Comparative Analysis of Feature Selection Techniques for Network Intrusion Detection

Vinay Jain
B.E Graduate, Netaji Subhas Institute of Technology, 2014-18

***Abstract*- Network traffic is growing exponentially with the increase in use of smart devices and the internet. Commensurately, the chances of network intrusion are also growing. Processing massive data in near real time is the bottleneck in the performance of intrusion detection systems. If the dimension of data in use could be reduced to including only those features which are significantly important for intrusion detection, this can help in increasing the performance of intrusion detection systems. This is where feature selection (FS) techniques play an important role because it provides the classifiers to be fast, cost-effective, and more accurate. In this paper, three feature selection methods are analyzed; FI (Feature Importance), RFE (Recursive Feature Elimination), ANOVA (Analysis of variance). Features from these FS methods are then learned using machine learning models; Random Forest (RF) and Multi-Layer Perceptron (MLP). Later, a comparative analysis of the accuracies and ROC curves for each method is done.**

Index Terms- NSL-KDD dataset, Feature Selection, Intrusion Detection, Classification models, ROC Curves

I. INTRODUCTION

The 21st century is a digital era where a heavy usage of devices that generate digital data which are processed through networks, is very pervasive. This results in increased network traffic, increased network attacks/intrusions and hence, increased importance of having network security. Intrusions are defined as attempts or action to compromise the confidentiality, integrity or availability of computer or network. Intruders do many attempts to gain access to the network and try to harm the organization's data. Computer systems also have some security vulnerabilities that are very difficult to overcome [1]. Therefore, Intrusion Detection Systems (IDS) [2], [6] play a very important role in networks as they act as devices made solely for detecting attacks or anomalies in the network. It is the networks defense mechanism against network attacks which basically differentiates benign intrusions from malicious intrusions. Due to these reasons, intrusion detection has been an important research issue.

High dimensional data are computationally expensive. Sometimes, there are some features of the dataset that are irrelevant with respect to the task at hand and thus, they do not contribute much to the dependent variable which is what we are trying to predict. This is where feature selection (FS) comes in handy. FS is the process of removing features from the original data set that do not have any significant impact on the outcome variable [3]. FS methods not only reduce the time complexity of the computation of the dataset, but also increases the accuracy of the predictions made in the dataset.

In this paper, NSL-KDD dataset, an improved version of KDD-CUP99 dataset [4], [7], is used for training our machine learning models, namely Random Forest (RF) and Multi-Layer Perceptron (MLP). Three FS techniques; Feature Importance, Recursive Feature Elimination, Analysis of Variance; are used to select a subset of features from all the features of the dataset. Then these selected features are used to train the models RF and MLP from train dataset and predict benign or malicious intrusion on test dataset. All three FS methods are used for each classification model. The models are evaluated by their accuracies for different number of features and by ROC curve analysis for the highest accuracy for every model.

The rest of this paper is organized as follows: Section 2 explains related work done, Section 3 gives a brief introduction to FS methods, Section 4 summarizes the experiment carried out and the results thus obtained and Section 5 concludes the work and suggests future work.

II. RELATED WORK

The authors of [8] use information gain property of decision trees on NSL-KDD dataset to segregate 16 features that have significant impact on the outcome variable. Later decision trees are used as classifiers on test data which gives an accuracy of 79.53% in 50.87 seconds. A comparative analysis with few other classifier models is also done.

The authors of [9] propose a selection algorithm (FMIFS) which selects optimal features for classifications analytically. Least Square Support Vector Machine (LSSVM) intrusion detection system is used as a classifier with different selection algorithms. LSSVM with FMIFS for KDD-CUP99 dataset gives the highest accuracy of 99.79%, highest detection rate (DR) of 99.46% and lowest false positive rate (FPR) of 0.13.

The authors of [10] perform a comparative analysis between various FS methods while using J48 model as the classifier. They propose a new method of selecting features from the union of two FS methods (OneR and Relief). The accuracy of J48 classifier (66.807%) was highest for the proposed method with only 12 features.

