DECISION TREE CLASSIFIER BASED INTRUSION DETECTION SYSTEM IN NETWORK TRAFFIC

A. Sunitha nandhini[1], P. Sudarsana rubine[2], P. Sri saranya[3]
[1]Assistant Professor,[2]student of Computer Science dept.,[3]student of Computer Science dept.
Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India

Abstract: A few examinations propose that by choosing pertinent highlights for intrusion detection system, it is conceivable to significantly enhance the detection accuracy and execution of the detection engine. With the development of new advances, for example, remote system or Big Data, the vast measure of network traffic is generated. The intrusion detection system powerfully gathers and breaks down the information created by the approaching movement. In any case, in an expansive dataset, not all highlights add to speak to the activity, along these lines diminishing and choosing various satisfactory highlights may enhance the speed and accuracy of the intrusion detection system. The proposed framework is decision tree classifier utilized for detecting the IDS assault Packets. KDD 1999 Dataset, scikit-learn is used, which is a machine learning library written in python. ANOVA-F test is used to calculate the variance of the attacks. These Algorithms are used to find the attacks that have caused security issues in the datasets during the network traffic. These outcomes loan to help highlights choice enhance the classifier execution.

Keywords: *scikit-learn, intrusion detection, one-hot-encoding, machine learning, the classifier*

I.INTRODUCTION

Numerous associations and organizations utilize Internet administrations to deal the items and commercial center to work together, for example, E-Bay and Amazon.com site. Together with the development of PC arrange exercises to affecting the accessibility, secrecy, and trustworthiness of basic data information. A system framework should utilize security instruments like a firewall, antivirus, IDS and Honey Pot for keeping the critical information from criminal undertakings. A firewall just isn't sufficient to keep systems from all assault composes and can't barrier the system against interruption endeavors amid the opening port. Henceforth a Real-Time Intrusion Detection System (RT-IDS),.a avoidance device that gives a caution flag to the PC client or system chairman for opposing action on the opening session, by reviewing perilous system exercises.

1.1 *Intrusion Detection System*

Regularly Intrusion Detection (ID) was directed physically by framework organization and entrusted with completely observing every movement on a reassure recognizing any oddities (Zhihua Zhang et al.,2017). An early type of ID demonstrated inadequate route because of the blunders it delivered. Mechanized log record per user was created to permit snappy look process for anomalies and unapproved workforce. The presentation of review logs helped to show ID into a measurable procedure; whereby organization ordered data and recognized the issues after episodes had just happened and not amid the procedure of an assault are occurred. Before the 90s‟ Intrusion recognition was a type of post-examination, at that point investigation of interruptions and changes in framework structure were just recognized long after the real occasion. Procedures are monotonous, moderate tedious and introduced capability of human blunders because of substantial contribution. Analysts built up an IDS that looked into review information and gave a headway generate to the principal form of continuous IDSs‟ taking into account assault pre-emption through strategies for ongoing reaction. As the world entered the mechanical age, the market interest for IT security expanded and IDS were additionally created and made accessible to numerous associations and organizations. New highlights were imagined like another ready strategy, updates to assault design definitions, committed easily to understand interfaces and anticipation systems that naturally halted assaults when it is recognized.

A number of information mining methods can be utilized as a part of interruption identification, however, each with its own particular leeway. The accompanying records a portion of the systems and the thought processes. An arrangement is which creates a grouping of tuples and used to distinguish singular assaults however it will deliver a high false alert rate. The happens of the issue might be decreased by applying tweaking systems, for example, boosting. Affiliation describes the connections inside tuples and discovery of abnormalities may happen when numerous tuples show already concealed connections.

Groups tuples that display comparable properties as indicated by pre-depicted measurements and furthermore utilized for the general investigation like order, or for identifying exceptions that might possibly speak to assaults are gathering.

1.2 Application of data mining in Intrusion Detection System

The objective of interruption discovery to identify security infringement in data frameworks (Huaifeng Zhang et al.,2014). Interruption discovery is a way to deal with security as it screens the data frameworks and raises alerts when security infringement are distinguished. A few cases of security infringement incorporate the manhandle of benefits or the utilization of assaults to misuse programming or convention vulnerabilities. Customarily, interruption recognition is characterized by two principal classes: Misuse identification and Anomaly location. Abuse identification works via hunting down the follows or examples of understood attacks. Only known assaults that leave trademark follows can be recognized. Oddity recognition utilizes a model of ordinary client or framework conduct and banners noteworthy deviations from the model as possibly malevolent. Model of typical client or framework conduct is normally known as the client or framework profile. A quality of oddity location is its capacity to identify beforehand obscure assaults.

