May - June 2020
ISSN: 0193-4120 Page No. 4413-4418

Detection of Cyber Attack in Network using Machine Learning Techniques

[1]Swamy Abhishek Revanna, [2]Ashwinkumar. U. M

[1,2]School of Computing & Information Technology,
REVA University, Bengaluru, Karnataka, India

Article Info
Volume 83
Page Number: 4413-4418
Publication Issue:
May - June 2020

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *12 May 2020*

Abstract

Contrasted with the past, improvements in PC and correspondence innovations have given broad and propelled changes. The use of new innovations give incredible advantages to people, organizations, and governments, be that as it may, messes some up against them. For instance, the protection of significant data, security of put away information stages, accessibility of information and so forth. Contingent upon these issues, digital fear based oppression is one of the most significant issues in this day and age. Digital fear, which made a great deal of issues people and establishments, has arrived at a level that could undermine open and nation security by different gatherings, for example, criminal association, proficient people and digital activists. Along these lines, Intrusion Detection Systems (IDS) has been created to maintain a strategic distance from digital assaults. Right now, learning the bolster support vector machine (SVM) calculations were utilized to recognize port sweep endeavors dependent on the new CICIDS2017 dataset with 97.80%, 69.79% precision rates were accomplished individually. Rather than SVM we can introduce some other algorithms like random forest, CNN, ANN where these algorithms can acquire accuracies like SVM – 93.29, CNN – 63.52, Random Forest – 99.93, ANN – 99.11.

Keywords: *data security, accessibility of information, digital fear, Intrusion Detection Systems.*

1. Introduction

PC wrongdoings keep on expanding throughout the years. They are not just confined to inconsequential acts, for example, assessing the login accreditations of a framework yet in addition they are significantly more dangerous. Data security is the way toward shielding data from unapproved get to, use, exposure, decimation, change or harm. The expressions "Data security", "PC security" and "data protection" are regularly utilized reciprocally. These territories are identified with one another and have shared objectives to give accessibility, secrecy, and honesty of data. Studies show that the initial step of an assault is disclosure [1]. Surveillance is made so as to get data about the framework right now. Finding an once-over of open ports in a structure gives incredibly essential information to an attacker. Consequently, there are lots of gadgets to recognize open ports [2], for instance, ant viruses and IDS. At this moment, learning and SVM AI computations were been applied to make IDS models to perceive port yield tries the models were presented with the explanation of used material and techniques.

2. Literature Review

Data security ideas involve of human, period, strategy, information, framework and innovation as is appeared in Figure 1. Classification, uprightness, and availability should be given by a protected framework. To begin with, the segregation of the data implies granting access just to the individual who needs to get to that data. Secondary, the honesty of the data is guaranteeing that the data is secured without bending and the primary structure is unblemished. At long last, the openness of data is the capacity to access and to use data at the ideal time

May - June 2020
ISSN: 0193-4120 Page No. 4413-4418

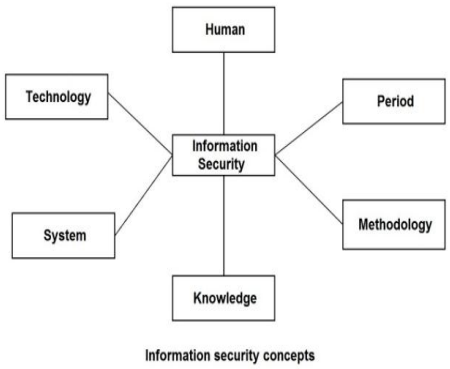

Figure 1: Information Security Concepts

As it is connoted by Stanford et al, there have been amazingly restricted working on that issue of distinguishing port sweeps [4].Robertson et al. utilized a limit technique to recognize the bombed association endeavors [5].Linear Discriminate Analysis (LDA) and Principal Component Analysis (PCA) were applied by Ibrahimi and Ouaddane to distinguish the interruption with NSL-KDD dataset [6]. Near outcomes of KDD99 and UNSW-NB15 informational indexes examining system practices were appeared by Moustafa and Slay [7]. Liuying et al. recognized and orchestrated pernicious models in sort out traffic reliant on the KDD99 dataset [8]. Blameless Bayes and Principal Component Analysis (PCA) were been used with the KDD99 dataset by Almansob and Lomte [9].Similarly, PCA, SVM, and KDD99 were used Chithik and Rabbani for IDS [10]. In Aljawarneh et al's. Paper, their assessment and examinations were conveyed reliant on the NSL-KDD dataset for their IDS model [11]

