Intrusion Detection
In
Internet Of Vehicles

Abhijeet Srivastava *2020202011*
M.Tech CSIS
IIIT-H

Rishi Tripathi *2020202003*
M.Tech CSIS
IIIT-H

***Abstract*—With the rise of the Internet of Vehicles (IoV), Vehicles have become a connected node in the open internet. Vehicles need not only to communicate with other vehicles but also the external world. Also, with the innovation of Autonomous vehicles secure networks are far more important than ever needed. If a vehicle is compromised in a cyber-attack it can lead to severe consequences. Intrusion Detection System aim us to monitor the network and prevent any potential attack. In our work we propose Anomaly Based Intrusion Detection System Based on Voting Based Machine Learning Classifier.**

***Index Terms*—Internet of Vehicles, Machine Learning, Intrusion Detection System, Classifier**

I. Introduction

The speed of adoption of vehicles has increased lately. Modern vehicles with improved capabilities and intelligent system are common in these time. While it has provided convenience and facilities, these vehicles are becoming a frequent target for hackers.

Due to the inevitable rise of self driving-cars(Autonomous vehicles), providing a secure ride preventing accidents is major challenge . Intra-vehicle networks (IVNs) and external vehicular networks are two types of vehicle networks that need to be monitored. Vehicles have a number of Electrical Control Units(ECUs) which are connected to controller area network bus for transmitting messages. In external network, vehicles connect to outer environment devices. The heart of IoV is the use of V2X in which data is exchanged not only among vehicles but also infrastructures and pedestrians with Vehicle-to-Infrastructure (V2I), Vehicle-to-Pedestrian (V2P). Since wireless sensors in the vehicles are carefully designed, but they have a lot of security vulnerabilities and attacks that are possible via the open network.

IOV suffers from all common attacks like Distributed Denial of Service (DDoS) attack, fake information attack, Sybil Attack, black hole attacks, wormhole attacks. Ex IOVs can be effected by suffer Sybil Attack, the hacker can simulate many virtual vehicle nodes at a certain location causing vehicle node to interpret it as a congestion. Hackers could also compromise the sensor of vehicle and task them to perform malicious behaviours causing accidents. There is a case, in which a jeep was compromised causing it to turn the steering and use of parking brakes abruptly which caused severe accidents.

Remote attacks are leading attacks, in which a remote hacker wants to invade and break the normal behaviour of system by sending malicious data to vehicle, pedestrian or traffic nodes. Use of intrusion detection system becomes imperative to be installed in each nodes which screens the data flowing in the node and flags the attack pattern.

Machine learning (ML) and data mining algorithms have been recognized as effective models to design IDSs. The popular types of intrusion detection systems are either anomaly detection or misuse detection systems. In industries misuse detection system are quite popular whereas anomaly detection are limited to academics for research and developments. But overall an Intrusion detection system needs to be accurate. That is the reason IDS should be trained on a effective database.

II. Literature Review

There have been numerous attempts to resolve a class imbalance dataset and develop a high performing IDS.

Hongpo Zhang et al [1] proposed a IDS based on convolutional neural network and have used SMOTE for class imbalance problem in the dataset.They have achieved detection rate of 99.85 for 15-class classification on the CICIDS2017 dataset.

Arif Yulianto et al [2] have developed a AdaBoost-based Intrusion Detection System (IDS) using Synthetic Minority Oversampling Technique (SMOTE), Principal Component Analysis (PCA), and Ensemble Feature Selection (EFS) for class imbalance and feature selection.Using this method they have achieved a accuracy, precision, recall, and F1 Score of 81.83%, 81.83%, 100%, and 90.01% respectively.

Razan Abdulhammed et al [3] have used Features Reduction Approaches to develop a IDS.They were able to reduce the CICIDS2017 dataset's feature dimensions from 81 to 10, while maintaining a high accuracy of 99.6% in multi-class and binary classification.

Aamer Hanif et al [4] have proposed three methods for resolution of class imbalance and feature selection for dataset preprocessing. They have used random forest as their classification machine learning algorithm.

III. Your Proposed Scheme

An intelligent IDS not only needs to have a high accuracy but also low false positive cases. Therefore we propose a voting based classifier for our IDS. A voting based classifier is a machine learning model that trains on an ensemble of numerous machine learning and predict the class by taking the highest majority votes. Its of two types, a hard voting classifier where a class is selected based on the majority vote. In soft voting classifiers each model generates the probabilities score and the average of all probabilities of individual models are used. We have experimented on both and preferred soft max.

A. Dataset And Preprocessing

We will be using CICIDS2017 Dataset which is open source dataset comprises of five days of internet traffic from Canadian Institute of Cybersecurity. The dataset contains attack traffic as well as normal traffic. The dataset is huge, comprises of 8 days of traffic collected at the Canadian Institute of Cybersecurity. So we took a subset of this dataset. In our dataset we have 56580 rows with 78 features.

Some of the attacks present in the dataset are **WebAttack**, **Bot**, **Infiltration**, **Benign**, **DoS**, **PortScan**, **BruteForce**. DoS attack is the most common type of attack which floods the network by sending a large number of irrelevant messages or requests. This can lead to vehicles services being unavailable to the user or completely shut down. PortScan attack is a type of attack that scan the range of port address in order to find the potential active open port that can be used to launch another attack on the vechicle.

We will be normalising our dataset using StandardScaler of sklearn library. StandardScaler subtracts the feature mean for each sample and then scales to unit variance by dividing it with the standard deviation. It uses the following equation.

$$z = \frac{x - \mu}{\sigma}$$

where μ and σ are the column(feature) mean and standard deviation respectively

The dataset have high class imbalance therefore we have used ADASYN to resolve the class imbalance problem.

