Grenze International Journal of Engineering and Technology, Jan Issue

An IDS in Cloud Environment by using Feature Selection based Hybrid Multilevel Classifier

Partha Ghosh[1], Prashant Kumar[2], Soumyadip Paul[3] and Santanu Phadikar[4]
[1-3]Department of Information Technology, Netaji Subhash Engineering College, MAKAUT, Kolkata, India
Email: partha.ghosh@nsec.ac.in, prashantkr001@gmail.com, sdpaul2000@gmail.com
[4]Department of Computer Science and Engineering, MAKAUT, Kolkata, India
Email: sphadikar@yahoo.com

***Abstract*—Cloud Computing has made resources and services accessible from any place of the world at any time. But with increasing popularity of Cloud Computing, it has also become vulnerable to intrusions. To defend the Cloud from attacks, Intrusion Detection Systems (IDSs) are developed. IDSs are deployed in the system or network to identify intrusions and report about them to the administrative system. Here the authors have proposed a novel IDS to provide security in the Cloud. For Feature Selection (FS), a Mutual Information-Information Gain based modified Ant Colony Optimization (MI-IG based modified ACO) method has been used which resorts to a feature scoring method. After FS, a hybrid multilevel Logistic Regression-Neural Network (LR-NN) classifier has been used for classification purpose where LR executes anomaly detection through binary classification and NN performs misuse detection. NSL_KDD dataset has been selected for conducting the experiments. From the experiments, it is observed that the proposed model has produced results with a high rate of accuracy. The results prove the excellence of the constructed IDS in Cloud Environment.**

***Index Terms*— Cloud Computing, Intrusion Detection System (IDS), Feature Selection (FS), Anomaly Detection, Misuse Detection and Ant Colony Optimization (ACO).**

I. INTRODUCTION

Cloud Computing is the technology that delivers data and services all over the world through Internet as per the demand of the users. With the advancement of Internet technology, it has also become prone to attacks or intrusions. To identify and inform about intrusions, Intrusion Detection Systems (IDSs) are developed. It analyses a network connection to detect whether it is an attack or not and reports to the administration system. There are two types of IDS - Host based Intrusion Detection System (HIDS) and Network based Intrusion Detection System (NIDS). HIDS runs on individual hosts or devices on the network and NIDS is deployed at the strategic points of the network. Two different kinds of approaches are made by the IDS to detect attacks – Anomaly detection and Misuse detection. Anomaly detection approach identifies the attacks which are unknown. In this technique, an IDS has a predefined behavior model of a normal network connection. If any data packet or request behaves differently it is labelled as abnormal and is reported. In case of misuse detection approach, an IDS is equipped with the descriptions of attacks known as signatures which are used to check whether they match with the analysed data and detect intrusion. All the features present in a network do not provide equally useful information. So, to make the IDS efficient under huge computational loads, considering only the

Grenze ID: 01.GIJET.8.1.13_1

informative features or Feature Selection (FS) is required [1]. After FS, classification is done to categorize network connections into attack or normal. Different classifiers follow different methods to make decisions. Constructing and using efficient classifiers that have good confusion resolving and decision making abilities is necessary. In this paper, authors have proposed a method for Feature Selection from the NSL_KDD dataset by applying a Mutual Information-Information Gain based modified Ant Colony Optimization (MI-IG based modified ACO). The IDS model also contains the construction of a hybrid multilevel Logistic Regression-Neural Network (LR-NN) classifier to obtain better classification accuracy where LR performs binary classification based anomaly detection and NN does misuse detection.