Furthermore, interruption discovery frameworks as in the Fig.1.2 (IDSs) are ordered by the sort of information data. It will prompt the qualification between have based and arrange based IDSs. Host-based IDSs investigate have bound review sources, for example, working framework review trails, framework logs and application logs. A system based IDSs dissect organize bundles are caught on a system (Yogita B. Bhavasar et al.,2013). They use the learning about clients verifiable in an information stockroom to decrease costs and enhance the estimation of client connections. Associations would now be able to center around the most essential (gainful) clients and prospects and configuration focused on promoting systems to best reach.

Figure 1.2 Intrusion Detection Environment

II. INTRUSION DETECTION SYSTEM METHODS

Mark based location is the way toward contrasting marks or examples of known assault with the occasions to recognize the conceivable incidents. Most regular type of mark based IDS utilized monetarily indicates each example of occasions that relates to an assault as a different mark. Anomaly-Based discovery analyzes meanings of what movement is viewed as ordinary against the occasions to distinguish noteworthy deviations. Abnormality based IDS utilizes profiles that speak to the ordinary conduct of framework, applications or system movement that are produced by the attributes of run of the mill action over timeframe .K-means is a parceling strategy in bunching procedure of information mining and grouping technique is utilized to segment the preparation information into k groups with the assistance of Euclidean separation similitude. A calculation to gathering or to order the articles in view of characteristics and highlights into k number of groups (Nadiammai et al.,2014). Fundamental strides for grouping the information by k-means are:

1.Select a number (k) of bunch focuses - centroids (arbitrary)

2.Assign each question its closest group focus (e.g. utilizing Euclidean separation)

3.Move each group focus to the mean of its relegated objects

4.Repeat stages 2,3 until union (change in group assignments not as much as an edge)

Classification tree analysis is utilized to distinguish the "class". Relapse tree examination is that the information is persistent and the tree is utilized to anticipate its esteem. The term Classification and Regression Tree (CART) investigation is utilized to elude both of the above methodologies. Arrangement and relapse trees are machine-learning techniques for building expectation models from the information. The Classification and Regression Trees (CART) approach is actually called as double recursive partitioning. The process is paired in light of the fact that parent hubs are constantly part into precisely two kid hubs and recursive on the grounds that the procedure is rehashed by regarding every tyke hub as a solitary parent. The key components of CART examination are an arrangement of guidelines for part every hub in a tree; choosing the tree is finished and appointing a class result to every terminal hub.

III. SOFTWARE REVIEW

3.1 Python Overview

Python is a dynamic, deciphered (bytecode-arranged) dialect. There are no sort statements of factors, parameters, capacities, or techniques in the source code, which makes the code short and adaptable and the incorporate time composes checking of the source code is lost. Python tracks the kinds of all esteems at runtime and hail code that does not bode well as it runs. Python is a programming dialect. It takes the message that is composed (more often than not alluded to as code), transforms it into guidelines for the PC, and runs directions.

For all intents and purposes, Python is simply one more program on PC. The principal thing to learn is the way to utilize and collaborate with it. There are in reality numerous approaches to do: the first to learn is to cooperate with python's translator, utilizing working framework's (OS) support. Given that Python encourages distinctive ways to deal with composing code, a legitimate follow-up question is: what is an alternate method to compose code? While there are a few responses to this scrutinize, the most well-known elective style of composing code is called useful programming. Practical programming gets its name from composing capacities which gives the fundamental wellspring of rationale in a program. Python is a programming dialect, which is utilized to guide a PC. Complete a hunt to make sense of the association and a sort of electronic mail called spam.

Python is less entangled than some other well-known content-based programming dialects like Java or C++. However, Python is sufficiently capable that it is utilized by organizations, for example, Google, Yahoo, Red Hat, and then some. Numerous colleges utilize Python in their first processing course.