ADASYN(Adaptive Synthetic) is an algorithm that generates synthetic data, and its does not follow traditional method of replicating minority classes that produces rows with less accurate samples.The essential idea of ADASYN is to use a weighted distribution for different minority class examples according to their level of difficulty in learning, it produces more synthetic data for minority classes that are harder to learn as compared to those minority classes that are easier to learn. ADASYN have shown improvements in machine learning algorithm that have a class imbalanced data set. It calculates the neighbourhood impurity for each of the minority observations by looking in the k neighbourhood of a samples and compute $r_i = \frac{\delta_i}{i}$, where δ_i are number of majority samples in k neighbourhood. It normalized the r_i values and convert it into a probability distribution. For each sample to be generated it computes $s_i = x_i + (x_{zi} x_i))$where perturbs the sample from the straight line.

B. Training

We will be using **Random Forest**, **XGBoost classifiers**, **LightGBM** for our voting classifer.

Random Forest: Random Forest classifier uses ensemble of decision trees in order to perform classification. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by voting is more accurate than that of any individual tree. Random forest algorithm can also measure and rank important features and could be used for feature selection. The algorithm works as follows

- Sample datasets into m samples with replacement $D_1, D_2, ...D_m$
- Train a full decision tree for each D_j train for max depth but only select subset of features $k \leq d$ without replacement for each split
- The final classifier is $h(x) = \frac{1}{m} \sum_{j=1}^{m} h_j(x)$

XGBoost: XGBoost is boosting classifier which on classification boosts attributes that led to it. Its an improvement to the gradient boosting classifier and considered as one of the best off shelf classifiers. It has its benefits It has an implicit regularization which reduces variance and penalize objective function with both L1 and L2 regularization. It also has inbuilt capacity to handle missing values. It used sparsity awareness splitting handles the case if the data has lot of zeros.

$$L(\phi) = \sum_i l(\hat{y_i}, y_i) \ + \sum_k \Omega(f_k)$$

where $\Omega(f) = \gamma T \frac{1}{2} \lambda ||w||^2$. The first part is loss function which computes pseudo residues while second is a regularize term.

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{t-1} + f_t(x_i)) + \sum_k \Omega(f_k)$$

It finds an optimal output value for the leaf that minimizes the equation and uses residues from previous steps, which on further simplification using second order approximation, loss value is obtained.

LightGBM: Its uses histogram based algorithm to split nodes. Continuous feature are stored in the discrete buckets. In LightGBM tree is grown in depth first fashion leaf-wise((best-first). It will select the leaf with max delta loss to grow and loss is less than algorithms that used level by level approach to grow.

Voting Classifier: Soft voting classifier is used and in the experiment, we found that setting relative weightage of random forest, XGBOOST and LGBM to [1,1,2], we found best accuracy.

IV. Analysis

On running the random forest classifier, we obtained $99.15 \mod$ accuracy and f1 score of $93.5 \mod$ Xg Boost gets we $97.99 \mod$ accuracy and f1 score of $97.7 \mod$, while the LGBM gets $99.19 \mod$ accuracy and f1 score of $90 \mod$. Finally the voting ensemble gets $99.21 \mod$ accuracy and f1 score of $99.1 \mod$. Experiments show that our IDS achieves very high detection accuracy and low false positives.

V. Comparative Study

The following table shows our model performance against the previous work.

TABLE I
Comparative Studies

Works	Dataset	method	classes	Classifier	Acc %	F1 %
[1] Hongpo Zhang et al (2020)	CICIDS2017	SGM	multiclass	RF	93.08	94.67
				MLP	99.60	99.69
				CNN	99.85	99.86
[2] Arif Yulianto et al (2019)	CICIDS2017	SMOTE	binary	AdaBoost	81.83	90.01
[3] Razan Abdulhammed et al (2019)	CICIDS2017	UDBB	multiclass	RF	98.8	98.8
				NB	97.6	97.7
				LDA	95.7	95.7
				QDA	98.9	99.0
[4] Aamer Hanif et al (2017)	Customer Churn	Undersampling Oversampling	binary	RF	98.5	
Our Work	**CICIDS2017**	**ADASYN**	**multiclass**	RF	99.15	93.5
				XgBoost	97.99	97.7
				LGBM	99.19	99.17
				Voting Classifier	99.20	99.1

VI. Future Work

In future work, the results of the proposed system on CICIDS2017 data set can be further improved by working on the feature engineering.

References

[1] Hongpo Zhang, Lulu Huang, Chase Q. Wu, Zhanbo Li, An effective convolutional neural network based on SMOTE and Gaussian mixture model for intrusion detection in imbalanced dataset, Computer Networks, Volume 177, 2020, 107315, ISSN 1389-1286, https://doi.org/10.1016/j.comnet.2020.107315.

[2] Yulianto, Arif & Sukarno, Parman & Suwastika, Novian. (2019). Improving AdaBoost-based Intrusion Detection System (IDS) Performance on CIC IDS 2017 Dataset. Journal of Physics: Conference Series. 1192. 012018. 10.1088/1742-6596/1192/1/012018.

[3] Abdulhammed, Razan Musafer, Hassan Alessa, Ali Faezipour, Miad Abuzneid, Abdelshakour. (2019). Features Dimensionality Reduction Approaches for Machine Learning Based Network Intrusion Detection. Electronics. 8. 322. 10.3390/electronics8030322.

[4] Hanif, Aamer & Azhar, Noor. (2017). Resolving Class Imbalance and Feature Selection in Customer Churn Dataset. 82-86. 10.1109/FIT.2017.00022.