and CFA algorithms to carry out feature selection.

2.2. Feature Selection Methods

2.2.1 Filter Method

Filter method removes the variables which are assigned a value with appropriate ranking criteria and having value below certain threshold. In simple terms, the filter method of feature selection uses ranking techniques. Filter methods are computationally cheaper. The filter-based method provides more

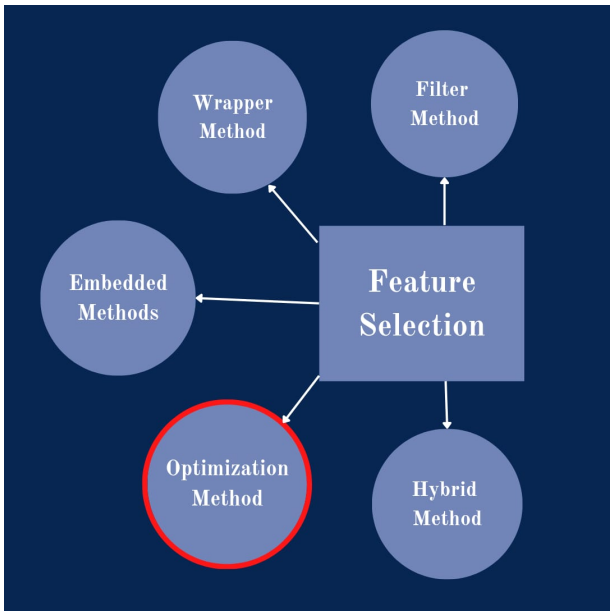

Figure 2. Feature Selection Methods

generality because it is independent of the supervised learning algorithm. They avoid over fitting of the model. But this method ignore dependencies between the features in the dataset and hence, the selected subset might not be optimal. Even there are chances of getting redundant results obtained. This is one of the advantage of the filter method and hence not preferred for bigger and complex datasets or problems.

2.2.2 Wrapper Method

It is also known as brute force method.This method is a heuristic type of learning algorithm.Apart from selecting the relevant features,it selects the optimal features from dataset. It uses the technique of backward elimination to remove insignificant and irrelevant features from all the other features. It creates the different combination of the features evaluates and the compares with the other combination by considering the set of features as a search problem.In order to avoid over fitting it

uses cross validation technique.This method gives more accurate result thought it is computationally expensive and more time consuming.It also maintains the dependencies between features.

2.2.3 Embedded Method

It also goes by the name ”hybrid model of feature selection” because it includes the benefits of both the filter and wrapper methods. The technique uses the learning process of the supervised learning algorithm to carry out feature selection. The embedded approach is divided into three categories: regularisation methods, built-in mechanisms, and pruning methods. In the pruning approach, all of the features are first processed during the training phase of the classification model's construction, and then the features with lower correlation coefficient values are iteratively deleted using a support vector machine. The C4.5 and ID3 supervised learning algorithms' training phases are used in part to pick features in the built-in mechanism-based feature selection technique. The fitting error is minimised using objective functions in the regularisation approach, and features with close to zero regression coefficients are removed.

3. Optimization Techniques for Feature Selection

Feature selection involves selection of best features. This task needs to be done correctly as the further result or classification and final outcome is based on it. Optimization technique or algorithm proves to be the better and accurate feature selection technique, especially when it comes to Intrusion Detection System. The computational capacity of the optimization algorithm is high and gives results in less time as well. Among different optimization algorithms available, Ant colony and Cuttle fish algorithms seems to do the feature selection task for IDS with more accuracy and has higher detection rate. Combination of ACO and CFA yields better outcomes. Their working proves to be effective for selecting the best features from given dataset. This in turn can detect accurate attacks.

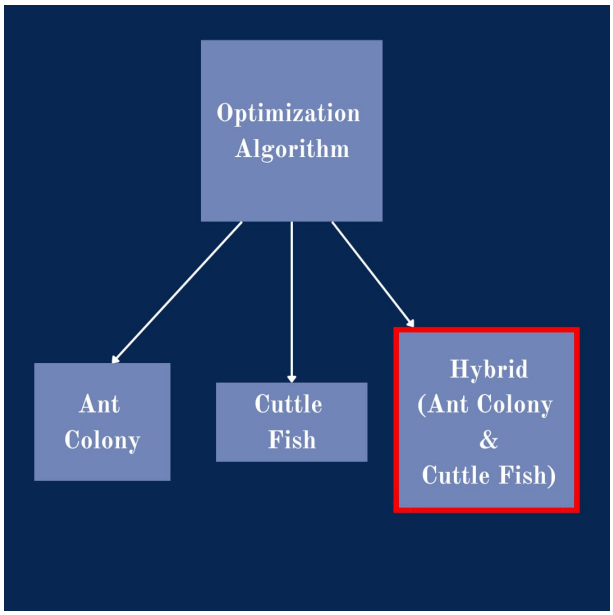

Figure 3. Optimization Techniques

3.1. Ant Colony Optimization

Ant Colony Optimisation algorithm is an example of swarm intelligence technique and it is a metaheuristic algorithm. This algorithm is a multi-agent system that draws its inspiration from ant behaviour. Ants are eusocial insects that prioritise collective survival over individual species survival. Because of this community survival of ants, this algorithm proves to solve optimization problems with good computational capacity and better accuracy. In essence, these ants locate the shortest route by dispersing a chemical compound called a pheromone. Pheromones are nothing but the organic

chemical compounds secreted by the ants. The ants scurry towards the direction of the strongest pheromone concentration. Over time, the shortest route deteriorates and. The shortest path will have a higher traversal rate. ACO's fundamental tenet is to follow the movement of the ants as they leave their nests in search of food, taking the quickest route possible. The general idea behind the ant colony optimization technique is this. The basic objective of the ACO algorithm is to represent a problem as the pursuit of the least expensive path through a graph.In this case, nodes can be thought of as features, and the edges between them signify the selection of the following feature. A minimal number of nodes, or features, must be visited in order to satisfy the traversal stopping requirement in order for the search for the best feature subset to be an ant traversal over the graph. Any feature can be chosen as the next option because nodes are fully connected. The transition rules and pheromone update rules of conventional ACO algorithms can be utilised on the foundation of this reformulation of the graph representation. Heuristic value and pheromone are not linked in this instance.