Aghdam and Kabiri [11] use ant colony optimization technique to reduce the number of features of KDD-CUP99 dataset. Nearest neighbor model is used as a classifier which gave an accuracy of 98.59% in detecting intrusion attempts and lower false alarm rate (2.59%) with reduced number of features (appx. by 88%).

Gharaee and Hosseinvand [12] propose a combination of Genetic algorithm with Support Vector Machine (SVM) to detect anomalies. The new fitness function of the genetic algorithm evaluates feature chromosomes considering their effectiveness on True and False Positive rates by using a SVM classifier. This method finds the optimal features for every intrusion type (Normal, DoS, Probe, U2R, R2L) and trains model on them accordingly. Results are increased accuracies and decreased FPR for every intrusion type. This study proposes a method which can achieve more stable features in comparison with other techniques.

Kyaw Thet Khaing [13] proposes an enhanced SVM model which comprises of Recursive Feature Elimination and k-Nearest Neighbor (KNN) method to perform a feature ranking and selection task of the new model. The proposed model is better than the conventional SVM model as it yields results with higher accuracies, lower false negative rate (FNR), improved precision and reduced time with 16 features only.

III. FEATURE SELECTION

Feature selection (FS) methods aims at segregating the significant features from the insignificant ones. Datasets with high dimensionality are usually computationally expensive and therefore, difficult to analyze easily. On the other hand, via FS methods the same dataset can be analyzed with reduced features or complexity; improving the accuracies of the models and hence, leading to a better intelligibility of the models [14], [15], [16].

There are three types of FS methods: Filter-based, Wrapper-based and Embedded methods [14], [15], [16]. Filter methods measure the relevance of features by their correlation with dependent variable while wrapper methods measure the usefulness of a subset of feature by actually training a model on it. Embedded methods combine the qualities' of filter and wrapper methods. It's implemented by algorithms that have their own built-in feature selection methods.

In this paper, the following FS methods have been used: Feature Importance (FI), Recursive Feature Elimination (RFE), Analysis of variance (ANOVA).

A. Analysis of Variance (ANOVA)

ANOVA [19] is a statistical technique to compare multiple population means through significance tests. It is a Filter-based approach. The impact of individual features' significance levels for the problem at hand can be computed via ANOVA. ANOVA is computed by comparing variance between different samples with variance within the sample, which is nothing but the F value.

$$F = \frac{MS_{EFFECT}}{MS_{ERROR}} \qquad (1)$$

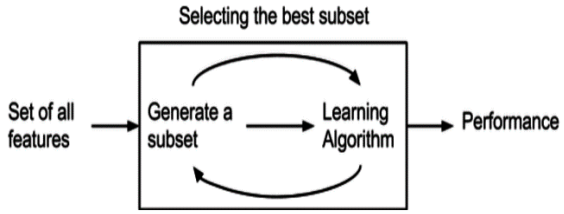
Refer [19] for further elaboration of (1).

B. Recursive Feature Elimination (RFE)

RFE computes the best subset of features from the total available features. It recursively tests a classifier model by including all features one by one and ranking them later on the basis of their performance.

It is a Wrapper-based method. Fig. 1 gives a brief explanation of RFE algorithm.

Fig. 1: RFE Flowchart



C. Feature Importance (FI)

It is a type of Embedded method which perform variable selection as part of the learning procedure and are usually specific to given learning machines. Examples are classification trees, random forests. The more important a feature is in determining the outcome variable, higher the score. The importance of features via RF [21] in a machine learning problem can be calculated by (2).

$$Imp(X_m) = \frac{1}{N_T} \sum_T \sum_{t \in T: v(s_t)=X_m} p(t) \Delta i(s_t, t) \quad (2)$$

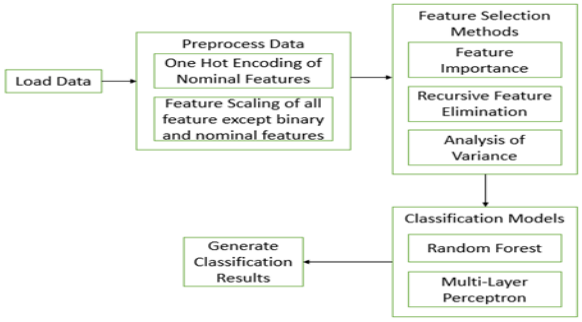
Refer [21], [22] for better understanding of FI using RF.