Composing inspects show that KDD99 dataset is continually used for IDS [6]–[10].There are 41 highlights in KDD99 and it was created in 1999. Consequently, KDD99 is old and doesn't give any data about cutting-edge new assault types, for example, multi day misuses and so forth. In this manner we utilized a cutting-edge and new CICIDS2017 dataset [12] in our investigation. There are distinctive yet constrained investigations dependent on the CI-CIDS2017 dataset. Some of them were talked about here. D.Aksu et al. demonstrated exhibitions of different AI calculations recognizing DDoS assaults dependent on the CICIDS2017 dataset in their past work [13]. They didn't make a difference all dataset and utilized restricted information 26.167 DDoS and 26.805 generous examples from the above dataset in their investigation. Additionally, they utilized that Fisher score include determination calculation to choose the best highlights. Hence, their past SVM models arrived at an extremely high precision result. In any case, they were intending to apply profound learning calculation as a component work to distinguish DDoS assaults. N. Marir et al. proposed a circulated examination to find anomalous movement in a huge scope arrange [14]. In another investigation, Resende et al. utilized hereditary calculations to identify interruptions on the CICIDS2017 dataset [15]

3. Material and Method:

In this section CICIDS2017 dataset, deep learning, SVM and Various Methodologies are being explained.

A. CICIDS2017 Dataset

The CICIDS2017 dataset is utilized in our examination. The dataset is created by the Canadian Institute for Cyber Security and incorporates different basic assault types. Right now, this is centered on port output endeavors. There are 692703 records comprising, and converted into 691406 for example source IP, source port, goal port, stream term, all out fwd parcels, all out in reverse bundles and so on. A piece of the records is as appeared in Table I. While making the dataset, Attack-Network and Victim-Network, totally were isolated two systems, were de-marked and executed by Sharafaldin H. et al [12]. They gathered information from July 3, 2017, to July 7, 2017, for the dataset.

B. SVM

Factual learning's and arched improvement, in light of the rule of basic hazard minimization, structure the premise of Support Vector Machine (SVM) calculations. Vapnik et al created SVM as an answer for various issues [16]. For instance, it very well may be utilized in various zones, for example, learning, design acknowledgment, relapse, grouping, and investigation.

Table 1: A Sample Set of Records from Dataset

Destination Port	Flow Duration	Total Fwd Packets	Total Bwd Packets
49666	3	2	0
49413	4	3	0
3268	4955405	33	24
49441	130	1	2
44459	28460	14	9
13792	29	1	1
13791	54	1	3
35215	35831	15	6
40372	4	3	0
49489	16	1	1
13810	3	2	0
5353	56005079	44	0
49469	3	2	0
49666	60590273	27	22
49471	3	2	0
49528	3	2	0

SVM is a directed learning strategy since it utilizes labeled information in a dataset as info. The quantity of yield classes changes relying upon that dataset. For instance, 2 classes of yield information are produced

May - June 2020
ISSN: 0193-4120 Page No. 4413-4418

when that dataset of 2 classes is given as that info. Subsequently, those classes arrange the examples given as that info. During the preparation procedure, a model is made by the information dataset and order is been performed by utilizing the model.

C. Deep Learning

Profound Learning calculations permit to separate highlights consequently from the given dataset and they comprise of a successive layer design. Applying non direct change capacities to the successive layered structure establish the premise of profound learning calculations. Expanding the quantity of layers will build the multifaceted nature of nonlinear changes to be developed. Profound learning calculations gain proficiency with the conceptual shrouded properties of that information got from the last layer in its theoretical portrayals obtained at various levels. In this way, the theoretical properties of the last layer yield are gotten by bringing the information into a significant level non direct capacity.

