International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 5 Issue 1, November-December 2020 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Comparative Study on Machine Learning Algorithms for Network Intrusion Detection System

Priya N, Ishita Popli

Masters in Computer Applications, Jain (Deemed-to-be) University, Bangalore, Karnataka, India

ABSTRACT

Network has brought convenience to the earth by permitting versatile transformation of information, however it conjointly exposes a high range of vulnerabilities. A Network Intrusion Detection System helps network directors and system to view network security violation in their organizations. Characteristic unknown and new attacks are one of the leading challenges in Intrusion Detection System researches. Deep learning that a subfield of machine learning cares with algorithms that are supported the structure and performance of brain known as artificial neural networks. The improvement in such learning algorithms would increase the probability of IDS and the detection rate of unknown attacks. Throughout, we have a tendency to suggest a deep learning approach to implement increased IDS and associate degree economical.

KEYWORD: *Intrusion Detection System, Machine Learning, NIDS*

How to cite this paper: Priya N | Ishita Popli "Comparative Study on Machine Learning Algorithms for Network Intrusion Detection System" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-5 | Issue-1, December 2020, pp.1212-1215, URL: www.ijtsrd.com/papers/ijtsrd38175.pdf

IJTSRD38175

INTRODUCTION

While technology is growing by day to detect successful attacks on networks and to elegantly detect intrusions, hackers use complex attack patterns to hack a network. The self and adoptive brain of the neural net and the intrusion prevention mechanisms are typically used to improve the detection capabilities. In the idea of signature, anomaly, host, and network, IDS are classification. The intrusion detection system attached to the signature draws patterns that could be matched by the information pattern.

The two problems are false positives and false negatives, since these IDS cannot accurately see all risks. Intrusion detection system primarily based on the network and host is often used within the network to understand the approach of users to avoid intrusions. It usually raises an alert while the IDS is watching a current attack between the computer device. The assaults are carried out by criminals in order to steal data from a repository or harass a financial institution or workplace.

IDS tracks a network or malicious activity system which defends a network from unwanted access from, even maybe, insiders. IDS tracks and manages a network or malicious activity system. The learning role of the intrusion detector is to construct a statistical model (i.e. a classification model) that can differentiate between 'weak relations.' Attacks fall under four main categories:

- DOS: denial, eg. syn flood; denial of service;
- R2L: illegal foreign-machine entry, e.g. password devaluation;
- U2R: unauthorized entry, such as multiple attacks of a buffer overload, to local super user (root) rights.
- Inspection: tracking and extra testing, i.e. port screening.

In the history of knowledge science and machine learning, privacy and security are major challenges. For example, a day where we do business, check questions, view images, toggle through different social media platforms; any channel which is stored and measured every day receives excellent information. This personal data is used in different machine learning systems to create a smoother navigation experience for the user.

For diverse applications, deep learning is used. Certain apps need personal data such as browsing history, business data, location, cookies, etc. This personal data is uploaded into ML algorithms in pure script form for the retrieval of the patterns and the creation of models. This is not just a case about risks associated with private data that are vulnerable to insider attack or external threats that indicate that the organizations that possess these data sets get compromised.

Machine learning supported software requires private data processed by servers. Private data such as corporate attacks, illicit transfers and unauthorized access etc. also being exploited by attackers for malicious purposes. Therefore, we built a data security forum to safeguard the privacy of the various cryptographic methods of information owners.

Related Work

Literature Review

Network intrusion detection systems (NIDS) are located within the network at a strategic location to track traffic from and to all users on the network. It analyzes the passage of traffic over the entire subnet and matches the traffic that has been transmitted on the subnet to the known attack library.

When an aggression has been detected or an odd behaviour, the warning is always sent to the boss. An instance of an NIDS will be to install it in the firewall subnet, to figure out whether anyone wants to disturb it. Ideally, all incoming and outgoing traffic will be screened, although this will create a bottleneck that may impact the overall network capacity.

NIDS is most frequently paired with other identification and prediction rates technologies. Artificial Neural Network-based IDS are able, due to its self-structure, to help identify intrusion patterns in an insightful manner, to evaluate enormous volumes of information. Neural networks allow IDS to anticipate attacks by error learning. IDS support the implementation of an early warning system, two levels are supported.

The first layer takes single values, while the second layer takes the output of the first layer as the input; the loop repeats and enables the device to detect new patterns within the network automatically. This technique supports 24 network attack outcomes, categorized into four categories: DOS, Study, Remote-to-Local, and User-to-root.

ML Algorithms

We apply different machine learning algorithms:

1. Guassian Naive Bayes
2. Decision Tree
3. Random Forest
4. Support Vector Machine
5. Logistic Regression

Deep learning algorithms for various classifying and predictive problems are commonly used and have yielded reliable results. A variety of ML algorithms are listed in particular.

1. Naive Bay of Guassian:

Naive Bayes often expand to real-time attributes by assuming natural distribution. Gaussian Naive Bayes is considered this extension of Bayes. The knowledge distribution is also determined by other functions, but Gaussian (or normal distribution) is the simplest to work out with since you can easily predict the mean and ultimately the variance from your training results.

When dealing with ongoing results, it is always believed that the continuous values for each class are distributed according to the Gaussian distribution. This model is also fit by simply looking for the mean and variation of the points on each mark, all required to describe such a distribution.

2. Decision Tree:

Decision tree may be a basic structure-like tree; model chooses each node. The root-to-leaf direction symbolizes rules for classification. One must chose at any node on which direction to drive into a leaf node. This call at each node depends on the trainings set's features/columns or the info information. It is useful for basic tasks in which one can understand when to categorize items by easy reasoning. It is very easy to illustrate to clients and straightforwardly to show how an election method functions as one of the most common examples of its ease and use.

