International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Survey on different Machine Learning algorithms that are compatible with CSE-CIC IDS 2018 Dataset

Radhika S K[1], Ashwini S P[2], Dr. Jalesh Kumar[3]

[1,2]Assistant Professor, Computer Science Department, JNNCE, Shimoga
[3]Professor, Computer Science Department, JNNCE, Shimoga

Abstract
Cyber security is the concept of applying various technologies, procedures and protocols in order to safeguard the network, programs, systems, devices and each and every data that belongs to any person or an organization. Due to insane growth in the field of artificial intelligence, every individual is prone to cyber attacks. In this regard, many intrusion detection systems (IDS) have been developed with the aid of different machine learning algorithms. This paper mainly devotes on these machine learning models: Decision Trees, Naive Bayes, Gradient descent, support vector machine and Random Forest describing different potential threats represented with CSE-CIC-IDS2018 dataset. Multiclass classifications have been included to check which of the machine models will effectively identify and prevent intrusion into the network. This can be procured by ML supported IDS. Paper discusses about the merits and demerits of different ML models and suggests that conclusion result

Keywords: Cyber security, Machine Learning, CIC-IDS-2018, Multiclass classification, Decision Tree, Random Forest, Gradient Boost, SVM, KNN, Naïve Bayes

1. Introduction

The rapid growth in the field of internet of things and artificial intelligence has made impeccable impact on the growth of devices in a network pertaining to residential, industrial or urban surroundings. With increase in the number of devices, there increases a potential threat to the network. Many cases can be observed that are destructive and disturbing damages done to the society. The cyber attackers mostly targets and breaches the information related to finance, medicine and industry that can hugely impact the entire nation. Some of the common attacks include Denial of service (DoS) attacks, malwares and phishing attacks. In this context, providing security against unauthorized access, medication and deletion of private data is very much necessary. Cyber security includes many aspects like antivirus, firewalls, intrusion detection systems (IDS). In this regard, developing efficient and robust security mechanisms has become crucial. To detect all the cyber attacks and to prevent them, different machine learning approaches have been employed.

Machine learning is a wing of artificial intelligence that will allow the systems to learn new things from the training given to it. These systems will not be trained to act a certain way for every task, they will be trained for some of the tasks and the rest of the tasks will be managed by them on its own on the basis of training data. There are various classifications of machine learning algorithms. Some of the promising

types of machine learning algorithms are: Supervised machine learning, unsupervised machine learning and reinforcement learning.

Supervised learning works with labeled datasets. Both training and testing datasets will be labeled in this case. Here, both input and output will be labeled and these algorithms will draw the points between inputs and outputs of labeled data to match the result. The two categories of supervised learning are classification and regression. Classification simply involves grouping of different data into its corresponding groups based on the training given to it. Some of the classification algorithms include Decision Trees, Naive Bayes, support vector machine and Random Forest. Regression refers to predicting continuous data and providing numerical values to the predicted data.

Unsupervised learning will discover different behaviors, patterns and relationships in the unlabelled data to render the result. It does not require any algorithm or training as in case of supervised learning. This in turn will discover the solution on its own with the aid of hidden patterns and similarities. Further unsupervised learning has two categories: clustering and association. Reinforcement learning is a method where it learns by making interactions with its environment and finds the result. This method increases its performance by feedback mechanism. It learns from the feedback and takes its performance to next level at every point of time.

The dataset used for cybersecurity for this survey is the CSE-CICIDS2018, published by the Canadian Institute for Cybersecurity - Intrusion Detection System 2018. This dataset consists of seven multiclass classification attack scenarios: DoS, Brute-force, Botnet, DDoS, Heartbleed , infiltration and Web attacks. The attacking infrastructure includes the victim organization has 5 departments and includes 30 servers and 420 machines. The dataset includes the captures network traffic and system logs of each machine, along with 80 features extracted from the captured traffic. The whole dataset contains 16,000,000 instances records with labelled columns. The CSV version of CSE-CIC-IDS2018 is provided.

In this survey ML models are discussed - Decision tree, Naïve Bayes,, Gradient Boost, KNN, RF and SVM are applied for classification of intrusion detection.

Decision tree

The decision tree is a model represented in a tree like structure where each node denotes a decision based on a specific feature. It's used for both classification and regression. The structure of tree start with the root node as initial input, as each node represents as a feature, and outcome of the decision is by branches.

Naïve Bayes

Naïve Bayes used in the classification of text and spam filtering.it assumption of feature independence might not often hold in all real-world datasets, still it performs well and is efficient, especially with high dimensional data.

Gradient boost

The gradient boost is a learning method that involves combining multiple weak learners or models sequentially. It works for training a new model that predict the errors made by the previous models, and then to minimize the overall error will combine all error. To get optimized performance this will iterate until models predicate.

K-Nearest Neighbour (KNN)

The KNN algorithm comes in both under classification and regression tasks. In classification k-nearest neighbours is calculated where as in regression mean of k-nearest points of data as an output. For new instance testing KNN calculates the distance between the item to be classified and other training data items. The classification results assigned with majority class among the k neighbours. An optimal k value is considered.

Random Forest (RF)

The Random Forest is a supervised machine learning algorithm used in both classification and regression. A randomly chosen part of the original dataset is used for each tree in the Random forest. All combined prediction is noted to determine the result of classification. To enhance the accuracy more number of trees is used with mitigates over fitting issues.

Support Vector Machine Algorithm (SVM)

The SVM algorithm comes under supervised machine learning, applicable to both regression and classification. A hyperplane that separates two classes in a given dataset, using distance between points belonging to different classes and identifies the points closest to the line known as support vector. To maximize the margin between itself and the support vectors.