3.2 Jupyter Notebook

A "notebook" or "notebook archives" signify reports that contain both code and rich content components, for example, figures, joins conditions. In view of the blend of code and content components, the reports are the perfect place to unite an examination portrayal and its outcomes and they can be executed and play out the information investigation continuously. A comfort (or 'terminal', or 'charge incite') is a printed approach to collaborate with OS, similarly as the 'work area', in conjunction with the mouse, is the graphical method to associate your framework. As a server-customer application, the Jupyter Notebook App permits to alter and run notebooks by means of a web program. The application can be executed on a PC without Internet access or it can be introduced on a remote server, where it can be gotten to through the Internet.

Jupyter has a lovely notebook that composes, executes the code, examine information, insert substance and offer reproducible work. Jupyter Notebook (beforehand alluded to as IPython Notebook) permits to effortlessly share code, information, plots, and clarification in a single journal. Distributing is adaptable and it can be as PDF, HTML, ipynb, dashboards, slides, and that's only the tip of the iceberg. Code cells depend on an info and yield organize.

IV. METHODOLOGY

Accomplishing high expectation exactness in distinguishing irregularities in organize movement is a noteworthy objective in outlining machine learning calculations and in building Intrusion Detection Systems. Refusal of Service (DoS) assault class that contains different sorts of assaults, for example, Smurf, Teardrop, Land, Back, and Neptune (Swati Paliwal et al.,2012). The proposed arrangement calculation utilizing choice trees are unit parcels structures that speak to choice sets. These decisions produce decides that are utilized to characterize information. To accomplish the target, a recursive component disposal process is utilized and connected with a choice tree-based classifier and the reasonable significant highlights are distinguished.

The approach is associated with the NSL-KDD dataset which is an improved version of the past KDD 1999 Dataset, scikit-realize which is a machine learning library written in python.The tests result on the NSL-KDD informational collection, which accomplished a high precision.

4.1 Scikit-Learn

Scikit-learn (once in the past scikit learn) is a free programming machine learning library for the Python programming dialect. It has highlights different order, relapse and grouping calculations including bolster vector machines, arbitrary woods, inclination boosting, k-means and DBSCAN, and is intended to interoperate with the Python numerical and logical libraries NumPy and SciPy.

The library is based upon the SciPy (Scientific Python) that must be introduced before utilizing scikit-learn. The stack that incorporates:

1.NumPy: Base n-dimensional exhibit bundle

2.SciPy: Fundamental library for logical figuring

3.Matplotlib: Comprehensive 2D/3D plotting

4.IPython: Enhanced intuitive reassure

5.Sympy: Symbolic arithmetic

6.Pandas: Data structures and examination

The pressures or modules for SciPy mind traditionally named scikit. The module gives learning calculations and is named scikit-learn

4.2 Machine Learning in Python

Machine learning is a branch in software engineering that reviews the plan of calculations that can learn. Normal undertakings are idea learning, workplace learning or "prescient demonstrating", bunching and finding prescient examples. These errands are found out through accessible information that was seen through encounters or directions. The expectation that accompanies the teach is that including the experience into its undertakings will, in the end, enhance the learning. In any case, the change needs to occur such that the learning itself ends up programmed with the goal that people don't have to meddle any longer is a definitive objective.

1.Learn how to utilize Python and its libraries to investigate your information with the assistance of matplotlib and one hot Encoding.

2.Preprocess the information with standardization and split your information into preparing and test sets.

3.Next, work with the notable one-of-K calculation to develop an unsupervised model, fit the model to the information, foresee values, and approve the model that have been manufactured.

4.Use choice tree characterization to build another model to group the current information.

V. IMPLEMENTATION

5.1 Data Preprocessing

All highlights are made numerical utilizing one-Hot-encoding. The highlights are scaled to stay away from highlights with expansive esteem that may weigh excessively in the outcomes. One-Hot-Encoding (one-of-K) is utilized to change every single clear-cut component into twofold highlights. Prerequisite for One-Hot-encoding: "The contribution to the transformer ought to be a lattice of whole numbers, signifying the qualities gone up against by straight out (discrete) highlights. The yield will be a scanty lattice where every segment relates to one conceivable estimation of one component. It is accepted that info highlights go up against values in the range[0, values)."

