International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 4 | ISSN : 2456-3307

# Redundant and Irrelevant Feature Detection System using Online Feature Selection Algorithm

**G. Supriya[1], Dr. A. Sureshbabu[2]**
[1]M. Tech, Department of CSE, JNTUA College of Engineering, Ananthapuramu, Andhra Pradesh, India
[2]Associate Professor, Department of CSE, JNTUA College of Engineering, Ananthapuramu, Andhra Pradesh, India

## ABSTRACT

Developing effective and adaptive security approaches in network has become more critical than ever before. Security techniques such as Intrusion detection system is highly recommended to provide security but such security defence systems facing decision making problems due to the presence of irrelevant and redundant feature. To handle this condition, online feature selection algorithm has been proposed in this paper. With the help of this proposed technique, intruder will be identified by classifying packets as anomaly and normal .
**Keywords:** Feature Selection, Intrusion detection, KDD Cup 99 Dataset.

## I. INTRODUCTION

Security techniques like authentication, firewall and data encryption are not satisfying challenges from intrusion skills in providing network security. Hence Intrusion Detection System (IDS) is highly recommended. C5 decision trees using Boosting technique and Kernel Miner are the earlier attempts on developing intrusion detection systems. Machine Learning techniques like Support Vector Machines (SVM) were also proposed to identify abnormal packets. These developments were concentrated mainly on four attacks that usually occurred in network. Dos attack, probing, U2R and R2L attack are those four types that were characterized by earlier intrusion detection systems. Apart from these attacks, now a days network traffic data is high in amount that became a challenge to the intrusion detection system to handle this "Big Data". Presence of huge data causes IDS to detect and classify the abnormal packets in slow way and also gives misclassified results. This also leads to high computational complexity due to the presence of large amount of features for such large data.

Hence, there is a requirement of removing unnecessary features prior to detecting anomaly packet. In this paper, a feature selection technique using online feature selection algorithm has been proposed to remove irrelevant features and also to identify intruder. This entire process will work on KDD-Cup 99 dataset which is a large network data that contains 41 huge features with millions of records.

The remaining sections of this paper deals with: Section 2 describes related work on LSSVM classifier and HFSA algorithm, KDD- Cup 99 dataset Working is presented in Section 3, the next Section 4 gives information of the experiments being conducted and the appropriate results and Section 5 concludes the paper.

## II. RELATED WORK

Following are related works towards Feature Selection methods over network data.

Due to the high grow in data dimensionality, feature selection has become an essential part in developing an efficient intrusion detection system. Mukkamala and Sung[2] proposed feature selection algorithm to reduce the features of KDD Cup 99 dataset. Chebrolu et al[3] proposed Markov blanket and decision tree analysis for feature selection that reduced 41 features to 12 features. Chen et al[4] makes use of Flexible Neural Tree(FNT) for developing Intrusion Detection System. Horng et

al[5] proposed SVM-based IDS that is a combination of hierarchal clustering and SVM.

Filter algorithms for feature selection that makes use of independent measures as a criteria for measuring the relation among different features have been used while dealing with large data in network in order to detect intruder. Let us discuss Hybrid feature selection algorithm(HFSA) along with LSSVM IDS classifier. HFSA that consists of two methods: Wrapper method, Filter method. Wrapper method partitions the lakhs of records into partitions such that each partition contains thousands of data. Filter method works on each partition that filters features using Mutual Information and Linear Correlation coefficient as a metric. Mutual Information gives relation between two features that indicates their statistical dependence. Features with high predictive power will have high mutual information.

Linear Correlation Coefficient is dependence measure that evaluates relationship between two features. LCC value close to zero indicates weak relationship between two same features. After having required features from HFSA, a LSSVM IDS takes input of these features of required dataset in order to detect normal and anomaly packet. Following are the related work towards Intrusion Detection System(IDS).

Mukkamala et al [1] surveyed the chance of assembling different learning methods include Neural Networks(NN),SVMs and Multivariate Adaptive Regression Splines (MARS) to detect Intrusions. In their investigation, performance in terms of classifying four attacks of each learning model has been compared. Toosi et al[6] made use of neurofuzzy classifiers in developing intrusion detection system. Detection decision was made using this classifier and the system was optimized by using genetic algorithm.

## III. PROPOSED SYSTEM

In Existing System, 41 features like Protocol type, duration, service, source, destination, flag, su-attempted etc,. are taken. By taking both necessary and unnecessary features that degrades the performance of the system in detecting intruder packet. To avoid this, online feature selection algorithm that automatically takes relevant features to detect intruder has been proposed in this paper. Moreover, earlier discussed approaches for building Intrusion detection systems were worked on datasets only. Hence, a real-time application has been proposed using Online feature selection algorithm.

**Algorithms:**

**Algorithm 1**: Modified perception for OFS

1: Input
- B.the number of selected features

2: Initialization
- $W_1 = 0$

3: For t=1,2….T do
4: Receive $X_t$
5: Make prediction Sgn ($X_t^T W_t$)
6: Receive $y_t$
7: If $y_t X_t^T W_t < 0$ then
8: $w_{t-1} = W_t - y_t x_t$
9: $w_{t-1} = \text{Truncate}(W_{t-1}, B)$
10: else
11: $w_{t-1} = W_t$
12: end if
13: end for

**Algorithm 2:** W=Truncate(W,B)
1: If $||W||_0 > B$ then
2: $W = w^{\wedge B}$ where $w^{\wedge B}$ is w^ with everything but the B largest elements set to 0
3: Else
4: W=w^
5: End if

In this approach, KDD Cup 99 dataset is used for training the system and identifies anomaly and normal packets just like in earlier approaches. The difference here is testing of this system will not be done on KDD cup but it makes use of Client and Server.