D. Methodology

The SVM, ANN, CNN, Random Forest and profound learning calculations were utilized to recognize port output endeavors dependent on the CICIDS2017 dataset. The flowchart of that invented strategy was introduced in figure. Most importantly, 692703 records from that dataset and afterward these records were almost all standardized. After standardization tests were part into two as a 75% preparing information and 25% testing information. Likewise, the SVM and profound learning IDS models were made dependent on the preparation information. At last, the models were tried with test information and the presentation of models was determined nearly. What's more, the profound realizing IDS model comprise of 7 shrouded layers of each layer incorporate the diverse number of neurons, for example, 100,70,40,150 and 6 individually. Contingent upon the quantity of neurons and shrouded layer model exhibitions were changed in this paper, we chose ideal numbers dependent on the model's precision. Then again, we didn't have any significant bearing any component choice calculation for SVM and we utilized all highlights. As a future work, we are going to utilize distinctive man-made reasoning ways to deal with characterize select this ideal qualities.

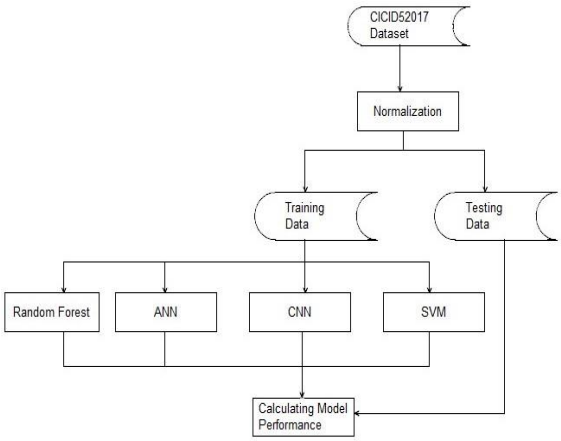

Figure 2: Flowchart of the method

First we have the dataset named CICID52017 and we need to perform normalization, data pre-processing. Later we need to split the data into training and testing. In training we need to train model by algorithms named Random forest, CNN, ANN, SVM. At last make use of 25% test data we need to evaluate our data.

As is shown in above "**Flowchart of the method**", important steps of the algorithm are given in below.

1) Normalization of every dataset.
2) Convert that dataset into the testing and training.
3) Form IDS models with the help of using RF, ANN, CNN and SVM algorithms.
4) Evaluate every model's performances.

In standardization, non-numeric name highlights were changed over into numeric structures. Likewise, irrelevant highlights, for example, Timestamp and a few examples that have NaN, boundlessness and void qualities were evacuated. Besides, we rescaled the all watched estimations of highlights to have the length of 1. As a subsequent advance, the standardized dataset was almost part into 75% preparing and 25% testing. Then In the third step, the IDS models were prepared and created to distinguish port sweep endeavors by utilizing the preparation information.

Consequently, the performances of those models were calculated.

Table 2: Confusion Matrix

Actual class\Predicted class	Normal (Benign)	Anomaly (port scan)
Normal (Benign)	TN	FP
Anomaly (port scan)	FN	TP

Table II can be explained in below items.

May - June 2020
ISSN: 0193-4120 Page No. 4413-4418

True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) statistics (Table II) are used for evaluation of model performances.

- TN: Actual Benign is classified as Benign.
- FP: Actual Benign is classified as Port Scan.
- FN: Actual Port Scan is classified as Benign.
- TP: Actual Port Scan is classified as Port Scan.

