Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection

Harpreeth Kaur J[1], Anjali Kumari Singh[2], Pratyusha Chowdhury[3], Ashok Bhandari[4]

Prof. Sathya Priya A[5]

[1-4]Student, Dept. of Computer Science Engineering, Cambridge Institute of Technology, Karnataka, India
[5]Professor, Dept. of Computer Science Engineering, Cambridge Institute of Technology, Karnataka, India

---

Abstract - *Anomaly based Intrusion Detection Systems (IDS) learn normal and anomalous behaviour by examining network traffic in various standardized datasets. Some of the challenges for IDS are huge amounts of data to process, low detection rates and high rate of false alarms. In this paper, an approach based on the Online Sequential Extreme Learning Machine (OSELM) is introduced for intrusion detection. The proposed technique uses Symmetric Uncertainty based Feature Selection to minimize the time complexity while irrelevant features are discarded using an ensemble of Filtered, Correlation and Consistency based feature selection techniques. Considering anomaly pattern detection as detecting a point in time where the behaviour of the system is unusual and significantly different from past behaviour. In this context, anomaly (pattern) detection and concept drift detection mean detecting the behaviours that deviate from normal behaviours. Based at the experimental result achieved, we finish that the proposed method is an effective approach for network intrusion detection.*

***Key Words*: intrusion, machine learning, neural network, deep learning, feature selection.**

1. INTRODUCTION

In the world of rapidly developing technology, networks are facing threats like viruses, worms, Trojan horses, spyware, adware, root kits, etc[1]. These intrusions need to be identified before any type of loss to the organizations. Even internal Local Area Network (LAN) is also seriously struggling with intrusions [2]. This is affecting productivity of computer networks in terms of bandwidth and other resources. Hackers use advance features like dynamic ports, IP address spoofing, encrypted payload etc., to avoid detection. This type of intrusions can be detected by discovering patterns in network traffic dataset [3]. Due to huge and imbalanced dataset machine learning based Intrusion Detection System (IDS) faces problem to process entire data. So, it is necessary to identify intrusions through network traffic behaviour. IDS is designed to defend the network from malicious activities. Anomaly based IDS learn normal behaviour from network traffic dataset to detect attacks [4].Soft computing based IDS embraces several computational intelligence methodologies, including artificial neural networks, fuzzy logic, evolutionary computation, probabilistic computing, artificial immune systems, belief networks etc. This This project presents an intrusion detection technique that considers various issues like hugeness of network traffic dataset, feature selection, low accuracy and high rate of false alarms. Online Sequential Extreme Learning Machine (OSELM) is used to process network traffic dataset to detect intrusions [5]. It is fast and accurate single hidden layer feed forward neural network (SHLFN) which can process network. It has proved its applicability in classification by performing in single iteration [6].

Table below shows the Network Traffic Attacks:

Attack group	Attacks
Probe	ipsweep, mscan, nmap, portsweep, saint, satan
DoS	apache2, back, land, mailbomb, Neptune, processtable, pod, udpstorm, smurf, teardrop
U2R	buffer_overflow, httptuneel, loadmodule, perl, rootkit, xterm, ps, sqlattack
R2L	ftp_write, imap, guess_passwd, named, multihop, phf, sendmail, snmpgetattack, snmpguess, spy, warezclient, worm, warezmaster, zsnoop, xlock

1.1 Problem Statement

Streaming data is currently generated from many sources, like sensor networks, the world wide web, internet traffic, and real-time surveillance systems. Hence, outlier detection on data streams is an important data mining task. Streaming data cannot be saved permanently to memory, and it is difficult to process data streams using traditional data mining algorithms. Another issue is with streaming data comes from concept drift, where the underlying data

distribution can change over time. Considering anomaly pattern detection as detecting a point in time where the behaviour of the system is unusual and significantly different from past behaviour. In this context, anomaly (pattern) detection and concept drift detection mean detecting the behaviours that deviate from normal behaviours.

1.2 Existing System

An intrusion detection technique that considers various issues like hugeness of network traffic dataset, feature selection, low accuracy and high rate of false alarms [2]. Online Sequential Extreme Learning Machine (OS-ELM) is used to process network traffic dataset to detect intrusions [5]. It is fast and accurate single hidden layer feed forward neural network (SHLFN) which can process network instances one by one or in chunks. It has proved its applicability in classification by performing in single iteration.

Disadvantages in Existing System:

1. The feature selection method is not good, irrelevant and redundant features are present.

2. The classifier does not work well for limited training data set.

1.3 System Model

The system proposed is composed of feature selection and learning algorithm show in Fig.1. Feature selection component are responsible to extract most relevant features or attributes to identify the instance to a particular group or class.

