Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol.6, No.1, December 2018, pp. 197~200
Available online at:http://pen.ius.edu.ba

Study and Analysis of Intrusion Detection System Using Random Forest and Linear Regression

Sathish Kumar. P[1], Arun Raaza[2]

[1]Assistant Professor, ECE, Vels Institute of Science, Technology & Advanced Studies (VISTAS),
[2]Deputy Director of CARD, ECE, Vels Institute of Science, Technology & Advanced Studies (VISTAS)

Article Info

Article history:

Received Jan 12th, 2018
Revised Apr 20th, 2018
Accepted June 26th, 2018

Keyword:

Centralized IDS
Decentralized
UNSW-NB15

ABSTRACT

The cyber security is the challenging job in present network system. There are number of existing Intrusion Detection Systems are available to overcome the issues, in this paper we proposed the linear regression and random forest technique is used. The latest UNSW-NB15 dataset is used for analyzing the proposed methods. Selecting significant features and removing irrelevant features by using proposed learning methods as well as identifying the best method by evaluating the results obtained.

Corresponding Author:

Sathish Kumar. P,
[1]Assistant Professor, ECE, Vels Institute of Science, Technology & Advanced Studies,

1. Introduction

Nowadays enhancement of network system leads to increase of cyber attacks. In early days, the intrusion detection is done by administrator manually. At present the large in network system, it is very difficult of using traditional method [1]. To avoid the complexity of manual monitoring, the machine leaning algorithms are implemented in Intrusion Detection System. First the supervised learning method is implemented in IDS and it gives the best result of identify the known attacks with low false positive rate, but it is restricted on detecting only known attacks. Then the unsupervised learning method is used to detect the unknown attacks but increase in false positive rate [2]. To overcome the issues, the new method of Hybrid is introduced. Hybrid method is the combination of supervised and unsupervised learning technique and it gives the best result of high detection rate and low false positive rate [1,4] .

2. **Architecture classification**

The architecture of IDS is important factor in every organization. The performance of IDS enhance based on the architecture of the system. It mainly categorized into three: (i) Centralized, (ii) Decentralized and (iii) Distributed.

i) *Centralized* IDS is the multiple systems connected into a central processing unit, in which multiple system behavior is monitored by a central processing unit. Due to the large number of systems connected, the central analysis unit is overloaded with data and Single Point of Failure (SPof).
ii) *Decentralized* means multiple system and multiple processing units are connected using a hierarchical structure. In this structure the main processing unit at the top, in which the

preprocessing of data is done by nearest processing unit and send back to the top of the main processing unit.

iii) *Distributed* structure is designed in basis of Peer to peer (P2P) architecture. In this structure there is no main processing unit, every system has own processing unit and work is assigned to the agents in distributed method.

3. Overview of ids

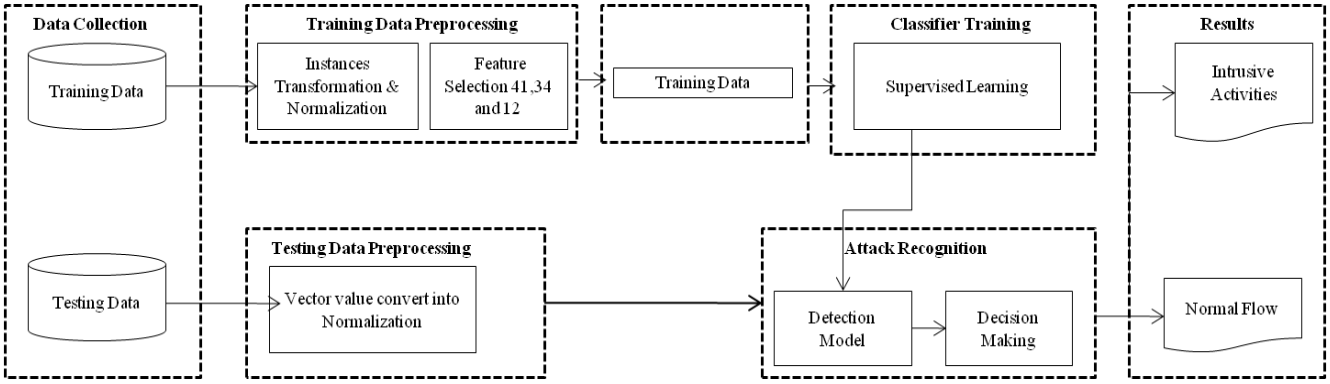

Figure 1. Overview of IDS

3.1 Data Collection

In this phase the data is collected to analyze the best proposed method. UNSW-NB15 dataset is used in our research for identifying the performance of proposed methods. This dataset consists of 49 features and nine attack types.

3.2 Data Preprocessing

The collected data is preprocessed in this phase. It consists of Data Transformation, Data Normalization and Feature Selection.

1) *Data Transformation* is the process of transferring the symbolic data into numerical values. In KDD Cup 99 dataset consists symbolic data such as ftp, http,and so on. This symbolic data is converted into numerical values.
2) *Data Normalization* is the process of converting the irregular values into a normal range [0-1].
3) *Feature Selection* is the method of identifying the significant features and removing the unwanted features in order to increase the speed of the process.

3.3 Classifier Training

The supervised learning classifier method is used in our research. After selecting the relevant features the proposed classifiers are implemented to identify the anomalies. The proposed linear regression and random forest classifiers are used.

3.4 Attack Recognition

In this phase the normal and attack type are identified using trained classifiers and also testing data is processed to identify the attacks.

