ATTACK AND ANOMALY DETECTION IN IOT SENSORS AND SITES

Team Number : 18

Team Members:
- Amith Bhat (181IT105)
- Harsh Agarwal (181IT117)
- Kumsetty Nikhil Venkat (181IT224)

INTRODUCTION

- With the increasing demand and growth in the Internet of Things (IoT) automated network system, the IoT models are getting larger and more complicated day by day.

- The growing complexity in IoT infrastructures is raising unwanted vulnerability to their systems. In IoT devices security breach and anomaly has become common phenomena nowadays.

- IoT devices use a wireless medium to broadcast data which makes them an easier target for an attack . Normal communication attack in the local network is limited to local nodes or small local domain, but attack in IoT system expands over a larger area and has devastating effects on IoT sites.

- Vulnerability in IoT nodes makes a backdoor for an attacker to gather confidential data from any important organization. For some stakeholders and entrepreneurs, data is the money for their business. For the government and some private agency, some data are classified and confidential.

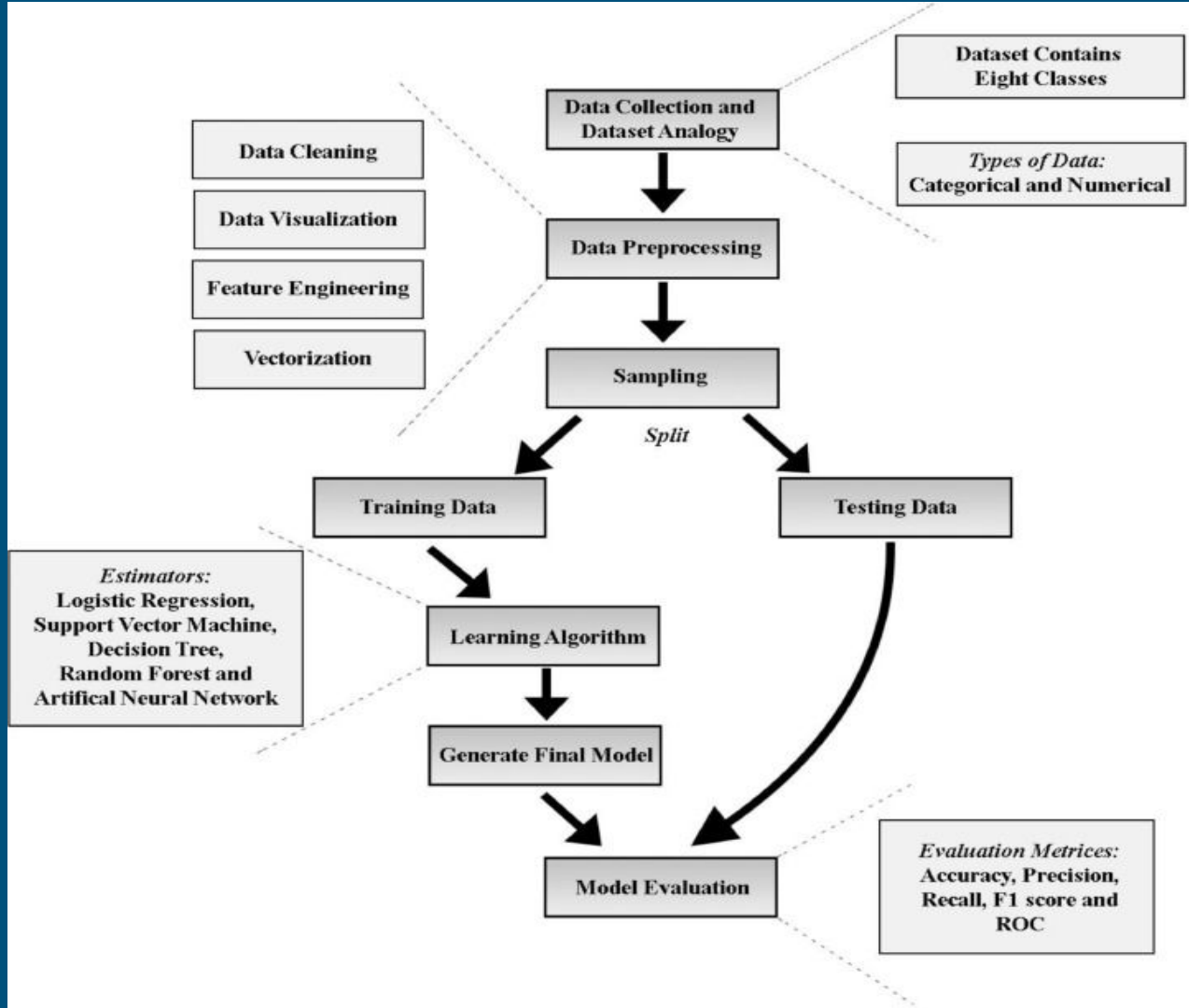
- Hence, a secured IoT infrastructure is necessary for protection from cybercrimes.