IV. EXPERIMENT AND RESULTS

A. Experiment Setup

Binary classification is performed, i.e. the classifiers have to classify intrusions into benign/normal or malignant/attack category. Also, each classification model (RF, MLP) performs classification for all the set of features selected by each FS method (FI, RFE, ANOVA). Other than complete set of features, number of features on which classification was performed varied from 5 to 40. RFE evaluated its subset of features using the random forest algorithm. FI also used random forest to rank the importance values to the features. Fig. 2 illustrates the experimental setup or the intrusion detection system used.

Fig. 2: Experiment Setup

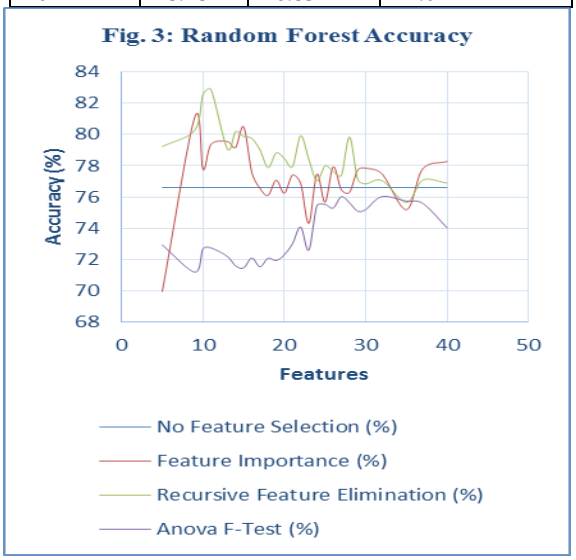


B. Result Analysis

Accuracy of RF on complete dataset (41 features) = 76.6%. The accuracies of RF calculated when trained on features from different FS methods is summarized in Table 1. Fig. 3 visualizes the Table 1.

Table 1: Accuracies of Random Forest for Feature Selection Methods.

Features	FI (%)	RFE (%)	ANOVA (%)
5	69.95	79.23	72.93
9	81.12	80.27	71.2
10	77.77	82.56	72.68
11	79.37	82.86	72.74
13	79.55	79.08	72.2
14	79.16	80.17	71.6
15	80.49	79.88	71.47
16	77.58	79.75	72.1
17	76.56	79.02	71.54
18	76.1	77.88	72.08
19	77.08	78.82	71.94
20	76.24	78.46	72.3
21	77.41	77.94	73.01
22	76.88	79.92	74.08
23	74.31	78.42	72.6
24	77.46	77.04	75.45
25	75.67	77.99	75.53
26	77.93	77.63	75.3
27	76.46	77.38	76.02
28	76.3	79.84	75.63
29	77.72	77.12	75.09
30	77.84	76.84	75.2
32	77.48	77.05	76.03
35	75.19	75.65	75.74
37	77.8	77.08	75.6
40	78.28	76.89	74.02



Accuracy of MLP on complete dataset (41 features) = 76.1%. The accuracies of MLP calculated when trained on features from different FS methods is summarized in Table 2. Fig. 4 visualizes the Table 2.