Accuracy, recall, precision and f1 score performance metrics are calculated using the statistics of the confusion matrix

Table 3: Performance Metrics

Measure	Formula
Accuracy	(TP+TN) / (TP+FP+FN+TN)
Recall	TP / (TP+FN)
Precision	TP / (TP+FP)
F1 score	2TP / (2TP+FP+FN)

The proportion of effectively anticipated perceptions is exactness, while accuracy implies a proportion of right positive perceptions. The review is an extent of accurately anticipated positive occasions. F1 score connotes the weighted normal of exactness and review.

Convolution layers are where channels are applied to the first picture, or to other element maps in a profound CNN. This is the place the vast majority of the client indicated parameters are in the system. The most significant parameters are the quantity of bits and the size of the bits.

The pseudo code of CNN algorithm is

```
function XCompressCU(*pCurCU)
    M ← FastCUMode(PO,QP)
    if M ≠ SPLIT then
        C2N ←CheckINTRA(pCurCU)
    else
        C2N ← ∞
    end if
    if M ≠ HOMO and DCUR < DMAX then
        CN ← 0
        for i = 0 to 3 do
            pSubCUi ← pointer to SubCUi
            CN ← CN+ XCompressCU(pSubCUi)
        end for
    else
        CN ← ∞
    end if
    CheckBestMode(C2N, CN)
end function
```

Here the accuracy of CNN is 63.52

```
print(classification_report(y_true=y_test, y_pred=y_pred_rounded))
```

	precision	recall	f1-score	support
0	0.64	1.00	0.78	87840
1	0.00	0.00	0.00	46085
2	0.00	0.00	0.00	1177
3	0.00	0.00	0.00	2072
4	0.00	0.00	0.00	1103
5	0.00	0.00	0.00	5
accuracy			0.64	138282
macro avg	0.11	0.17	0.13	138282
weighted avg	0.40	0.64	0.49	138282

```
acc_cnn = accuracy_score(y_true=y_test, y_pred=y_pred_rounded)*100
print("Accuracy in percentage : %f"%acc_cnn)
```

```
Accuracy in percentage : 63.522367
```

E. Random Forest

Random Forest: group model settled on of numerous choice trees utilizing bootstrapping, irregular subsets of highlights, and normal democratic to make forecasts. This is a case of a stowing troupe.

Training the Random Forest model

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report,confusion_matrix,accuracy_score
```

```
rfc = RandomForestClassifier(n_estimators=600)
```

```
rfc = rfc.fit(x_train,y_train)
```

```
predictions_rf = rfc.predict(x_test)
```

```
acc_rf = accuracy_score(y_true=y_test, y_pred= predictions_rf)
print("Overall accuracy of RF model using test-set is : %f" %(acc_rf*100))
```

```
Overall accuracy of RF model using test-set is : 99.936940
```

A random forest decreases the variance of a single decision tree leading to best predictions on new data.The accuracy of random forest is 99.93

The pseudo code of random forest algorithm is

```
Input:
    Data set T
    Set of p original features F = {f1,f2,..,f p}
Output:
    Subset of features
Code:
    Final ranking R
    Repeat for i in {1: p − 1}
        Rank set F using random forest
        f* ← last ranked feature in F
        *R(p −i +1) ← f*
        * F ← F − f*
```

SVM or Support Vector Machine is a straight model for gathering and to check backslide issues. It can deal

TEST Engineering & Management

May - June 2020
ISSN: 0193-4120 Page No. 4413-4418

with straight and non-direct issues and capacity commendably for some practical issues. The chance of SVM is fundamental: The computation makes a line or a hyper plane which segregates the data into classes. The accuracy of svm is 93.29

F. Support Vector Machine(SVM)

SVM Train Model

```
from sklearn.svm import SVC
from sklearn.metrics import classification_report,confusion_matri

model_svm = SVC()

model_svm.fit(x_train, y_train)

SVC(C=1.0, break_ties=False, cache_size=200, class_weight=None, c
    decision_function_shape='ovr', degree=3, gamma='scale', kerne
    max_iter=-1, probability=False, random_state=None, shrinking=
    tol=0.001, verbose=False)

predictions_svm = model_svm.predict(x_test)

acc_svm = accuracy_score(y_true=y_test, y_pred= predictions_svm)
print("Overall accuracy of SVM model using test-set is : %f" %(ac