OBJECTIVES

- The primary goal of the system is to develop a smart, secured and reliable IoT based infrastructure which can detect its vulnerability, have a secure firewall against all cyber attacks and recover itself automatically.

- Here, an ML-based solution is proposed which can detect and protect the system when it is in the abnormal state. For this task, several ML classifiers such as Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Artificial Neural Network (ANN) have been analysed and compared.

METHODOLOGY



DATASET

- We have used an open source dataset which is considered as the benchmark dataset for projects in this field.
- The dataset contains 357,952 samples and 13 features, divided into 8 types.
- In the dataset, around 10000 samples are anomalous, while the rest are normal data. This ratio has been kept to mimic real-life situation to largest extent.
- The 8 classes are : Denial of Service, Data Type probing, Malicious Control, Malicious Operation, Scan, Spying, Wrong Setup, and Normal.
- Denial of Service has the highest proportion of samples among the anomalous ones(57%), again to mimic the real life situation.

DATA PREPROCESSING

Dealing with Missing and Unexpected Values :

The following rules were implemented to deal with missing and unexpected values :

1. NaN data in the “Accessed Node Type” column was replaced with “Malicious” value.
2. Any numbers in the value column in text form(such as “twenty”) were replaced by their numeric form (20.0)
3. Any other types of erroneous data was dropped.

FEATURE VECTOR

- The next part of the preprocessing is the feature vectorization. The 13 features in our dataset are mainly of two types : Categorical and Numerical.

- Categorical data can be converted into vectors in many ways such as label encoding and one hot encoding. In this project label encoding technique have been used to convert the data into a feature vector.. If one hot encoding were applied to these features, the number of features would have increased with a significant number, and the resulting dataset would have lots of dimensions.

Work Done

Logistic Regression (LR):

It is a discriminative model which depends on the quality of the dataset. Given the features X = X 1 , X 2 , X 3 , . . . , X n (where, X 1 −X n = Distinct features), weights W = W 1 , W 2 , W 3 , . . . , W n , bias b = b 1 , b 2 , . . . , b n and Classes C = c 1 , c 2 , . . . , c n (in our case, we have eight classes) the equation for estimation is given in following.

$$\text{Predicted Value: } p(y = C|X;W,b) = \frac{1}{1 + exp(-W^{transpose}X - b)}$$

Work Done (CONTD.)

SVM -

Support Vector Machine is another discriminative model like LR. It is a supervised learning model for analyzing the data used for classification, regression, and outliers detection. SVM is most applicable in the case of Non-Linear data. Given Input X , Class or Label C and LaGrange multipliers **α**; weight vector can be calculated by following equation:

$$\Theta = \sum_{i=1}^{m} \alpha_i c_i x_i$$

The target of the SVM is to optimize the following equation:

$$\text{Maximize}_{\alpha_i} \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j c_i c_j < x_i x_j >$$

In the above equation “<x i , x j >” is a vector which can be obtained by different kernels like polynomial kernel, Radial Basis Function kernel and Sigmoid Kernel.

Work Done (CONTD.)

Decision Tree -

A Decision Tree starts with a single node and then it branches into possible outcomes. Each of these outcomes lead to additional nodes, which branch off into other instances. Given, features x, impurity measure I(data), the number of samples in parent node Pn, the number of samples in left child LCn and the number of samples in right child RCn; DT's target is to maximize following Information Gain given as follows:

$$\text{Information Gain}(P_d, x) = I(P_d) - \frac{LC_n}{P_n} I(LC_d) - \frac{RC_n}{P_n} I(RC_d)$$

Impurity Measure I(data) can be calculated in three techniques Gini Index IG, Entropy IH and Classification Error IE:

$$I_H(n) = -\sum_{i=1}^{c} p(c|n) log_2 p(c|n)$$

$$I_G(n) = 1 - \sum_{i=1}^{c} p(c|n)^2$$

$$I_E(n) = 1 - max\{p(c|n)\}$$

Work Done (CONTD.)