Table 2: Accuracies of Random Forest for Feature Selection Methods

Features	FI (%)	RFE (%)	ANOVA (%)
5	42.98	72.02	43.08
9	75.62	83.52	74.38
10	73.42	80.2	74.02
11	75.41	83.56	75.79
13	77.53	77.69	69.48
14	76.56	76.11	73.47
15	72.65	75.56	71.71
16	74.07	75.96	73.97
17	75.95	74.81	74.17
18	74.03	77.19	76.7
19	75.35	79.27	74.88
20	74.28	75.01	75.54
21	74.14	78.72	75.4
22	75.14	78.47	74.53
23	73.87	76.14	79.01
24	75.13	76.35	75.95
25	75.93	79.96	76.15
26	76.68	77.91	74.58
27	78.02	76.62	73.28
28	74.57	77.85	75.3
29	74.61	78.83	79.89
30	75.78	80.43	74.99
32	77.92	79.41	75.87
35	76.51	76.65	73.15
37	78.07	73.54	77.84
40	76.98	76.76	76.03

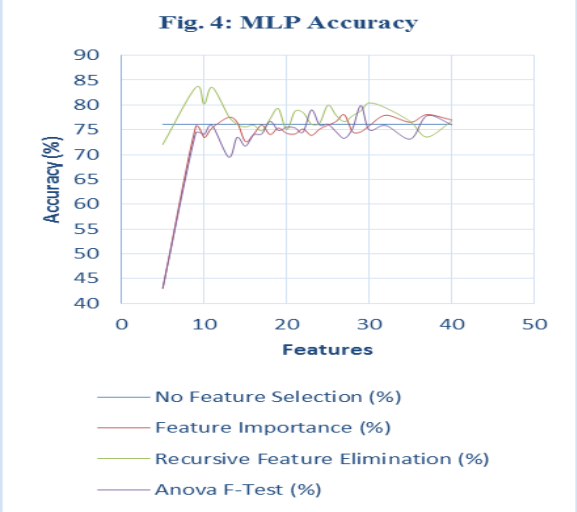


Table 3 shows the time needed by different FS methods to select different number of features. The computation time for RFE reduces with increase in number of features to be selected. Also for RFE, time needed to compute the features is significantly higher than the other two FS methods. Fig. 5 is a visualization of computation time for FI and ANOVA from Table 3.

Table 3: Computation Time for Feature Selection Methods

Features	FI (seconds)	RFE (seconds)	ANOVA (seconds)
5	0.23	118.25	0.44
10	0.25	117.57	0.42
15	0.24	110.20	0.44
20	0.30	106.94	0.45
25	0.32	98.00	0.47
30	0.36	94.03	0.44
35	0.35	91.78	0.46
40	0.40	85.26	0.47

It is visible from the Fig. 5 that computation time for both (FI and ANOVA) increases with increase in number of features to be selected but the rate of increase of computation time is higher in FI than in ANOVA.

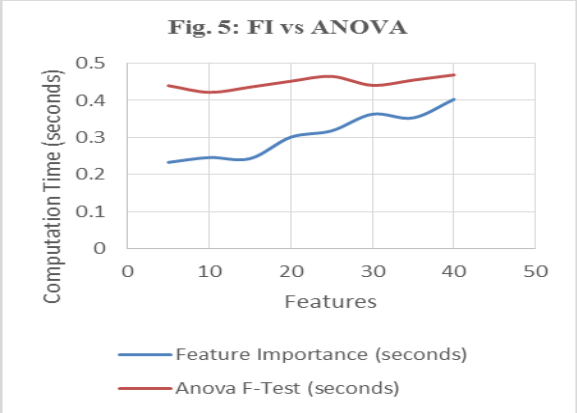


Table 4 is basically the summary of Table 1 and 2. It shows the highest accuracies of classification methods along with the number of features used for different FS methods.

Table 4: Highest accuracy of Classification Methods for Feature Selection Methods

	FI		RFE		ANOVA	
	Features	Acc. (%)	Features	Acc. (%)	Features	Acc. (%)
RF	9	81.12	11	82.86	32	76.03
MLP	37	78.07	11	83.56	29	79.89

Fig. 6 and Fig. 7 represent the ROC curves of RF and MLP when run on the complete dataset (41 features) and with different FS methods discussed so far.