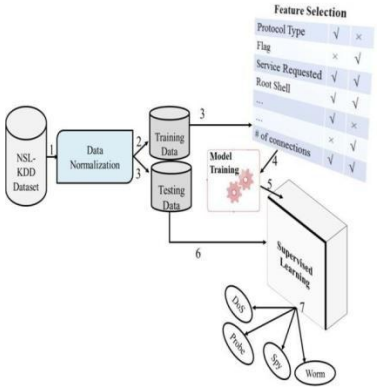

Fig 1: Proposed supervised machine learning classifier System.

A. Feature Selection.

Feature selection is an important part in machine Learning to reduce data dimensionality and extensive research carried out for a reliable feature selection method. For feature selection filter method and wrapper method have been used. In filter method, features are selected on the basis of their scores in various statistical tests that measure the relevance of features by their correlation with dependent variable or outcome variable. Wrapper method finds a subset of features by measuring the usefulness of a subset of feature with the dependent variable. Hence filter methods are independent of any machine learning algorithm whereas in wrapper method the best feature subset selected depends on the machine learning algorithm used to train the model.

2. PROPOSED SYSTEM

We propose solution for both the problems

1. Symmetric Uncertainty based Feature Selection.

We propose a new formula for feature selection based on Symmetric Uncertainty instead of CFS method mentioned in the base paper. The symmetric uncertainty is better indicator for the relation between the input variables and output classification result (intrusion or no intrusion). So applying this method, we can get the best features which are relevant and not redundant.

2. Classifier.

For improving the accuracy of the classifier for limited data set, we use a method called data expansion. We learn MLR (multivariate Linear Regression) based on input dataset and then use the MLR model learnt to generate further dataset and trained the classifier. By this way classifier can be trained well and accuracy is improved.

Aim / Objective of the Work

The networks are facing threats like viruses, worms, Trojan horses, spyware, adware, root kits, etc. These intrusions need to be identified before any type of loss to the organizations is the major objective of the work.

2.1 Methodology

The modules are showed in diagram, are:

IDS: An intrusion detection system (IDS) is a software program that monitors network or analyze system activities for malicious activities and produces electronic reports to a management station. Intrusion detection technique considers various issues such as huge of network traffic dataset, feature selection, low accuracy and high rate of false alarms.

Feature selection: this module takes training data set as input, training data set is helpful for multi stage classifier modeling.

Multistage Classifier Modeling: This module contains naïve bayes and MLR modelling.

MLR Modeling: MLR (multivariate Linear Regression) supported input files set then use the MLR model learnt to get further dataset and trained the classifier.

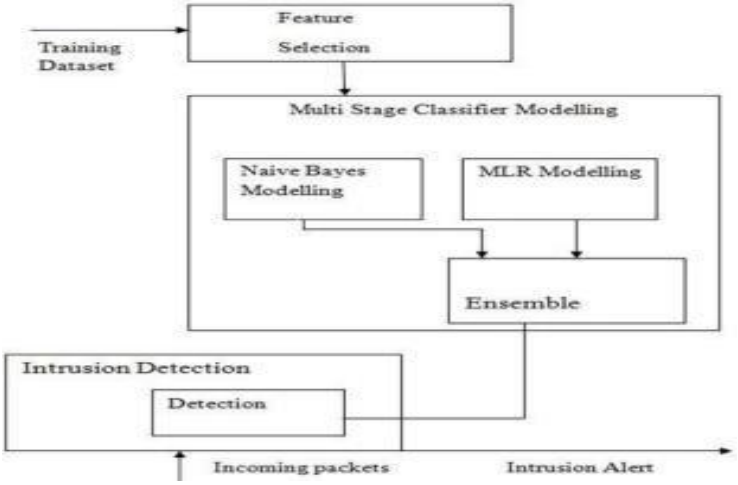

Fig 2: Methodology

3. CONCLUSION

During this paper, we introduced an adaptive ensemble model for classification and novel class detection in concept drifting data streams. More specifically, novel class instances in data streams are often automatically detected in our approach. Our work addresses challenging issues in data stream classifications like infinite length, limited labeled data. The experimental results proved that this ensemble classifier correctly detects the advent of novel class instances and also greatly improves the classification accuracy rates under different circumstances.

REFERENCES

[1] H. Song, M. J. Lynch, and J. K. Cochran, "A macro-social exploratory analysis of the rate of interstate cybervictimization," American Journal of Criminal Justice, vol. 41, no. three, pp. 583–601, 2016.

[2] P. Alaei and F. Noorbehbahani, "Incremental anomaly- based intrusion detection system using limited labeled data," in Web Research (ICWR), 2017 3th International Conference on, 2017, pp. 178–184.

[3] Srinivas Mukkamala, Guadalupe Janoski, Andrew Sung Intrusion Detection: Support Vector Machines and Neural Networks.

[4] Wei Li "Using Genetic Algorithm for NetworkIntrusion Detection.

[5] Cheng, Tay, &Huang, 2012 "Online sequential extreme learning Machine"(OS-ELM).

[6] Liu, Chen, Liao, & Zhang, "Intrusion detection techniques".