Random Forest Classifier -

Random forest algorithm creates the forest with many decision trees. It is a supervised classification algorithm. It is an attractive classifier due to the high execution speed . Many decision trees ensemble together to form a random forest, and it predicts by averaging the predictions of each component tree. It usually has much better predictive accuracy than a single decision tree

Work Done (CONTD.)

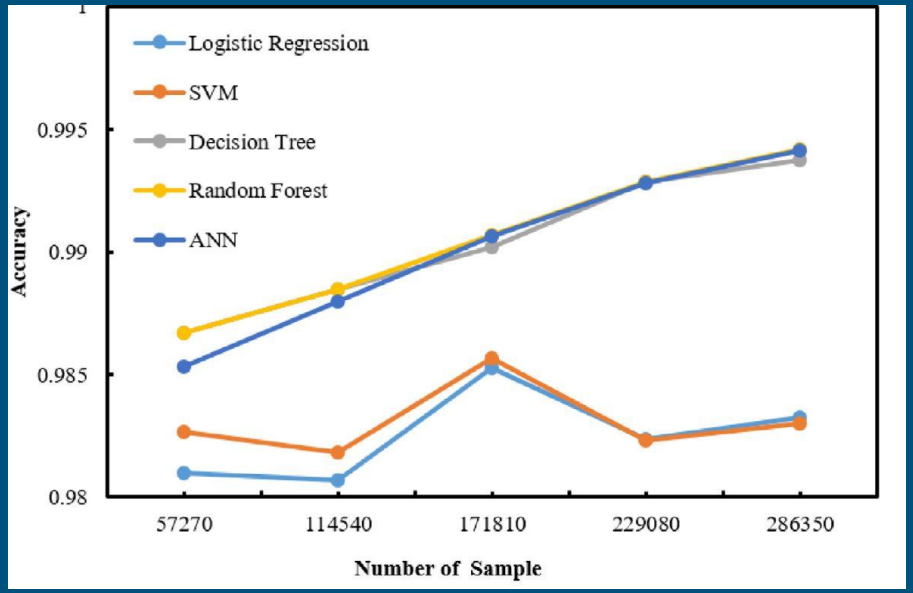
Artificial Neural Network -

Artificial Neural Network (ANN) is a deep learning method which trains the model based on raw data. Compared to other classifiers it has a large number of parameters for tuning which makes it a complex structure. It also takes a long time to optimize error than other techniques. Each single Neuron Node of ANN is trained with feature set X = X1, X2, X3, . . . , Xn (where, X1 − Xn = Distinct features). The features are multiplied by some random weights, W = W1,W2,W3, . . . ,Wn and added with bias values, b = b1, b2, . . . , bn. The values are then given as input in non-linear activation function

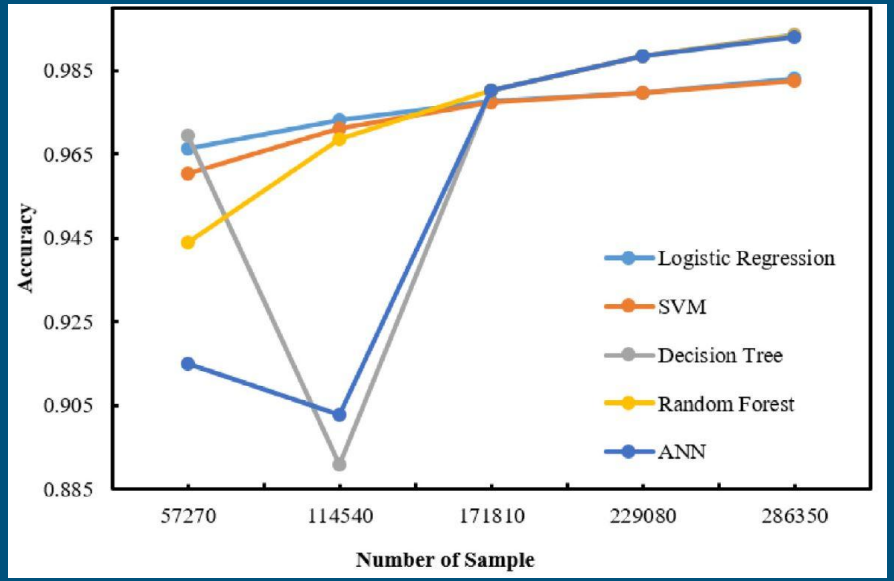
RESULTS

- Five-fold cross-validation was performed on the dataset using each of these techniques.
- RF and ANN have performed best both in training and testing accuracy.