The following metrics are commonly used to evaluate the performance of classification models:

- **Accuracy:** this will calculate the precision of a classifier. The number positive prediction is determined by this. It is the ration of the number of predictions in true positive to the sum of number of predictions made by the model.
- **Precision:** the ratio of the number of positive predictions to the total number of positive values predicted by the classifier. The accuracy of positive predictions.
- **Recall:** the total number of true prediction divided by the sum of all samples belonging to the positive class.
- **F1 score:** to calculate the precision and recall simultaneously by generating harmonic mean of the two metrics using F1 scores.

2. Related work

In[1] authors **Khalil IBRAHIMI, Mohammed JOUHARI, Zineb JAKOUT** worked on CICIDS2018 dataset evaluate and compare the performance of four supervised ML models – Decision Tree, Random Forest, Naïve Bayes and Gradient Boost in the cyber security on the 10% of complete dataset.

Pre-processing steps followed by authors in [1]

- To streamline the dataset, first empty cell across the sample is removed.
- To maintain data integrity in the dataset missing values are replaced with appropriate median or mean of the attributes.
- The data types of some attributes are converter to context to maintain consistency in processing and analysis.

They worked on multi-classification context of dataset to distinguishing among several subcategories of attacks as list in the introduction section. The performance evaluation of each classifier is summarised as Decision Tree proved an accuracy of 0.97, precision of 0.99, Recall of 0.98 and F1-score of 0.99, for Naïve

Bayes proved an accuracy of 0.16, precision of 0.17, Recall of 0.67 and F1-score of 0.27, for Random Forest proved an accuracy of 0.98, precision of 0.99, Recall of 0.98 and F1-score of 0.99 and Gradient Boost proved an accuracy of 0.99, precision of 0.99, Recall of 0.98 and F1-score of 0.99. Authors of [1] showed a high performance metrics close to 99% for Decision Tree, Random Forest and Gradient Boost.

In [2] authors **Emmanuel Chinanu Uwazie, Morufu Olalere, comAfolayan A. Obiniyi and Perpetua N. Achi** worked on CICIDS2018 dataset , classifiers are examined and performance of the RF, KNN and SVM is compared in the cybersecurity to detect the different types of attacks.

Pre-processing steps followed by authors in [2] are

- Feature selection to identify the most relevant features from the datasets while reducing dimensionality. This leads to most informative attributes in dataset for leading to classification efficiency.
- For uniformity normalization method is applied to improve classification accuracy.

Considering the multiclassification context of dataset attacks three algorithms, RF, KNN and SVM algorithm were also used to predicat classes on the dataset. In dataset, KNN outperformed other algorithms an accuracy of 0.984220385, next closely by RF with an accuracy of 0.970445802 and lowest performance with SVM an accuracy of 0.96324545. The precision, Recall and F1-score are concluded according to the classification of attacks.

In [3] authors **Turukmane, Anil V., and Ramkumar Devendiran** Proposed a method detects intrusion on its own very swiftly by making use of hybrid machine learning algorithms. It involves a method called as mud ring assisted multilayer support vector machine in order to segregate the various potential cyber attacks. Its basic working models takes into account of CSE-CIC-IDS 2018 and UNSW-NB15 dataset. This dataset is fed to pre processing unit. Working stages are given as follows:

- **Preprocessing and normalization of data:** In this stage, all the null values are handled accordingly and even normalization process is performed. For the normalization step, min=max normalization procedure is carried out where any minimum value will be set to 0 and in the same fashion, any maximum value is reset to 1.
- **Balancing of data:** After preprocessing stage is finished, the next step is data balancing. To prevent biasing to any of the class, under sampling or oversampling can be performed. Here oversampling is performed in this study. Oversampling is making new instances of the minority class in order to balance the data.
- **Feature extraction:** Once the data is balanced, next step is to extract the potential features by a method called as modified singular value decomposition. In this step, a single value will be decomposed into three matrices for further processing.
- **Feature selection:** Further step is feature selection, which is accomplished by opposition based northern goshawk decomposition.
- **Applying Multi SVM:** After selecting prominent features, next step is applying hybrid machine learning algorithm called as mud ring assisted multilayer support vector machine. After applying this hybrid algorithm, cyber attacks will be classified.

More accurate result can be obtained after fine tuning. The proposed method has rendered accuracy of 99.89%, precision of 99.914%. Recall value was found to be 99.25% and F1 score was 99.214%.

3. Conclusion

A comparison of machine learning algorithms for the classification of intrusion detection using CIC-IDS-

2018 dataset were conducted in this survey paper. Three distinct papers are survived. A Multi-SVM out performance other algorithm with an accuracy of 99.914%, followed closely by Gradient Boost with an accuracy 99%, then RF and KNN with an accuracy of 98%, SVM with an accuracy of 96% and lowest performance Naïve Bayes with an accuracy of 16%. By reviewing recent works have been discussed above says that Multi-SVM yield an high accuracy of classification of intrusion detection using CIC-IDS-2018.

Reference

1. Ibrahimi, Khalil, Mohammed Jouhari, and Zineb Jakout. "Enhancing Intrusion Detection Systems Using Machine Learning Classifiers on the CSE-CIC-IDS2018 Dataset." 2024 11th International Conference on Wireless Networks and Mobile Communications (WINCOM). IEEE, 2024.
2. Uwazie, Emmanuel Chinanu, et al. "Comparison of Random Forest, K-Nearest Neighbor, and Support Vector Machine Classifiers for Intrusion Detection System." 2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG). IEEE, 2024.
3. Turukmane, Anil V., and Ramkumar Devendiran. "M-MultiSVM: An efficient feature selection assisted network intrusion detection system using machine learning." Computers & Security 137 (2024): 103587.