Fig. 6:ROC Curve Analysis For Random Forest

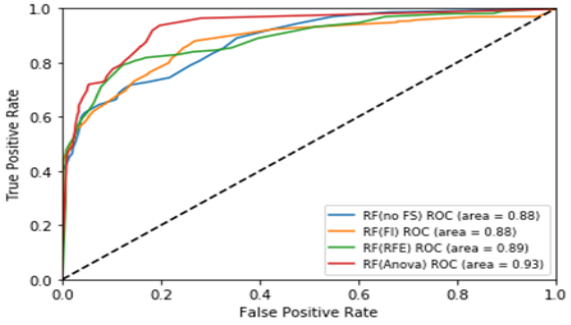
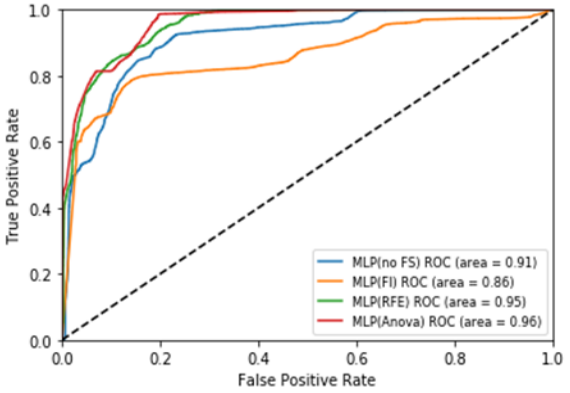


Fig. 7:ROC Curve Analysis For Multi-Layer Perceptron



According to Fig. 6 and Fig. 7, although all FS methods (FI, RFE, ANOVA) have a good area under the curve (AUC) score, the best FS method for both classification models seems to be ANOVA as it has the greatest AUC score (0.93 and 0.96 for RF and MLP respectively). However if we also look at Table 4, we see that RFE method, with only 11 features, gives the highest accuracy for both classification models (82.86% and 83.56% for RF and MLP respectively). Since RFE has the second best AUC score (0.89 and 0.95 for RF and MLP respectively) and gives the highest accuracy along with 73.17% reduced number of features for both classification models, RFE poses itself as a viable FS method for effective intrusion detection.

V. CONCLUSION AND FUTURE WORK

Feature selection methods play an important role in reducing any high dimensional data which is computationally expensive and time consuming to a relatively low dimensional data that with less number of features is capable of giving better results and in reduced amount of time. RF and MLP were the classification models used for detecting benign and malignant intrusions. FI, RFE and ANOVA were the FS methods discussed in this paper of which RFE proves to be a viable and better solution than FI and ANOVA for intrusion detection. With only 11 features i.e., a 73.17% reduction in number of features being used for classification, RFE gives the highest accuracy for both classification models (82.86% and 83.56% for RF and MLP respectively). The only disadvantage with RFE is that its computation time for selecting features is very high as it is a wrapper-based FS method and thus, requires constant evaluation for every set of features via a classification model, which was random forest in this case. The prediction time in testing phase however, is same for all feature selection methods for a particular model.

This work can be extended to other FS methods and other classification models with the aim being finding the optimal combination of machine learning models and FS methods in order to build an intrusion detection system that is capable of yielding excellent results with reduced time and reduced number of features.

ACKNOWLEDGMENT

The effort of M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani to propose a new dataset, NSL-KDD dataset, which overcomes all the problems of the KDD-CUP99 dataset and thus, makes it fit for intrusion detection purposes is gratefully acknowledged. Special thanks to University of New Brunswick, Canada for making this dataset open source, without which this analysis would have been impossible.

REFERENCES

[1] C. E. Landwehr, A. R. Bull, J. P. McDermott, and W. S. Choi, “A taxonomy of computer program security flaws,” ACM Comput. Surv, vol. 26, no. 3, pp. 211–254, 1994.

[2] R.Bace and P. Mell, “NIST Special Publication on Intrusion Detection Systems”, 2001.

[3] Yang, Yiming; Pedersen, Jan O., “A comparative study on feature selection in text categorization” In: ICML, 1997, pp. 412-420.

[4] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, “A Detailed Analysis of the KDD CUP 99 Data Set,” Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2009.