Training



Testing



RESULTS

- DT and RF have more accuracy, precision, recall, and F1 score values than other techniques. ANN also performed well in the case of evaluation. However, DT and RF are a little more accurate than ANN.

Evaluation Metrics		Classifiers				
		LR	SVM	DT	RF	ANN
Training	Accuracy	0.983	0.982	0.994	0.994	0.994
	STD(+/-)	0.0012	0.0015	0.00081	0.00081	0.0013
	Precision	0.98	0.98	0.99	0.99	0.99
	Recall	0.98	0.98	0.99	0.99	0.99
	F1 Score	0.98	0.98	0.99	0.99	0.99
Testing	Accuracy	0.983	0.982	0.994	0.994	0.994
	STD(+/-)	0.0055	0.0064	0.016	0.014	0.021
	Precision	0.98	0.98	0.99	0.99	0.99
	Recall	0.98	0.98	0.99	0.99	0.99
	F1 Score	0.98	0.98	0.99	0.99	0.99

RESULTS

- From the confusion matrices, it can be concluded that RF is the best technique for Attack and Anomaly detection of IOT sensors.

LR

	DoS	D.P	M.C	M.O	SC	SP	W.S	NL
DoS	775	0	0	0	0	0	0	403
D.P	0	0	0	0	0	0	0	63
M.C	0	0	10	0	0	0	0	159
M.O	0	0	0	78	0	0	0	77
SC	5	0	2	0	0	0	0	298
SP	0	0	0	0	0	0	0	120
W.S	0	0	0	0	0	0	0	28
NL	34	0	0	9	0	0	0	69528

SVM

	DoS	D.P	M.C	M.O	SC	SP	W.S	NL
DoS	775	0	0	0	0	0	0	403
D.P	0	0	0	0	0	0	0	63
M.C	0	0	10	0	0	0	0	159
M.O	0	0	0	33	0	0	0	122
SC	0	0	2	0	0	0	0	303
SP	0	0	0	0	0	0	0	120
W.S	0	0	0	0	0	0	0	28
NL	34	0	0	0	0	0	0	69537

DT

	DoS	D.P	M.C	M.O	SC	SP	W.S	NL
DoS	775	0	0	0	0	0	0	403
D.P	0	63	0	0	0	0	0	0
M.C	0	0	169	0	0	0	0	0
M.O	0	0	0	155	0	0	0	0
SC	0	0	2	0	305	0	0	0
SP	0	0	0	0	0	120	0	0
W.S	0	0	0	0	0	0	28	0
NL	18	0	0	0	0	2	0	69551

RF

	DoS	D.P	M.C	M.O	SC	SP	W.S	NL
DoS	775	0	0	0	0	0	0	403
D.P	0	63	0	0	0	0	0	0
M.C	0	0	169	0	0	0	0	0
M.O	0	0	0	155	0	0	0	0
SC	0	0	2	0	305	0	0	0
SP	0	0	0	0	0	120	0	0
W.S	0	0	0	0	0	0	28	0
NL	18	0	0	0	0	2	0	69553

ANN

	DoS	D.P	M.C	M.O	SC	SP	W.S	NL
DoS	775	0	0	0	0	0	0	403
D.P	0	63	0	0	0	0	0	0
M.C	0	0	169	0	0	0	0	0
M.O	0	0	0	155	0	0	0	0
SC	0	0	2	0	305	0	0	0
SP	0	0	0	0	0	120	0	0
W.S	0	0	0	0	0	0	28	0
NL	18	0	1	0	0	2	0	69550

CONCLUSION

- In this study, RF performs comparatively better than all the other algorithms with the accuracy of 99.3%.

- It was found that the Random Forest algorithm is the best one to be applied on these kinds of datasets because RF predicted D.P, M.C, M.O, SC, SP, W.S attacks accurately compared to other approaches. It also predicted the DoS and Normal samples more accurately than any other ML model.

FUTURE WORK

- While this project has focused on classical ML models such as SVM, ANN, etc., we would like to work on designing and implementing a new algorithm specifically tailored for IoT systems.

- Also, the dataset used was created using virtual environment data, hence we would like to perform this project again using empirical, real-world data.