Overall accuracy of SVM model using test-set is : 93.291949
```

The pseudo code of svm algorithm is

ACO_R-SVM Algorithm
Input: k, m, q, C, γ, and termination criterion
Output: Optimal value for SVM parameters and classification accuracy
Begin
  Initialize k solutions
  call SVM algorithm to evaluate k solutions
    T = Sort (S_1, ..., S_k)
  while classification accuracy ≠ 100% or number of iteration ≠ 10 do
    for i = 1 to m do
      select S according to its weight
      sample selected S
      store newly generated solutions
      call SVM algorithm to evaluate newly generated solutions
    end
    T = Best (Sort S_1, ... $S_k + m$), k)
  end
End

G. Artificial Neural Network (ANN)

Counterfeit Neural Network (ANN) utilizes the preparing of the cerebrum as a premise to create calculations that can be utilized to display complex examples and expectation issues. ... In our cerebrum, there are billions of cells called neurons, which forms data as electric signs. The pseudo code of ANN algorithm is

// Algorithm: Pseudocode of the random perturbation
For all particles i
Generate a random number($r_1 \in [0,1]$);
  If ($r_1 > 0.5$) **then** select particle i;
    For all dimensions of selected particle i
      Generate a random number ($r_2 \in [0,1]$);
      If ($r_2 > 0.5$) **then** $\overline{v}_{id} = v_{max} \times (2r_3 - 1) + v_{id}$ ($r_3 \in [0,1]$);
      If ($\overline{v}_{id} > V_{max}$) **then** $\overline{v}_{id} = V_{max}$;
      If ($\overline{v}_{id}(t) \leq v_{min}$) **then** $\overline{v}_{id}(t) = V_{min}$;
  End for d;
End for i;

```
print(classification_report(y_true=y_test, y_pred=y_pred_rounded))
```

	precision	recall	f1-score	support
0	1.00	0.99	0.99	109875
1	0.98	1.00	0.99	57507
2	0.98	0.99	0.98	1437
3	1.00	0.97	0.98	2630
4	0.97	0.98	0.98	1398
5	0.00	0.00	0.00	5
accuracy			0.99	172852
macro avg	0.82	0.82	0.82	172852
weighted avg	0.99	0.99	0.99	172852

```
acc_ann = accuracy_score(y_true=y_test, y_pred=y_pred_rounded)*100
print("Accuracy in percentage : %f"%acc_ann)