In this way, the highlights first should be changed with Label Encoder, to change each class to a number.

```
for col_name in df.columns:
    if df[col_name].dtypes == 'object' :
        unique_cat = len(df[col_name].unique())
```

print("Feature'{col_name}'has{unique_cat} categories".format(col_name=col_name,unique_cat=unique_cat))

it is equally circulated and in this way it is expected to make fakers for all.

print()

print('Distribution of categories in service:')

print(df['service'].value_counts().sort_values(ascending=False).head())

```
Training set:
Feature 'protocol_type' has 3 categories
Feature 'service' has 70 categories
Feature 'flag' has 11 categories
Feature 'label' has 23 categories

Distribution of categories in service:
http         40338
private      21853
domain_u      9043
smtp          7313
ftp_data      6860
Name: service, dtype: int64
```

5.2 Feature Selection

Highlight Selection utilizing ANOVA F-test wipes out excess and unessential information by choosing a subset of significant highlights that completely speaks to the given issue. Univariate highlight determination with ANOVA F-test. Analyzes each element exclusively to decide the quality of the connection between the component and marks as shown in the Fig.5.2. Utilizing Second Percentile strategy (sklearn.feature_selection) to choose highlights in light of percentile of the most noteworthy scores. At the point when the subset is found, Recursive Feature Elimination (RFE) is connected. Highlight determination in view of one-way ANOVA F-test insights plot is connected to decide the most imperative highlights. The component determination in view of one-way ANOVA F-test is utilized to decrease the high information dimensionality of the element space before the grouping procedure.

	protocol_type	service	flag
0	tcp	ftp_data	SF
1	udp	other	SF
2	tcp	private	S0
3	tcp	http	SF
4	tcp	http	SF

Figure 5.2 Feature Selection

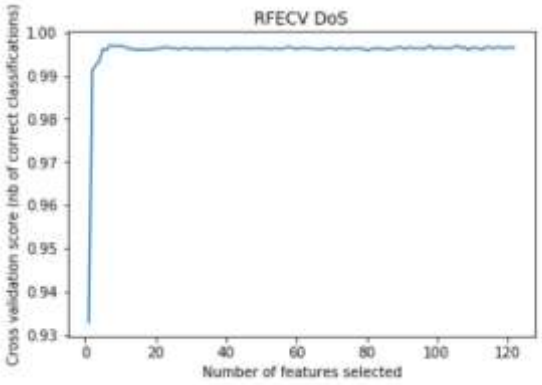

Figure 5.3 RFECV Dos

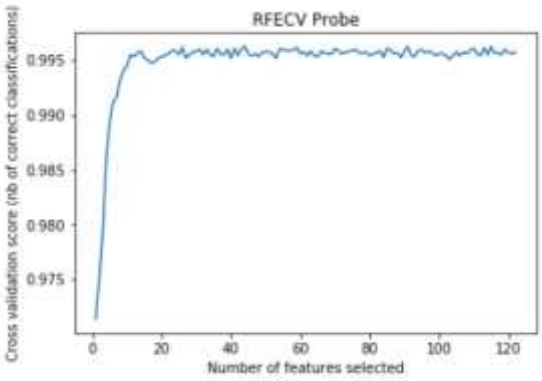

Figure 5.3 RFECV Probe

5.3 Decision Tree Classification

The estimations of the credits can be utilized to arrange the information things of the choice tree. The pre-ordered information is being utilized to develop a Decision tree. The information things can be isolated from classes and are parceled.

The procedure proceeds over and over for every subset and when every one of the information has a place with a similar class, the procedure closes. The specificity of a character is meant by a hub of a choice tree. Each hub has edges and is in the long run named according to the estimation of the trait in parent hub. A leaf or a hub is associated with an edge. The polluting influence of the information things is estimated by the idea of entropy utilizing data hypothesis. At the point when every one of the information things has a place with one class, the estimation of entropy is littler. Then again, the estimation of entropy is higher when the information things have more classes.