II. Related Works

Cloud Service Providers need to implement elasticity in the network to provide the customers with their needed information or services within a reasonable time [2]. But it cannot be maintained always because of fluctuations in the network and insufficiency of resources may arise at times. Malicious software and hackers make this work harder by attacking the network connections to steal or damage data. So, to make cloud more secure by identifying intrusions and reporting them, IDSs are being developed. A network connection contains features about itself. Selecting out only the relevant features from the network is an effective way to reduce computation time as well as errors of an IDS. After FS, classification is important to test the accuracy of the method. Amrita and P. Ahmed in 2013, used Correlation-based Feature Selection (CFS), Consistency-based Feature Selection (CON) and IG to obtain relevant features and tested the reduced feature set on Naïve Bayes and C4.5 classifiers [3]. Nature inspired algorithms are also used in feature selection processes. In 2005, Hai-Hua Gao et al. proposed an IDS using modified ACO for FS and Support Vector Machine (SVM) for classification purpose [4]. In 2017, Md. S. Akhtar et al. presented a two-step Particle Swarm Optimization (PSO) based feature selection method and then applied Maximum Entropy (ME), Conditional Random Field (CRF) and SVM for classification and testing [5]. It is evident from the above-mentioned articles that reduced feature sets containing only the relevant features perform better than the complete feature set, but there is a high percentage of false alarm rates which must be reduced. In 2014, P. Ghosh et al. developed a hybrid multilevel classification model where K-Nearest Neighbor (KNN) for anomaly detection and Neural Network (NN) for misuse detection are applied on the reduced feature set obtained by Rough Set Theory (RST) and IG [6]. After FS, choosing suitable classifiers is needed to produce better detection accuracy. M. Bahrololum et al. in 2009, used Self Organizing Map (SOM) based unsupervised NN for differentiating attacks into different categories and then used Backpropagation based supervised NN for clustering [7]. In 2013 Eduardo de la Hoz et al. proposed a method set up by an ensemble of support vector classifiers and used dimensionality reduction techniques for generating a set of distinguishable features [8]. The classification results can be used to identify new attacks and incorporate the results with future IDSs for their betterment. The authors of this paper are inspired by the above mentioned papers and have tried to develop an efficient IDS.

III. Preliminary Study

The authors have used a meta-heuristic algorithm and concepts of Information Theory in this paper. The basic ideas of these are described below.

In 1996, Marco Dorigo et al. proposed a novel meta-heuristic algorithm named Ant Colony Optimization (ACO) [9]. The algorithm represents the behavior of ants while they search for food by different paths starting from their nest (source). With time all the ants travel by the shortest path between source and destination (food). In the modified ACO algorithm for feature selection, the sources and destinations may be different for each ant, here the nodes chosen in between the source and destination are important. Each feature is considered as a node of a complete graph and some ants are generated which are then randomly put on the nodes. For an ant a currently present on a feature i, the probability of going to feature k is given as:

$$p_a(i,k) = \begin{cases} \dfrac{[\tau_{ik}]^{\alpha}[\eta_{kj}]^{\beta}}{\sum_{k \in V_a(i)}[\tau_{ik}]^{\alpha}[\eta_{kj}]^{\beta}} & \text{if } k \in V_a(i) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

τ_{ik} is the pheromone concentration between feature i and feature k, η_{kj} is the feature score for feature k for decision class j, $V_a(i)$ is the set of nodes that ant a has not visited yet. α and β are two parameters to regulate τ and η. The feature k will be selected by equation given as:

$$k = \begin{cases} argmax_{k \in V_a(i)}\, p_a(i,k) & \text{if } q \le q0 \\ v & \text{otherwise} \end{cases} \quad (2)$$

q is a random number generated between 0 and 1, q_0 is a threshold value, v is any random node from the set $V_a(i)$. The pheromone concentration is updated using the formula:

$$\tau_{ik} = (1 - \rho)\tau_{ik} + \rho\Delta\tau_{ik} \quad (3)$$

ρ is the evaporation rate of pheromone and $\Delta\tau_{ik}$ is given by:

$$\Delta\tau_{ik} = \begin{cases} \frac{Q}{c} & \text{if } \{ik\} \in S \\ 0 & otherwise \end{cases} \quad (4)$$

where S is the set of paths that have been chosen by an ant a, Q and c are two constants used for regulation of pheromone concentration. The process is continued for some fixed iteration or until certain number of features are chosen by an ant a for a particular decision class j.