Accuracy in percentage : 99.115428
```

The accuracy of ANN is 99.11

H. Experimental Results

The PC which has Intel(R) Core(TM) i5 CPU M 460 @2.53 GHz, 4 GB Ram limit was used for tests. 692703 records, which were taken from the institutionalized dataset, were secluded into two sets with 75% planning and 25% testing extents, for instance, 518555 models for getting ready and 172852 models for testing. Execution estimation of the SVM, ANN, CNN and Random Forest learning models are presented in Table IV

Table 4: Performance Metrics of used Classification Technique based on Cicids2017 Dataset

Algorithm name	Precision	Recall	F1score	Accuracy
RF	1.00	1.00	0.99	99.93
ANN	1.00	0.99	0.99	99.11
CNN	0.64	1.00	0.78	63.52
SVM	0.92	0.95	0.98	93.29

Table IV shows the accuracy of all the algorithms with the help of factors which are mentioned in the above table like precision, f1 score, recall and this can be done calculating accuracy using formulae which will be coming up in next lines. These calculations helps in choosing the best algorithm based on the accuracy scores.

May - June 2020
ISSN: 0193-4120 Page No. 4413-4418

4. Conclusion And Future Works

Right now, estimations of help vector machine, ANN, CNN, Random Forest and profound learning calculations dependent on modern CICIDS2017 dataset were introduced relatively. Results show that the profound learning calculation performed fundamentally preferable outcomes over SVM, ANN, RF and CNN. We are going to utilize port sweep endeavors as well as other assault types with AI and profound learning calculations, apache Hadoop and sparkle innovations together dependent on this dataset later on. All these calculation helps us to detect the cyber attack in network. It happens in the way that when we consider long back years there may be so many attacks happened so when these attacks are recognized then the features at which values these attacks are happening will be stored in some datasets. So by using these datasets we are going to predict whether cyber attack is done or not. These predictions can be done by four algorithms like SVM, ANN, RF, CNN this paper helps to identify which algorithm predicts the best accuracy rates which helps to predict best results to identify the cyber attacks happened or not.

References

[1] K. Graves, Ceh: Official certified ethical hacker review guide: Exam 312-50. John Wiley & Sons, 2007.

[2] R. Christopher, "Port scanning techniques and the defense against them," SANS Institute, 2001.

[3] M. Baykara, R. Das¸, and I. Karado ˘gan, "Bilgi g ¨uvenli ˘gi sistemlerinde kullanilan arac¸larin incelenmesi," in 1st International Symposium on Digital Forensics and Security (ISDFS13), 2013, pp. 231–239.

[4] S. Staniford, J. A. Hoagland, and J. M. McAlerney, "Practical automated detection of stealthy portscans," Journal of Computer Security, vol. 10, no. 1-2, pp. 105–136, 2002.

[5] S. Robertson, E. V. Siegel, M. Miller, and S. J. Stolfo, "Surveillance detection in high bandwidth environments," in DARPA Information Survivability Conference and Exposition, 2003. Proceedings, vol. 1. IEEE, 2003, pp. 130–138.

[6] K. Ibrahimi and M. Ouaddane, "Management of intrusion detection systems based-kdd99: Analysis with lda and pca," in Wireless Networks and Mobile Communications (WINCOM), 2017 International Conference on. IEEE, 2017, pp. 1–6.

[7] N. Moustafa and J. Slay, "The significant features of the unsw-nb15 and the kdd99 data sets for network intrusion detection systems," in Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS), 2015 4th International Workshop on. IEEE, 2015, pp. 25–31.

[8] L. Sun, T. Anthony, H. Z. Xia, J. Chen, X. Huang, and Y. Zhang, "Detection and classification of malicious patterns in network traffic using benford's law," in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017. IEEE, 2017, pp. 864–872.

[9] S. M. Almansob and S. S. Lomte, "Addressing challenges for intrusion detection system using naive bayes and pca algorithm," in Convergence in Technology (I2CT), 2017 2nd International Conference for. IEEE, 2017, pp. 565–568.

[10] M. C. Raja and M. M. A. Rabbani, "Combined analysis of support vector machine and principle component analysis for ids," in IEEE International Conference on Communication and Electronics Systems, 2016, pp. 1–5.

[11] S. Aljawarneh, M. Aldwairi, and M. B. Yassein, "Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model," Journal of Computational Science, vol. 25, pp. 152–160, 2018.

[12] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization." in ICISSP, 2018, pp. 108–116.

[13] D. Aksu, S. Ustebay, M. A. Aydin, and T. Atmaca, "Intrusion detection with comparative analysis of supervised learning techniques and fisher score feature selection algorithm," in International Symposium on Computer and Information Sciences. Springer, 2018, pp. 141–149.

[14] N. Marir, H. Wang, G. Feng, B. Li, and M. Jia, "Distributed abnormal behavior detection approach based on deep belief network and ensemble svm using spark," IEEE Access, 2018.

[15] P. A. A. Resende and A. C. Drummond, "Adaptive anomaly-based intrusion detection system using genetic algorithm and profiling," Security and Privacy, vol. 1, no. 4, p. e36, 2018.

[16] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273–297, 1995.

[17] R. Shouval, O. Bondi, H. Mishan, A. Shimoni, R. Unger, and A. Nagler, "Application of machine learning algorithms for clinical predictive modeling: a data-mining approach in sct," Bone marrow transplantation, vol. 49, no. 3, p. 332, 2014.