International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6S3, September 2019

High-Performance Feature Selection Model for Network Intrusion Detection System

L. Dhanabal, S. P. Shantharajah

Abstract:-Network intrusions detection is a continuous vigilant task and to efficiently analyze the traffic in the corporate network to detect network intrusions. The efficiency of the Network Intrusion Detection System (NIDS) performance can be improved by adopting feature selection or reduction process to suit the present day high speed real time networks. This work is focused on identifying the key features of the audit dataset used to build an efficient light-weight NIDS. The NSL KDD dataset is used in this work titled Attribute Richness Based Feature Selection (ARFS) in order to analyze its performance.The obtained results are compared with the Correlation-based Feature Selection (CFS) and Information Gain (IG) feature selection methods. The proposed feature selection method produced better detection rate comparatively.

I. INTRODUCTION

Networking of computers has led to automation of various processes in the world ranging from ecommerce to national security. Dependency on networked computers for all walks of life has lured the attackers to launch attack on them.Protecting real-time high-speed networked computers gains high priority in the agenda of security specialists. Robust NIDS are required to thwart the intrusions launched by professional intruders. Data sets are used to train and test NIDS. A few of the key features in the benchmark dataset are sufficient to build NIDS models [1]. Inclusion of feature selection / reduction process have resulted in implementation of light-weight and robust NIDS. In this research article the results of the ARFS is analyzed against conventional feature selection methods using the naïve bayes classifier. Section 2 discusses about related work. A discussion on feature selection methods is done in section 3. In the section 4, a discussion is made on classification techniques. An overview on the functionality of the proposed method is given in the section 5. Experimental setup is discussed in the section 6. An elaborate discussion on the result is done in section 7. This work is concluded in section 8 and a discussion on our future research is also done.

II. RELATED WORK

The research communities have contributed extensive work in the area of feature selection in audit data sets. Feature selection process has proved to be inevitable in the preprocessing stage of NIDS model development [2].

[3] Suggests a Markov model and decision tree technique for selection of optimal features to used in developing a Bayesian network and regression tree based IDS.

A novel feature selection method suggested by [4], in which, it involves in the removal of one un-important feature for every instance from the audit dataset by an integrated SVM and neural network method. 34 vital features out of 41 features are selected for the attack detection process.

[5] Suggests an Co-relation based feature selection for making the NIDS to effective in intrusion detection

An hybrid approach is suggested by [6] which involves combination of Bayesian Network (BN) and Classification & Regression Tee (CART) used in an NIDS.

III. FEATURE SELECTION METHODS

Feature selection methods relieve the NIDS from the curse-of-dimensionality issue. Any feature selection method consists of the following generic sequence:

a. Candidate Sub-set generation
b. Subset evaluation function
c. Termination Condition
d. Validation process

Feature selection methods are basically classified as the individual evaluation and subset evaluation.The limitation of the first method is overcome by the second method, which is further classified as filter and wrapper method.

IV. CLASSIFICATION TECHNIQUES

There are many classification techniques in practice to detect anomalous traffic. Naïve Bayes classification method is a probabilistic approach [7]. The probability of the presence of an attribute is totally independent of the rest of the attributes. In a given setof k features / attributes, the naïve bayes classifier attempts 2k! autonomous supposition. The contributions in [8] investigates the various scenarios in which the naïve bayes classifier's manifests best performance.

V. FUNCTIONALITY AND OVERVIEW OF PROPOSED METHOD

In this contribution, the benchmark dataset [9] NSL KDD is used. This dataset has 41 features, out of which 22 features are selected by the proposed method based on the richness of information contained in it. The richness in information is assessed by considering the three vital metrics, namely Accuracy, True Positive Rate (TPR) and False Positive Rate (FPR) of the classification systems. In this approach and sequential search is made on the audit data set to find the set of vital features which improves the detection rate of the classifier.

Manuscript published on 30 September 2019.
***** Correspondence Author (s)
Dr. L. Dhanabal, Kumaraguru College of Technology, Coimbatore, Tamilnadu, India.
Dr. S. P. Shantharajah, School of Information Technology and Engineering, VIT University, Vellore, Tamilnadu, India.

The process begins with all features in the dataset and excludes one feature iteratively from the dataset till the accuracy of classifier doesn't go below the threshold value.If a feature is removed and accuracy deteriorate, then that particular feature is deemed to be important, if accuracy increases, the feature is deemed to be unimportant and if there is no change in the accuracy, that particular feature is of less importance.

The algorithm of the ARFS :

Input :

C = Complete set of 41 features of NSL-KDD dataset

A= Accuracy measure of the classifier

E = RMSE

AVG_TPR= Average of TPR

// Naïve Bayes classifier produces the values for A, E and AVG_TPR when the full dataset is used. This //value is utilized as the threshold value for the feature selection process

Begin

Initialize D = {C}

For each feature {ci } from C

K=D-{ci }

Call Naïve Bayes classifier with the K features

If E >= A ANDRMSE <=E AND A_TPR>AVG_TPR THEN

K=K- {ci }

D=D-{ci }

C=K // The optimal feature set

End

VI. EXPERIMENTAL SETUP

A description on experimental setup and result analysis is done in this section. The algorithm for ARFS is implemented in Java. The conventional feature selection methods such as CFS and IG are also implemented in java. Naïve Bayes classifier with the complete training set and 10-fold cross validation is chosen to validate the performance of the CFS, IG and ARFS. The training dataset is divided into 10 equal sized disjoint sub-sets. One among the sets is used for the testing purpose. The rest of the other nine sets are used for realizing the classification model. The testing process is repeated for equal sized disjoint subset.