The concepts of MI and IG are also used in the proposed model for FS. MI is the measure of dependence between two features i.e., quantity of information obtained about a feature i by observing another feature k [10]. MI calculates the difference between the joint distribution of the pair of features (i,k) and the product of the marginal distribution of feature i and feature k. IG is the decrease in entropy of a feature after dividing the records into some splits [11]. Entropy of a split can be described as the uncertainty of possible outcomes. The equation for entropy is as follows:

$$E = -\sum_{l}^{c} p_l \, log_2 p_l \quad (5)$$

where l is a unique value in a split, c is the set of all types of different values of l, p_l is the probability of occurrence of a feature value of type l in a particular split. The Information Gain of a particular feature is calculated by subtracting the sum of entropies of all splits from the entropy of the complete split.

The concepts mentioned above are used in this paper to develop a MI-IG based modified ACO method for FS. After FS, a hybrid multilevel LR-NN classifier is developed to create an efficient IDS.

IV. Proposed Model

In this section, the authors have proposed a method to develop a proficient IDS. A reduced feature set is a basic need of an IDS to analyze network connections in lesser time. At first, a feature subset is generated by executing MI-IG based modified ACO where feature scores are availed as heuristic information. After that, hybrid multilevel LR-NN classifier is deployed to identify attacks. Anomaly detection is performed by LR through binary classification and misuse detection is done by NN. Fig. 1 shows the flow of the proposed model.

The proposed model is subdivided into two parts. In the first phase a FS method to produce a feature subset has been described and in the second phase a hybrid multilevel classifier has been introduced to generate better classification results.

In the first phase, a feature subset has been generated primarily by using a modified ACO algorithm and then further reduced by applying the concept of MI and IG. To get the heuristic information as mentioned in the modified ACO algorithm, a method to calculate feature scores has been executed. Feature score for a feature i corresponding to a particular decision class j is presented as:

$$S(i,j) = \frac{V_{ij}}{N_i} \quad (6)$$

where $V_{ij} = |v_{ij}| - |I_i|$. v_{ij} is the set of unique elements in i that implies to decision class j and I_i is the intersection of all v_{ij}. V_{ij} is the difference between the cardinalities of v_{ij} and I_i, N_i is the number of unique values in feature i.

The algorithm for calculating the feature scores is given below:

Algorithm (1): Feature Score Calculation
Input: KDDTrain+ Dataset
Output: Feature Score of each feature for each decision class
Steps:
for feature i do:
find N_i
for decision class j do:
find set $v_{ij} = \{x \mid x \rightarrow j\}$
generate a set $I_i = \cap\, v_{ij}$
for decision class j do:
find the score $S(i,j)$ for feature i and decision class j by using equation (5)