3.2. Cuttle Fish Optimization

Cuttle fish optimization algorithm is also a meta-heuritic algorithm. It is based on the colour changing behavior of the cuttlefish. So, using this colour changing mechanism it, CFA can solve many global optimization problems. This algorithm can be used as a search strategy for finding the optimal features from all the present features. These colour variations are typically visible in cuttlefish due to light reflection from various cell layers, including chromatophores, leucophores, and iridophores. Two main processes involved in CFA are reflection and visibility. The cuttlefish attempts to mimic the patterns that occur in its environment by using reflection to emulate the mechanism of incoming light reflection in the aforementioned layers, and visibility is recommended to symbolise the matching pattern clarity. The feature subset generated via CFA has a decreased false alarm rate, higher detection rate, and accuracy rate.

Sr. No.	Algorithm	Detection Rate(%)	Accuracy(%)
1	ABC-17	-	93.25
2	Artificial Bee Colony	95.75	92.59
3	PSO	74.43	98.12
4	Hybrid Kernel PCA	-	87.00
5	CFA	92.05	92.80
6	ACO-CFA	98.01	98.55
7	ACO	-	97.80
8	Bat	97.40	97.20

3.3. Hybrid ACO-CFA

The table above shows that hybridization of ACO and CFA proves to be the better technique for intrusion detection system as it can perform the feature selection with higher accuracy and detection rate rather than using it individually.When these two optimization algorithms combined together yields relevant and optimal feature which could detect the further attack correctly in the network.

4. Conclusion

This paper summarizes and gives overview of the concept of IDS and CIDS. Basically, CIDS is more effective as it combines the intrusion detection capacity of individual IDS and therefore increases its capability in terms of detection. To enhance the working of CIDS we need some strong mechanism for attack detection and sending the alarms to other networks or systems in the distributed network. Hence, Machine Learning can be the option, in fact the better option for detection of the type of attacks. For this we need to carry out certain Machine learning phases including normalization, feature selec-

tion and classification and the result of all this will the attack detection. Among all the mention steps Feature Selection is the very important step and the paper gives the survey of different techniques for feature selection. So we further came at a conclusion that the optimization algorithms can be the great alternative for feature selection. For that we need the optimization algorithm with greatest accuracy. Hence, we have gone through and compared many optimization techniques. After studying different optimization algorithms, we have found out that the ACO and CFA gives better accuracy with higher detection rate.

Acknowledgments

We are extremely grateful to Dr.S.R.Devane, Principal, MVPS's Karmaveer Adv.Baburao Ganpatrao Thakare College Of Engineering and Dr.V.R.sonanwane HOD, Head of Department, Department of Information Technology, for their indispensable support, suggestions. We would like to thank our mentor, Ms. Tejaswini S. Pawar, Information Technology, for her insightful advice and direction for the project's preparation. If we do not express our gratitude to the writers of the references and other works consulted for this project, we will be in breach of our obligation. We want to thank them all again from the bottom of our hearts.

5. References

[1] Mehdi Hosseinzadeh Aghdam, Peyman Kabiri, et al. Feature selection for intrusion detection system using ant colony optimization. Int. J. Netw. Secur., 18(3):420–432, 2016.

[2] VR Balasaraswathi and M Sugumaran. A hybrid algorithm using ant colony optimisation and cuttle fish algorithm for feature selection of intrusion detection.

[3] L Dhanabal and SP Shantharajah. A study on nsl-kdd dataset for intrusion detection system based on classification algorithms. International journal of advanced research in computer and communication engineering, 4(6):446–452, 2015.

[4] Sheren Sadiq Hasan and Adel Sabry Eesa. Optimization algorithms for intrusion detection system: A review. 2020.

[5] Shailendra Sahu and Babu M Mehtre. Network intrusion detection system using j48 decision tree. In 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pages 2023– 2026. IEEE, 2015.

[6] Chibuzor John Ugochukwu, EO Bennett, and P Harcourt. An intrusion detection system using machine learning algorithm. LAP LAMBERT Academic Publishing, 2019.

[7] Emmanouil Vasilomanolakis, Shankar Karuppayah, Max M¨uhlh¨auser, and Mathias Fischer. Taxonomy and survey of collaborative intrusion detection. ACM Computing Surveys (CSUR), 47(4):1–33, 2015.

[8] Sharmila Kishor Wagh, Vinod K Pachghare, and Satish R Kolhe. Survey on intrusion detection system using machine learning techniques. International Journal of Computer Applications, 78(16), 2013.

[9] Yu-Sung Wu, Bingrui Foo, Yongguo Mei, and Saurabh Bagchi. Collaborative intrusion detection system (cids): a framework for accurate and efficient ids. In 19th Annual Computer Security Applications Conference, 2003. Proceedings., pages 234–244. IEEE, 2003.