3. Random Forest:

Random forests are a community method of regression, classification and other training tasks, created by the development of the decision-making trees and the performance of the class mode or average prediction of individual trees. Random decision forests are right to over fit their preparation habit for decision trees.

4. Support Vector Machine:

Support vector models with associated learning algorithms that analyze the data used for classification and regression are supervised in machine learning. An SVM model can also display the examples-points inside the space mapped to separate the samples of the various groups by a transparent variance. Help vectors are data points next to the hyper plane and influence the hyper plane's location and path.

5. Logistic Regression:

Logistic regression may be a statistical model which, even if there are more complicated extensions, uses a Logistic function to design the simple binary variable. Logistic regression calculates the parameters of a logistic model in a multivariate analysis (a sort of binary regression). The logistic regression model itself merely predicts the input likelihood of output and does not statistics.

Comparative Analysis

Features	Guassian Naive Bayes	Decision Tree	Random Forest	Support Vector Machine	Logistic Regression
Developed	Reverend Thomas Bayes	J. Ross Quinlan in 1975	Breiman in 2001	Vapnik in 1995	statistician DR Cox in 1958
Definition	Naive Bayes are also generalized by assuming a Gaussian distribution of practical attributes.	Decision Tree can be a simple structure tree and models select any node.	Random forests that produce several decision-making trees during preparation.	Supporting vector machines are supervised learning models that analyze knowledge for classification and regression using similar learning algorithms.	A mathematical model using a logistic function may be a logistic regression to design the binary variable in its fundamental form.

Applications	*News classification, identification of e-mail spam, facial recognition, medical diagnosis and forecast of weather.*	*Data mining is widely used to build algorithms for a prediction attribute.*	In finance, it is used to classify faithful clients and illnesses.	*Text and image classification, face identification and classification.*	*The risk of an occurrence is expected.*
Strength	1. *The test data set is simple and fast to estimate.* *2. It is strong in multi-class forecasting.*	1. Simple to comprehend, to interpret, to screen. 2. Performs easily and rapidly on massive datasets.	1. *The probability of excess fitting is lower.* *2. Less variation in the use of many trees.*	1. *SVM is able to model non-linear limits of judgment.* *2. They are resistant against overfit.*	1. *Outputs provide a clear understanding of probabilities.* *2. In logistic models, new data is quickly modified.*
Weakness	*Also referred to as the "zero frequency."*	Decision-tree students can produce unnecessarily complex trees that do not generalize info well.	*It is complex and difficult to implement.*	*It doesn't balance bigger datasets well.*	*They are not sufficiently scalable to capture more complex links.*
Strong Area	1. Text Data 2. Word- based Classification	1. Classification 2. Regression 3.Complex Non-Linear Classification	1. Complex Non-Linear Classification 2. Continuous Values	1. Complex Non- Linear Classification 2. Multi- Class Classification	1. Linear Model 2. Binary Classification
Core Idea	1. Bayes Theorem 2. Conditional Probability	1. Entropy 2. Cross- Entropy	1. Ensemble Learning 2. Weak Learner and Strong Learner	1. Kernel Methods 2. Margin Maximization	1. Event Occurs Probability 2. Odds Ratio
Computational Time	Less	Less	More	More	Less
Hyper parameters	No Hyperparameter	The decision tree contains the next tree node selection, maximum depth and minimum leaf of samples.	It comprises the number of functions that each tree considers when dividing a node.	The kernel is the SVM hyperparameter.	Learning pace must be balanced to ensure high precision.
Accuracy	81.69 %	99 %	92.49 %	90.6%	Depends on correct predictions.
Efficient	Most Efficient	Most Efficient	Less Efficient	It is memory efficient.	Less Efficient
Advantages	1. It needs a limited number of test results. 2. It's quick to introduce Naive Bayes too.	1. No data preprocessing required. 2. It provides understandable explanation over the prediction.	1. It reduces overfit and improves precision. 2 It dynamically handles the missed data values.	1. SVM model is stable. 2. Both classification and regression problems can be overcome by SVM.	1. Easy, fast and simple classification method. 2. Can be used for classifying multiclasses.
Implementations	Python/R	Python/R	Python/R	Python/R	Python/R

Conclusion

The detection of Network Intrusions in the area of network defense is highly significant. Although lots of research has been done in recent years in network intrusion detection, in particular multi-class network intrusion detection there is little or no in-depth investigation. A study of numerous Machine Learning models (Guassian Naive Bay, Decision Tree, Random Forest, Support Vector Machine and Logistic Regression) shows that the model Tree most closely conforms to our knowledge, given its precision and the complexity of time.

References

[1] A Deep Learning Approach for Intrusion Detection System in Industry Network, Ahmad HIJAZI, EL Abed EL SAFADI, Jean-Marie FLAUS.

[2] A Survey on Machine Learning and Deep Learning Methods for Intrusion Detection Systems, Hongyu Liu and Bo Lang.

[3] A Deep Learning Approach for Network Intrusion Detection System, Quamar Niyaz, Weiqing Sun, Ahmad Y Javaid, and Mansoor Alam.

[4] Research of Network Intrusion Detection Based on Convolutional Neural Network, Jianbiao Zhang

[5] Comparison of Naive Bayes, Decision Tree, Random Forest, Support Vector Machine and Logistic Regression Classifiers, Tomas Pranckevicius

[6] Model Application: How to compare and choose the best ML model, Akira Takezawa

[7] Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery, Phan Thanh Noi and Martin Kappas.

[8] Comparative Study on Classic Machine learning Algorithms, Danny Varghese

[9] Comparison of Machine Learning Classification Models for Credit Card Default Data, Vijaya Beeravalli.

[10] Text categorization with support vector machines: learning with many relevant features, Joachimss