Eurasian Journal of Analytical Chemistry
ISSN: 1306-3057OPEN ACCESS 2018 13(6):684-688

An Optimized Decision Tree Approach for Intrusion Detection

G.Parimala, M.Jayanthi, M.Sangeetha

Received 14 June 2017 ▪ Revised 23 August 2017 ▪ Accepted 24 September 2017

Abstract: *Nowadays, with rapid development in networking infrastructures and with an increase in Internet usage, network security has become an important issue for discussion. Some major challenges with regard to network security are DOS attack, Botnets etc., and sometimes vulnerabilities in network design can also serve as intrusion points for intruders. Therefore, this paper focuses and ensures on optimum network security by setting some thresholds on generic based feature selection mechanism in order to block and overcome attacks like DOS, R2L and U2R etc. In order to verify our approach, a broadly known intrusion dataset named NSL-KDD is used. For detecting the attacks in a network efficiently and also to reduce the false alarm rate, we optimize the decision trees by using Ant Colony Optimization (ACO) algorithm. In order to reduce the data set size we have used ACO algorithm for feature selection. This would provide a more efficient and reduced version of a decision tree and it will also help to identify the exact attack categories. Thus, this approach will prove to be quite an efficient way to identify intrusions in a network for the detection of any abnormal activity on the network. Thus, the proposed system will (1) immediately block an intruder if any of the threshold values set are exceeded. (2) it will list the exact type of attack used by an intruder to get access to the network (3) it also ensures optimum network security.*

Keywords: *ACO, Decision Tree, NSL-KDD.*

INTRODUCTION

Network security is compromised when an intrusion takes place. An Intrusion Detection System (IDS) is used to identify any set of actions or malicious activities which can compromise network security policy [2].There are three types of attacks such as scanning attacks, denial of service attacks and penetration attacks.

Scanning Attacks: It refers to gathering information about a system on the network. Using scanning attacks the attacker can get the information about the topology of the network, server information, version of the softwares and the softwares running on the server. The attacker can also get other information such as the kernel and the Operating System (OS) used on the hosts system. Using these information the attacker can prepare for more specific exploit attacks. *Denial of Service Attack :* In this type the attacker sends excessive messages to the server to authenticate requests that have invalid return addresses. In this way the attacker continuously sends messages with invalid return addresses and hence the server is kept busy thereby keeping the network busy. It is very easy to initiate DoS attacks it just requires a few commands to be run for example using a ping command which will send a number of repeated requests to the host server and if the attackers bandwidth is greater than the victims bandwidth then the victim can easily be hacked. *Penetration Attack :*In this type of attack the attacker can get access to all the system resources, administrator privileges or data. This attack may be a result of a software flaw, bug, loop holes in server or system level architecture[1].

G.Parimala, M.Jayanthi, M.Sangeetha, Assistant Professor,Department of Information Technology,SRM Institute of Science and Technology, Kattankulathur,TamilNadu,India. parimala.g@ktr.srmuniv.ac.in, jayanthi.m@ktr.srmuniv.ac.in, sangeetha.km@ktr.srmuniv.ac.in

NSL-KDD is a benchmarked and well known dataset as there are a limited a number of data sets available for testing efficiency of network systems. The latter KDD dataset had its own issues, thus NSL-KDD data set was proposed. The main advantage of this data set is that it has a reasonable number of records including test and train sets as well, thus it allows to run experiments on the entire data set instead of randomly selecting the parts and running experiments [2].

Ant colony optimization (ACO) is one of the most recent techniques for approximate optimization. The inspiring source of ACO algorithms are realant colonies. The ACO algorithm is mainly inspired by the ants foraging behavior [3]. Ant colony optimization (ACO) is a population-based meta heuristic that can be used to find approximate solutions to difficultoptimizationproblems.ACO Concept : (1) Ants navigate from nest to food source (2) Each ant moves at random and leaves behind a pheromonetrail (3) Shortest path is discovered via pheromone trails (4) More pheromone on path increases probability of path being. Thus ACO algorithm provides good solutions for a given optimization problem. In recent times ACO is used vastly in the Artificial Intelligence (AI) mainly in the swarm intelligence field with some meta-heuristic optimizations.