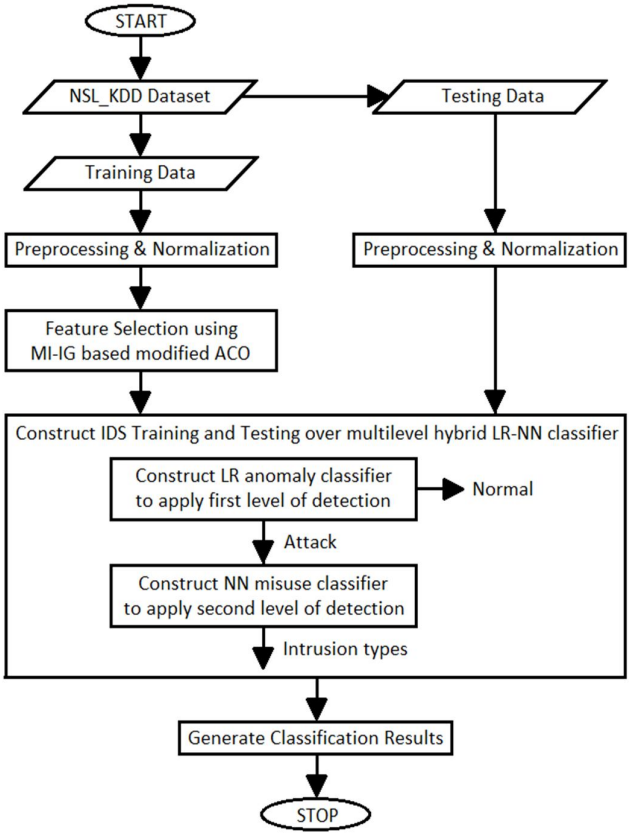

Figure 1. Flowchart of the proposed model

The feature scores are then used as heuristic information in the modified ACO algorithm. For each decision class *j*, a different path P_j is constructed by the modified ACO algorithm. A feature subset *FS1* is generated by the union of all paths and another empty set *FS2* is initiated. The Information Gain (IG) of all features and Mutual Information (MI) between the features present in *FS1* are calculated. The features in *FS1* that have MI values less than a threshold value with all other features are included in *FS2* and excluded from *FS1*. Now, the features in *FS2* having IG value less than another predefined threshold value are deleted from *FS2*. The remaining features in *FS1* are then sorted in decreasing order according to their IG. For each of the remaining feature *i* in *FS1*, a set R_i is incorporated with feature *i*. R_i contains the features that are highly correlated with feature *i*. The first feature *i* of *FS1* is included in *FS2*. For feature *i* and *k* in *FS1*, if R_k excluding *i* is not a proper subset of R_i then newly found features in R_k are included in R_i, *k* is included in *FS2* and excluded from *FS1*. Again, if R_k excluding *i* is a proper subset of R_i, but *k* has IG greater than a threshold value then also *k* is included in *FS2* and excluded from *FS1*. Otherwise, *k* is only excluded from *FS1*. When for a feature *i*, R_i is iterated completely, *i* is excluded from *FS1*. This process is continued until *FS1* is completely exhausted.
The algorithm for the process mentioned above is given below:

Algorithm (2): Optimal Feature Subset Generation
Input: KDDTrain+ Dataset
Output: Reduced feature set containing relevant features only
Steps:
for decision class *j* do:
construct a path P_j by modified ACO algorithm
generate feature subsets $FS1 = \cup P_i$ and $FS2 = \emptyset$
calculate IG of all features and MI between all features in *FS1*
for feature *i* in *FS1* do:
if $MI(i,k) < \text{threshold}_{MI}$; $\forall k \in FS1, k \neq i$:

$FS2 = FS2 \cup \{i\}$
$FS1 = FS1 - \{i\}$
for feature i in $FS2$ do:
if $IG(i) < \text{threshold}_{IG1}$:
$FS2 = FS2 - \{i\}$
sort the remaining features in $FS1$ according to their IG in decreasing order
for feature i in $FS1$ do:
incorporate a set $R_i = \{k \mid MI(i,k) > \text{threshold}_{MI} ; \forall k \in FS1, k \neq i\}$ with i
for feature i in $FS1$ do:
$FS2 = FS2 \cup \{i\}$
for feature k in R_i do:
if $(R_k - \{i\}) \not\subset R_i$:
$R_i = R_i \cup R_k$
$FS2 = FS2 \cup \{k\}$
else if $(R_k - \{i\}) \subset R_i$ and $IG(k) > \text{threshold}_{IG2}$:
$FS2 = FS2 \cup \{k\}$
$FS1 = FS1 - \{k\}$
$FS1 = FS1 - \{i\}$
$FS2$ is the optimal feature subset