VII. RESULTS AND PERFORMANCE EVALUATION

Two standard approaches such as CFS, IG and the proposed ARFS for feature selection was experimented with the NDS KDD dataset with 41 features and their results were obtained. The reduced set features are tabulated in the Table 1.

Table 1: No. of features selected by CFS, IG and ARFS Feature selection

Feature selection technique	No. of attributes selected	Selected attributes
CFS	10	3,4,5,6,12,26,29,30, 37,38
IG	20	3,4,5,6,12,23, 24 ,25,26,29,30, 31 ,32,33,34,35 , 36 ,37,38,39
ARFS	22	1, 3. 4. 5, 6, 10, 12, 13, 14, 17, 18, 22, 23,24,30,33, 34, 35, 36, 37, 39, 40

Confusion matrix shown in Table 2 was used to evaluate the accuracy manifested by the classifier.

Table 2: Confusion Matrix

Actual	Predicted	
	Normal (-ve)	Attack (+ve)
Normal (-ve)	True Negative (TN)	False Negative (FN)
Attack (+ve)	False Positive (FP)	True Positive (TP)

Accuracy (A), which is the total number of correct predictions, is a commonly used metric in the evaluation of a NIDS. It is evaluated using the formula

Accuracy (A) = TP + TN / TP + TN + FP + FN

Recall (R) or Sensitivity or True Positive Rate (TPR) or Detection Rate (DR) is the capability of the NIDS to detect the network connection as normal. The formula used is

Recall (R) = TP / TP + FN

False Positive Rate (FPR) is capability of the NIDS to identify normal as attack. The formula used is

False Positive Rate (FPR) = FP / TN + FP

Table 3: Performance of standard feature selection methods and proposed ARFS method

Feature Selection Method / Attack Class	ARFS		CFS		IG	
	TPR	FPR	TPR	FPR	TPR	FPR
Normal	99.5	1.02	93.07	1.74	97	0.02
DoS	99.72	1.31	92.14	1.49	96.7	0.001
Probe	99.06	0.20	90.35	0.4	98.8	0.011
U2 R	72.01	0..09	56.75	0.10	64	0.004
R 2 L	98.05	0.10	78.10	0.16	96.1	0.005

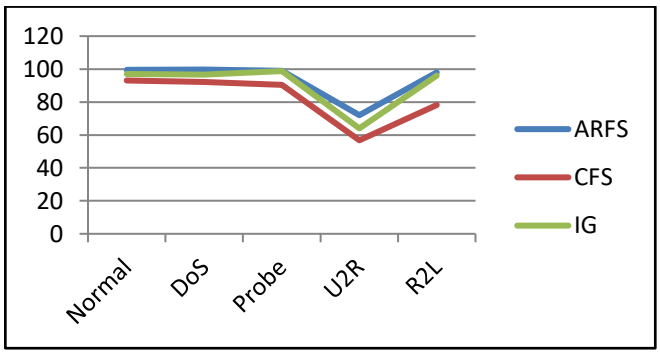

Figure 1: TPR comparison of various feature selection method along with ARFS

The Table 3 shows that the FPR calculated for Normal is 99.5, DoS is 99.72, Probe is 99.06, U2R is 72.01, R2L is 98.05, which is comparatively superior to the other conventional methods.

VIII. CONCLUSION & FUTURE WORK

The attained results shown in Figure 1 helps us conclude that NSL KDD is the best candidate audit dataset to evaluate the performance feature selection models. The proposed ARFS for feature selection method has helped us to realize a light-weight NIDS model with a better accuracy rate and low false positives. The tabulated results and comparison charts shown in the previous section have given a clear understanding on the NSL KDD dataset to the research community.

Our future will work will be a hybrid feature selection method that will further Improvise the performance of individual, hybrid and distributed NIDS.

REFERENCES

1. Liu, H & Motoda, H (Eds.) 2007, 'Computational methods of feature selection'. CRC Press.
2. Mukkamala, S, Sung, AH & Abraham, A 2005, 'Intrusion detection using an ensemble of intelligent paradigms'. Journal of network and computer applications, vol. 28, no. 2, pp. 167-182.
3. Chebrolu, S, Abraham, A & Thomas, JP 2005, 'Feature deduction and ensemble design of intrusion detection systems'. Computers & security, vol. 24, no. 4, pp. 295-307.
4. Sung, AH & Mukkamala, S 2003, 'Identifying important features for intrusion detection using support vector machines and neural networks'. Applications and the Internet, 2003. Proceedings. 2003 Symposium on IEEE, pp. 209-216.
5. H Nguyen, K Franke, S Petrovic - Improving Effectiveness of Intrusion Detection by Correlation Feature Selection, 2010 International Conference on Availability, Reliability and Security,IEEE Pages-17-24
6. S Chebrolu, A Abraham, J P. Thomas Feature deduction and ensemble design of intrusion detection systems, Computers & Security, Volume 24, Issue 4, June 2005, Pages 295-307
7. Wafa' S.Al-Sharafat, and Reyadh Naoum "Development of Genetic-based Machine Learning for Network Intrusion Detection" World Academy of Science, Engineering and Technology 55, 2009.
8. Ms.Nivedita Naidu, Dr.R.V.Dharaskar "An effective approach to network intrusion detection system using genetic algorithm", International Journal of Computer Applications (0975 - 8887) Volume 1 - No. 2, 2010.
9. NSL-KDD dataset for network -based intrusion detection systems" available on http://iscx.info/NSL-KDD/.