After being trained by KDD Cup 99dataset, the proposed system starts a Server to start its testing phase. The Server checks the Clients by taking Five most important features like (source_ IP_address,service,dst_host_srv_count,destination bytes,src_bytes). Checking of IP address allows the difference among abnormal and normal .Normal Users makes use of TCP protocol as service whereas Intruders makes use (ARP) Active Route Poisining or UDP protocol. Number of attempts made by the client

can also allow the system to detect abnormal user. These checking of clients will be done in runtime only hence the technique is referred as online feature selection. Thus one can detect the intruder and can also disconnect the abnormal user using online feature selection algorithm.

## IV. EXPERIMENTAL RESULTS

Online feature selection algorithm proposed in this paper being evaluated using KDD Cup99 dataset and the results are presented here the system will be trained by KDD Cup99 datasets and the testing will be done in online by having server and client model.

Online feature selection algorithm automatically selects required features and the system will be trained with these features so that in testing phase the proposed system successfully detects the intruder from the input. The server side operation using online feature selection algorithm is shown below diagram

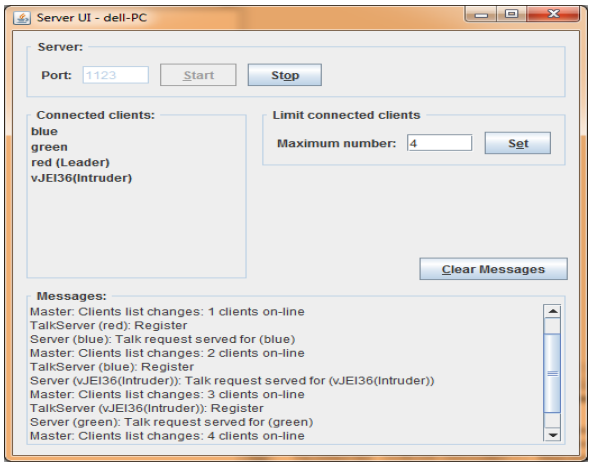


**Figure2.** Server Side operation indicating connected clients

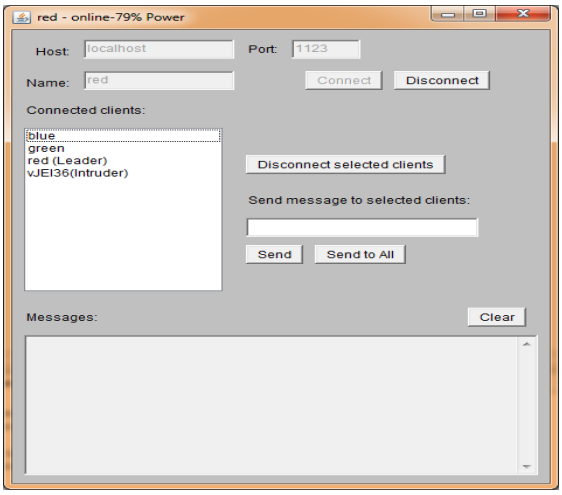


**Figure1.**Client side operation indicating leader, intruder

## V. CONCLUSION

The effect of redundancy problem doesn't guarantee uniqueness of data, which is a very crucial issue when we are dealing with decision making systems like Military databases or Medical Databases. KDD Cup 99 data set is used, this data set may have 41 features if we want to send a data from sender to receiver all the features will be sent among the data. To handle this redundant feature Online feature selection algorithm has been proposed. This approach automatically selects required features and the system will be trained with these features so that in testing phase the proposed system successfully detects the intruder from the input.

## VI. REFERENCES


[1]. S.Mukkamala,A.H.Sung,A.Abraham,Intrusion detection using an ensemble of intelligent paradigms, Journal of network and computer applications 28 (2) (2005) 167-182.

[2]. S. Mukkamala, A. H. Sung, Significant feature selection using computational intelligent techniques for intrusion detection, in Advanced Methods for Knowledge Discovery from Complex Data, Springer, 2005, pp. 285-306.

[3]. S. Chebrolu, A. Abraham, J. P.Thomas, Feature deduction andensemble design of intrusion detection systems, Computers &Security 24 (4) (2005) 295-307.

[4]. Y. Chen, A. Abraham, B. Yang, Feature selection and classification flexible neural tree, Neurocomputing 70 (1) (2006) 305-313.

[5]. S.-J. Horng, M.-Y. Su, Y.-H. Chen, T.-W. Kao, R.-J. Chen, J.-L.Lai, C.D.Perkasa, A novel intrusion detection system based on hierarchical clustering and support vector machines, Expert systems with Applications 38 (1) (2011) 306-313.

[6]. A. N. Toosi, M. Kahani, A new approach to intrusion detection based on an evolutionary soft computing model using neuro-fuzzy classifiers, Computer communications 30 (10) (2007) 2201-2212.