Decision tree comes under supervised learning techniques. A decision tree is a powerful form of multi variable analysis. They serve as the best substitute for traditional statistical form of analysis, in a variety of data mining tools and techniques for example neural networks, multidimensional forms of reporting and analysis used in the field of business intelligence. Decision trees find various ways of splitting a data set into many different branch like segments starting with the root node. It further on proceeds from the root node based on decision rules which maybe association rules, fuzzy logic or any other user defined set of rules which are used for decision making which finally gives the user a clear solution. Nowadays decision trees are used for predictive modeling machine learning the modern name given to it is Classification And Regression Trees (CART). One such method that is Ant Colony Optimized Decision Tree (ATM) approach is discussed on in this paper.

The Ant Tree Miner algorithm is again a combination of ACO and Decision Trees. In ATM algorithm each attribute is considered as an ant and a graph is being built with initial weight of the edges being zero. Thus, a number of sub-tress are constructed and for each iteration the heuristic function is updated and based on it a global optimal tree is constructed. Thus, the ATM algorithm performs better than the general ID3 or the standard Decision tree algorithm and better optimized results can be achieved. Thus, this paper focuses on providing an optimized decision tree using ACO algorithm in order to reduce the tree size and provide efficient results. The rest of the paper is organized as follows: Section 2 explain in detail about the related works. Section 3 explains the proposed model. Section 4 contains the results. Section 5 explains the conclusion.

LITERATURE SURVEY

There are many works have been proposed by many researchers in this direction in the past. Among them, in [3] NSL-KDD data set is analyzed and used to study the effectiveness of the various classification algorithms (available in the WEKA tool) in detecting the anomalies in the network traffic patterns. In [4] ACO algorithm is used to construct decision trees. Main aim is to provide the rationale for using swarm intelligence (i.e., ACO) in the process of constructing decision trees. The Ant Tree Miner (ATM) algorithm, is capable of giving good results obtained without heuristics. In [5] efficient HIDS Correlation based Partial Decision Tree Algorithm (CPDT) is implemented. The CPDT combines Correlation feature selection for selecting features and Partial Decision Tree (PART) for classifying the normal and the abnormal packets.

In [6] the attributes of the NSL-KDD dataset is divided into categories (i.e) numerical and categorical data and then a rule- based model is designed using Classification and Regression Tree (CART) for generating rules to identify intrusions. In [7] a machine learning approach known as Genetic algorithm is proposed to identify intrusions in network, here the author also solves the fidelity problem. Fidelity in IDS is nothing but the misinterpreting or the missing out the important events in the results obtained. In [8] Feature reduction or otherwise known as dimensionality reduction is performed on the NSL-KDD dataset using PCA and Kernel PCA methods. The Kernel PCA method for dimensionality reduction is more accurate and efficient than the PCA method is clearly observed from the results. After this preprocessing step the reduced dataset is classified using k nearest neighbor (K-NN) to check whether the samples obtained are normal or anomalous network connections. In [9], a fuzzy logic based decision tree approach is pro- posed. Decision trees do not provide sharp decision boundaries which may not be implicated to all knowledge inference systems. A fuzzy decision tree approach overcomes this draw- back of decision tree

without disturbing the attribute values. Fuzzy SLIQ based decision tree algorithm is used to construct decision rules.

PROPOSED SYSTEM

The system proposed in this paper is researched upon to reduce the increased dimensions and memory requirements of the decision tree and to greatly reduce the false alarm rate of a detected intrusion. The first step to this approach is the reduction in dimensionality of the NSL-KDD dataset using the Principal Component Analysis(PCA).

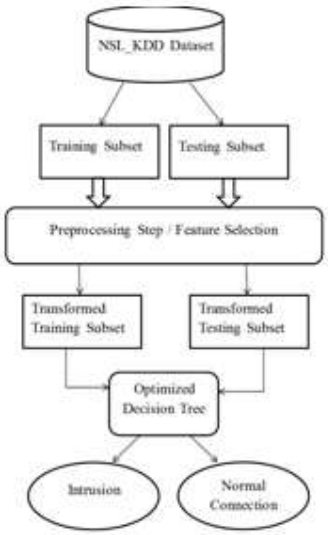

Figure 1: Proposed IDS

The preprocessing that is the result of ACO which generates an optimized data set which serves as in input to the Ant Tree Miner Algorithm (ATM). The biggest advantage of having an optimized data set for input is enabling the system for an optimal feature selection and faster processing in order to detect an intrusion. The second step in the system is traversal of the data set in the decision tree using ACO algorithm ACO (Ant Colony Optimization) algorithm. ACO algorithm makes the best possible traversal of the decision tree to give exact approximations and optimized results and is not prone to failure. The proposed system is trained with training data sets that learn the activity of all nodes in the network and assign threshold values to each node. When the system detects an intrusion in a test data set by exceeding threshold values, it immediately blocks the intruder. In addition to that, it will also list the exact type of attack used by an intruder to get access in to the network.