4. Proposed methodology

i) UNSW-NB15 Dataset

The UNSW-NB15 dataset was latest published dataset in 2015 which consists of 45 features and nine attack types. These features are classified into six set such as Basic Feature, Content Feature, Flow feature, Time

Feature, Additional Generated Feature and Labelled Feature. UNSW-NB15 dataset contain training and testing dataset. The nine attack types of UNSW-NB15 dataset are Dos, Shell code, Exploit, Fuzzer, Worms, Backdoor, Reconnaissance, Generic and Analysis.

ii) Linear Regression Method

Linear Regression is one of algorithm in supervised machine learning. This algorithm is referred as the combination of input variable (x) in order to predict the output variable (y). If there is single input value (x), then the linear model is represented as Simple Linear Regression. If there are multiple input values (x), the model is represented as Multiple Linear Regression. In this model both input and output variable is in numeric data. This method is defined as the relationship between independent variable and dependent variable. Here independent variable is considered as input values and dependent variable is considered as output values.

iii) Random Forest Method

Random Forest is the supervised machine learning algorithm, in which the number of trees created to predicts the results with high accuracy. This method is used in both classification and regression problems. The main advantage in this model is handle the missing values as well as it would not overfit the model.

5. Experiment and analysis

The proposed algorithms Random Forest and Linear Regression are implemented in UNSW-NB15 dataset for selecting the relevant features in order to get accurate detection rate. The results obtained by running the proposed machine learning algorithm using Weka are shown in table 1. In order to find the best proposed method in weka, the attribute evaluator CfsSubsetEvaluator and BestFirst method is selected and get the 11 significant features in Random Forest and Linear regression Method.

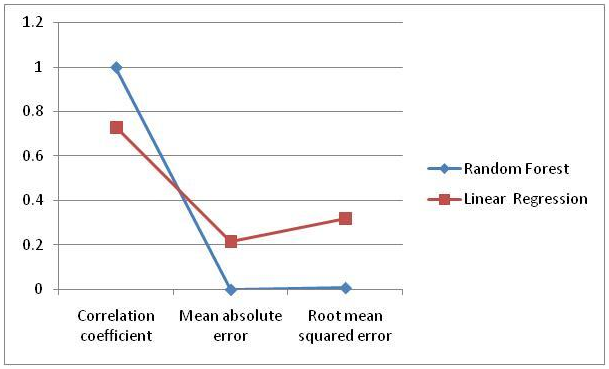

Figure 2: Performance of Random Forest and Linear Regression

	Random Forest	**Linear Regression**
Correlation coefficient	0.9999	0.7311
Mean absolute error	0.0003	0.2144
Root mean squared error	0.0078	0.3181
Relative absolute error	0.067%	49.3157%
Root relative squared error	1.5705%	68.2274%
Time Taken	24.65 sec	0.89 sec

By comparing the metrics of mean absolute, root mean squared, relative absolute, root relative squared as shown in Table 1 and Figure 2, Correlation coefficient is 0.999 and 0.731, Mean absolute error 0.0003 and 0.2144, Root mean squared error is 0.0078 and 0.3181, Relative absolute error 0.067% and 49.3157% Root relative squared error is 1.5705% and 68.2274%. We suggested the Random Forest method is performed better than compared to linear regression method.

6. Conclusion

In this method, the UNSW-NB15 dataset is used to analyze the best supervised learning method by comparing Random Forest and Linear Regression Algorithm. The Significant features are selected by implementing the proposed learning method using the Weka tool. By evaluating the results we conclude that the Random Forest Method is provided the better performance.

Reference

[1]. Antonia Nisioti, Alexios Mylonas, Paul D. Yoo and Vasilios Katos, "From Intrusion Detection to Attacker Attribution: A Comprehensive Survey of Unsupervised Methods" , IEEE Communications Surveys & Tutorials, Vol. 20, No. 4, Fourth Quarter 2018.

[2] An Improved intrusion detection Algorithm Based on GA and SVM PEIYING TAO, ZHE SUN, AND ZHIXIN SUN, IEEE ACCESS Volume 6, 2018, PP 13624 to 13631.

[3]. HAST-IDS Learning Hierarchical Spatial- Temporal Features using Deep Neural Networks to Improve Intrusion Detection IEEE Access 2017 Volume 6, 2018 pp No: 1792 to 1806.

[4] David J. Weller-Fahy, Member, IEEE, Brett J. Borghetti, and Angela A. Sodemann, Member, IEEE," A Survey of Distance and Similarity Measures Used Within Network Intrusion Anomaly Detection", IEEE COMMUNICATION SURVEYS & TUTORIALS, VOL. 17, NO. 1, FIRST QUARTER 2015.

[5] R.Heady, G.Luger, A.Maccabe, M.Servilla. "The architecture of a network level intrusion detection system". Tech.rep., Computer Science Department, University of New Mexico, New Mexico, 1990.

[6] https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/

[7]. G.V.Nadiammai, M.Hemalatha, "Effective approach toward intrusion detection system using data mining techniques",Egyptian Informatics Journal 2014 No.15, pp37-50.

[8] F.Amiri, M.Rezaei Yousefi, C. Lucas, A. Sha kery and N.Yazdani, "Mutual information-based feature selection for intrusion detection systems", J. Netw. Comput. Appl., vol.34, no. 4, pp. 1184-1199, 2011.[9] Zargari S. and Voorhis D., "Feature Selection in the Corrected KDD-dataset", 2012 Third International Conference on Emerging Intelligent Data and Web Technolgies, Bucharest, 2012, pp. 174-180.

[10] Y.Kalpana, S.Purushothaman, S.Rajeswari, "Internation Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 9 NO.27(2014).