Occasionally look over logs or movement catches and report distinguished high need malevolent activity. In the wake of playing out the choice tree investigation, a prescient model that can separate distinctive sorts of movement. For a model prepared to distinguish amiable and malevolent movement

DoS

```
# Apply the classifier we trained to the test data (which it has never seen before)
clf_DoS.predict(X_DoS_test)
```

Predicted attacks	0	1
Actual attacks		
0	9499	212
1	2830	4630

Probe

```
Y_Probe_pred=clf_Probe.predict(X_Probe_test)
# Create confusion matrix
pd.crosstab(Y_Probe_test, Y_Probe_pred, rownames=['Actual attacks'], colnames=['Predicted attacks'])
```

Predicted attacks	0	2
Actual attacks		
0	2337	7374
2	212	2209

Figure 5.3 Calculation of False alarm

DoS

```
from sklearn.model_selection import cross_val_score
from sklearn import metrics
accuracy = cross_val_score(clf_DoS, X_DoS_test, Y_DoS_test, cv=10, scoring='accuracy')
print("Accuracy: %0.5f (+/- %0.5f)" % (accuracy.mean(), accuracy.std() * 2))
precision = cross_val_score(clf_DoS, X_DoS_test, Y_DoS_test, cv=10, scoring='precision')
print("Precision: %0.5f (+/- %0.5f)" % (precision.mean(), precision.std() * 2))
recall = cross_val_score(clf_DoS, X_DoS_test, Y_DoS_test, cv=10, scoring='recall')
print("Recall: %0.5f (+/- %0.5f)" % (recall.mean(), recall.std() * 2))
f = cross_val_score(clf_DoS, X_DoS_test, Y_DoS_test, cv=10, scoring='f1')
print("F-measure: %0.5f (+/- %0.5f)" % (f.mean(), f.std() * 2))
```

```
Accuracy: 0.99639 (+/- 0.00341)
Precision: 0.99505 (+/- 0.00477)
Recall: 0.99665 (+/- 0.00483)
F-measure: 0.99585 (+/- 0.00392)
```

Figure5.3 Accuracy, Precision, Recall, F-Measure

VI. CONCLUSION

In network traffic, interruption location assaults are distinguished which is rising security issue and it is conceivable to impressively enhance the identification precision and execution of the recognition engine. With the development of new advances, for example, remote system or Big Data, a substantial measure of system activity is created and the interruption discovery framework should progressively have gathered and dissected the information delivery by the approaching activity. The proposed framework choice tree classifier is utilized for recognizing the IDS assault Packets. KDD 1999 Dataset, scikit-discover that is a machine learning library written in python.ANOVA-F tests are utilized to ascertain the difference in the assaults identified. These outcomes loan to distinguish the interruption location of most elevated need and watches the genuine positive, genuine negative, false positive and false negative. Highlights determination enhances the classifier execution of every interruption location assaults.

VII.REFERENCES

[1] Discovering Informative Knowledge in Complex Data by Longbing Cao, Senior Member, IEEE, Huaifeng Zhang, Member, IEEE, Yanchang Zhao, Member, IEEE, Dan Luo, and Chengqi Zhang, Senior Member,

[2] Denial-of-Service, Probing & Remote to User (R2L) Attack Detection using Genetic Algorithm by Swati Paliwal, Assistant Professor, Dept. of C.S.E., Sharda University, Gr.Noida, Ravindra Gupta, Assistant Professor Dept. of C.S.E., ASSIST, Sehore, India in International Journal of Computer Applications (0975 – 8887) Volume 60– No.19, December 2012

[3] Effective approach toward Intrusion Detection System using data mining techniques by G.V. Nadiammai, M. Hemalatha in Egyptian Informatics Journal (2014) 15, 37-50

[4] Ham Kember. Data mining concepts and techniques. Studying various algorithms.

[5] Intrusion Detection System Using Data Mining Technique: SupportVector Machine by Yogita B. Bhavsar, Kalyani C.Waghmare in International Journal of Emerging Technology and Advanced Engineering Website: ww.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 3, March 2013)

[6] I. Onat and A. Miri, ``An intrusion detection system for wireless sensor networks," in Proc. IEEE Int. Conf. Wireless Mobile Comput. Netw. Com- mun., Aug., pp. 253259 (2005)

[7] Network Intrusion Detection Using Improved Decision Tree Algorithm by K.V.R. Swamy, K.S. Vijaya Lakshmi in (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (5), 2012,4971 – 4975

[8] Error-Aware Data Mining by Xindong Wu and Xingquan Zhu in IEEE transactions on systems, man, and cybernetics—part a: systems and humans, vol. 38, no. 4, July 2008

[9] Zhihua Zhang, HongliagZhu, Shoushan Luo, and Xiaoming Liu," Intrusion Detection Based on State Context and Hierarchical Trust in Wireless Sensor Networks," IEEE Translations and content mining, April 23,(2017)