RESULT AND DISCUSSION

Before applying ACO for feature selection we apply Principal Component Analysis (PCA) to emphasize variation and bring out patterns in a dataset then, we plot the PCA graph which becomes constant after a certain level denoting that the variables after a certain point are redundant.

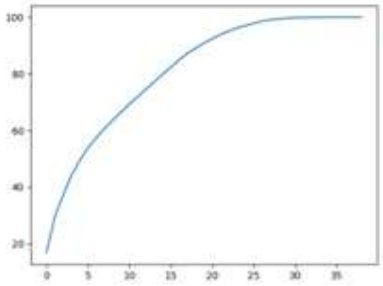

Figure 2: Principle Component Analysis (PCA)

We then use the original NSL-KDD dataset and apply ACO and based on the output of the ACO algorithm we select 21 features and then apply Ant Tree Miner (ATM) algorithm to classify that a

particular connection is an anomaly or not and the accuracy we obtained using 21 features was similar to that of using the entire dataset which contains 42 features.

Figure 3: Accuracy Using 42 Features

Figure 4: Accuracy using 21 features

Table 1 shows the classification accuracy when consider the 41 features (full features) and the selected features which are selected by the ACO algorithm. Totally 10 experiments have been conducted for training and testing datasets. Five experiments were conducted separately for the training and testing datasets.

Table 1: Classification Accuracy

NSL-KDD Dataset	Accuracy Using 41 Features	Accuracy Using 21 Features
NSL-KDD Train (Avg. Of 5 Experiments)	99.71%	99.73%
NSL-KDD Test (Avg. Of 5 Experiments)	99.25%	99.6%

From table 1, it can be seen that the overall performance of the proposed model with selected features is achieved better performance than the proposed model with full features when used the training and testing datasets in each five experiments. This is due to the fact that the use of useful and important features for decision making over the attacks.

CONCLUSION

In conclusion, the system that is proposed overcomes the disadvantages of the previously existing systems of Intrusion Detection System by using a preprocessed and optimized data set that requires lesser memory in order to obtain a better feature selection and reduced false alarm rate. Again the same accuracy is obtained by using 21 features instead of using all the 42 features provided in the original NSL-

KDD dataset which ultimately improve the IDS's performance to a great extent. Thus, the expected results are achieved and the objective of the research is met in the way it is proposed.

REFERENCES

[1] P. D. Verma, r. D. Dangar, r. R. Dangar, b. N. Suhagia (2012) A Pharmacognostical Study on Stem of Capparis decidua Edgew. . International Journal of Pharmacy Research & Technology, 2 (2), 28-32.

[2] Blum, Christian. "Ant colony optimization: Introduction and recent trends." Physics of Life reviews 2.4 (2005): 353-373.

[3] G. V. Subbarao , b. Rajkapoor. (2012) Studies in Formulation Development of Itraconazole Granules Using HPMC E 5 and HPC . International Journal of Pharmacy Research & Technology, 2 (2), 33-36.

[4] Kozak, J., &Boryczka, U. (2016). Collective data mining in the ant colony decision tree approach. Information Sciences, 372, 126-147.

[5] Catherine, F.L., Pathak, R., &Vaidehi, V.(2014, April). Efficient host based intrusion detection system using Partial Decision Tree and Correlation feature selection algorithm. In Recent Trends in Information Technology (ICRTIT), 2014 International Conference on (pp. 1-6).IEEE.

[6] Ji, S. Y., Jeong, B. K., Choi, S., &Jeong, D. H. (2016). A multi-level intrusion detection method for abnormal network behaviors. Journal of Network and Computer Applications, 62, 9-17.

[7] Sharma, V., &Nema, A. (2013, April). Innovative Genetic Approach for Intrusion Detection by Using Decision Tree. In Communication Systems and Network Technologies (CSNT), 2013 International Conference on (pp. 418-422). IEEE

[8] Elkhadir, Z., Chougdali, K., &Benattou, M. (2016). Intrusion Detection System Using PCA and Kernel PCA Methods. In Proceedings of the Mediterranean Conference on Information & Communication Technologies 2015 (pp. 489-497). Springer International Publishing.

[9] Kamadi, V. V., Allam, A. R., &Thummala, S. M. (2016). Acomputational intelligence technique for the effective diagnosis of diabetic patientsusing principal component analysis (PCA) and modified fuzzy SLIQ decision tree approach. Applied Soft Computing, 49, 137-145.