[5] J. McHugh, “Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory,” ACM Transactions on Information and System Security, vol. 3, no. 4, pp. 262–294, 2000.

[6] Srinoy, S., Chimphlee, W., Chimphlee, S., & Poopaibool, Y., “A fusion of ICA and SVM for

detection computer attacks." In: Proceedings of the 5th WSEAS international conference on Applied computer science. World Scientific and Engineering Academy and Society (WSEAS), 2006, pp. 986-990.

[7] L.Dhanabal, Dr. S.P. 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 Vol. 4, Issue 6, June 2015.

[8] Kajal Rai, M. Syamala Devi and Ajay Guleria, "Decision Tree Based Algorithm for Intrusion Detection", Int. J. Advanced Networking and Applications, Volume: 07, Issue: 04, pp. 2828-2834 (2016).

[9] Ananda Kumar K S, Arpitha K, Latha M N and Sahana M, "A Novel Approach for Intrusion Detection System Using feature Selection algorithm", International Journal of Computational Intelligence Research, Volume 13, Number 8, 2017, pp. 1963-1976.

[10] Krishan Kumar, Gulshan Kumar and Yogesh Kumar, "Feature Selection Approach for Intrusion Detection System", International Journal of Advanced Trends in Computer Science and Engineering, Vol. 2, No.5, pp. 47-53 (2013).

[11] Mehdi Hosseinzadeh Aghdam and Peyman Kabiri, "Feature Selection for Intrusion Detection System Using Ant Colony Optimization", International Journal of Network Security, Vol.18, No.3, PP.420-432, May 2016.

[12] Hossein Gharaee and Hamid Hosseinvand, "A New Feature Selection IDS based on Genetic Algorithm and SVM", 2016 8th International Symposium on Telecommunications.

[13] Kyaw Thet Khaing, "Enhanced Features Ranking and Selection using Recursive Feature Elimination (RFE) and k-Nearest Neighbor Algorithms in Support Vector Machine for Intrusion Detection System", International Journal of Network and Mobile Technologies, Vol. 1, Issue 1, June 2010.

[14] S.Vanaja, K.Ramesh kumar, "Analysis of Feature Selection Algorithms on Classification: A Survey", International Journal of Computer Applications (0975 – 8887), Volume 96– No.17, June 2014, pp. 28-35.

[15] Vipin Kumar, Sonajharia Minz, "Feature Selection: A literature Review", Smart Computing Review, vol. 4, no. 3, June 2014, pp. 211-229.

[16] Richard Zuech and Taghi M. Khoshgoftaar, "A Survey on Feature Selection for Intrusion Detection", In: Proceedings of the 21st ISSAT International Conference on Reliability and Quality in Design, August 6-8, 2015, pp. 150-155.

[17] Andrew H. Sung & Srinivas Mukkamala, "The Feature Selection and Intrusion Detection Problems", Advances in Computer Science, 2004.

[18] Veeran Ranganathan Balasaraswathi, Muthukumarasamy Sugumaran and Yasir Hamid, "Feature selection techniques for intrusion detection using non-bio-inspired and bio-inspired optimization algorithms", Journal of Communications and Information Networks, Vol.2, No.4, Dec. 2017.

[19] Tomasz NIEDOBA and Paulina PIĘTA, "Applications Of Anova In Mineral Processing", Mining Science, vol. 23, 2016, pp. 43−54.

[20] Eva Ostertagová and Oskar Ostertag, "Methodology and Application of One-way ANOVA", American Journal of Mechanical Engineering, 2013, Vol. 1, No. 7, pp. 256-261.

[21] Gilles Louppe, LouisWehenkel, Antonio Sutera and Pierre Geurts, "Understanding variable importance in forests of randomized trees", In NIPS'13 Proceedings of the 26th International Conference on Advances in Neural Information Processing Systems, 2013, vol. 1, pp. 431-439.

[22] Md. Taufeeq Uddin* and Md. Azher Uddin, "A Guided Random Forest based Feature Selection Approach for Activity Recognition", 2nd International Conference on Electrical Engineering and Information Communication Technology, May 2015.