After generating optimal feature subset, the records are classified in the second phase. To improve the classification accuracy, a hybrid multilevel LR-NN classifier is constructed. LR is applied at first to categorize incoming and outgoing data packets into attack or normal. LR uses binary classification for anomaly detection i.e., LR has the normal behavioral model of a network connection and initiates an alarm when it is deviated. If a data packet is normal then it reaches its destination without any restrains. Otherwise, an alarm is generated and the data packet is sent to the next module of classification. NN is used for misuse detection where the attacks are sub-categorized into four types - DoS, Probe, R2L and U2R. NN uses a set of signatures which is used to match with the features of a network connection to find out the type of attack. The effectiveness of the proposed IDS can be estimated from the classification results.

V. Result And Analysis

To evaluate the accuracy of the proposed model, NSL_KDD dataset has been chosen. Preprocessing and Normalization is required before using the NSL_KDD dataset. Some of the features of the dataset consist of non-numeric values, preprocessing is done to make them numeric. Feature values are scaled to a range through the process of normalization. Only 19 features out of total 41 features are obtained through MI-IG based modified ACO algorithm. To improve the classification accuracy a hybrid multilevel LR-NN classifier is constructed. From the results it is evident that the reduced feature set demonstrates better performance than the complete feature set and the hybrid multilevel LR-NN classifier has a higher accuracy than LR or NN individually can have. The hybrid multilevel LR-NN classifier is also tested on the feature subset containing 19 features having the highest IG values for comparing with the proposed FS method. Results obtained from the experiments are given below:

TABLE I. SELECTED FEATURES BY THE PROPOSED MODEL AND IG

Method	No. of Features	Features
MI-IG based modified ACO	19	2, 3, 4, 5, 6, 23, 25, 26, 27, 29, 31, 33, 35, 36, 37, 38, 39, 40, 41
IG	19	3, 4, 5, 6, 12, 23, 25, 26, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39

TABLE II. CLASSIFICATION ACCURACY OF DIFFERENT CLASSIFIERS ON DIFFERENT FEATURE SETS

No. of features	Hybrid LR-NN	LR	NN
MI-IG based modified ACO (19)	84.7986 %	76.10894 %	80.61125 %
IG (19)	82.092796 %	74.44109 %	78.57079 %
All (41)	81.61817 %	75.34599 %	77.027 %

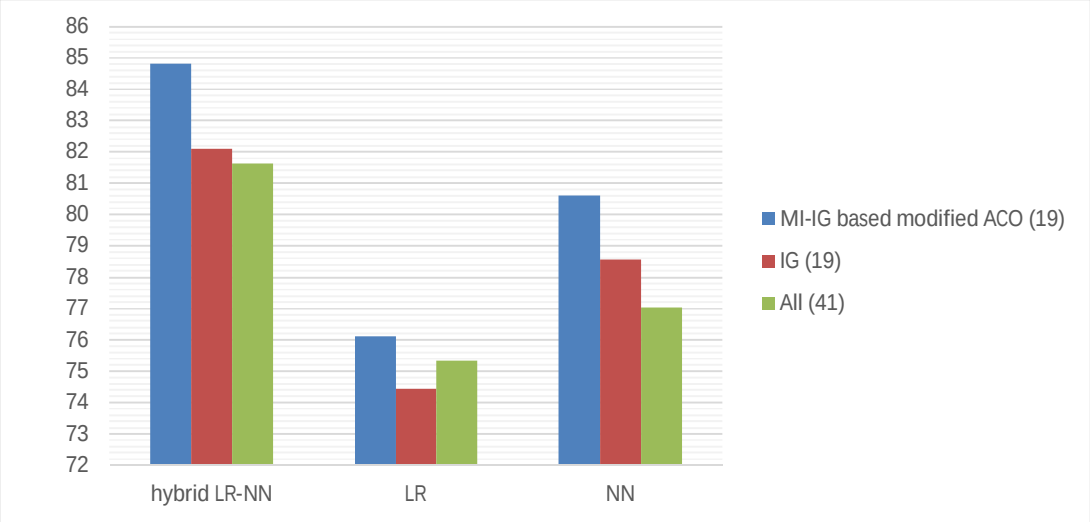

Figure 2. Comparison between Classification Accuracy of Different Classifiers on Different Feature Sets

VI. Conclusions

In this paper, the authors have tried to develop a model for detecting and analysing intrusions with a higher rate of accuracy. The goal is accomplished through MI-IG based modified ACO algorithm and hybrid multilevel LR-NN classifier. 19 features are obtained by the FS method and the hybrid classifier is then applied on it for evaluating the proficiency. The results show the excellence of the proposed model. From the experiments it can be concluded that FS reduces time and space complexity by a significant margin and a competent classifier can increase the detection accuracy considerably.

References

[1] H. J. Liao, C. H. Richard Lin, Y. C. Lin, and K. Y. Tung, "Intrusion detection system: A comprehensive review," *Elsevier, Journal of Network and Computer Applications*, vol. 36, pp. 16–24, 2013, doi: 10.1016/j.jnca.2012.09.004.

[2] S. Sakr, "Cloud-hosted databases: Technologies, challenges and opportunities," *Springer, Cluster Computing*, vol. 17, pp. 487–502, 2014, doi: 10.1007/s10586-013-0290-7.

[3] Amrita and P. Ahmed, "A hybrid-based feature selection approach for IDS," *Springer, Networks and Communications (NetCom2013), Lecture Notes in Electrical Engineering*, vol. 284, pp. 195–211, 2014, doi: 10.1007/978-3-319-03692-2_16.

[4] H.-H. Gao, H.-H. Yang, and X.-Y. Wang, "ANT COLONY OPTIMIZATION BASED NETWORK INTRUSION FEATURE SELECTION AND DETECTION," *IEEE, Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou*, pp. 18–21, 2005.

[5] M. S. Akhtar, D. Gupta, A. Ekbal, and P. Bhattacharyya, "Feature selection and ensemble construction: A two-step method for aspect based sentiment analysis," *Elsevier, Knowledge-Based Systems*, vol. 125, pp. 116–135, 2017, doi: 10.1016/j.knosys.2017.03.020.

[6] P. Ghosh, C. Debnath, D. Metia, and R. Dutta, "An Efficient Hybrid Multilevel Intrusion Detection System in Cloud Environment," *IOSR Journal of Computer Engineering (IOSR-JCE)*, vol. 16, no. 4, pp. 16–26, 2014.

[7] M. Bahrololum, E. Salahi, and M. Khaleghi, "ANOMALY INTRUSION DETECTION DESIGN USING HYBRID OF UNSUPERVISED AND SUPERVISED NEURAL NETWORK," *International Journal of Computer Networks & Communications (IJCNC)*, vol. 1, no. 2, 2009.

[8] E. de La Hoz, A. Ortiz, J. Ortega, and E. de La Hoz, "Network Anomaly Classification by Support Vector Classifiers Ensemble and Non-linear Projection Techniques," *Springer, LNAI 8073*, pp. 103–111, 2013.

[9] M. Dorigo, V. Maniezzo, and A. Colorni, "The Ant System: Optimization by a colony of cooperating agents," *IEEE, IEEE Transactions on Systems, Man, and Cybernetics-Part B*, vol. 26, no. 1, pp. 1–13, 1996.

[10] F. Amiri, M. Rezaei Yousefi, C. Lucas, A. Shakery, and N. Yazdani, "Mutual information-based feature selection for intrusion detection systems," *Elsevier, Journal of Network and Computer Applications*, vol. 34, pp. 1184–1199, 2011, doi: 10.1016/j.jnca.2011.01.002.

[11] H. Uğuz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," *Elsevier, Knowledge-Based Systems*, vol. 24, pp. 1024–1032, 2011, doi: 10.1016/j.knosys.2011.04.014.