Network Intrusion Detection System Using Supervised Machine Learning

Karpe Akashay,Gunjal Aniket, Dhage Saurabh, Adhav Aniket
Prof. Rohini S Hanchate

**D Y Patil Institute of Engineering and Technology Ambi,
Maharashtra 410501.**

ABSTRACT

A singular supervised machine learning machine is developed to classify network traffic whether or now not it is malicious or benign. to search out the only version considering detection success rate, Combination of supervised learning algorithmic rule and have choice technique are used.

Through this study, it is that Artificial Neural Network (ANN) primarily based mainly system gaining knowledge of with wrapper function desire out plays help support vector machine (SVM) technique while classifying community site visitors. to decide the performance, NSL-KDD dataset is employed to categorise community site visitors exploitation SVM and ANN supervised machine gaining knowledge of techniques. Comparative observe indicates that the projected model is affordable than alternative existing models with relevancy intrusion detection success fee.

Keywords: *Network Intrusion Detection, KDD-99 Dataset, KNN, Support Vector Machine, Machine Learning, Naïve Bayes*

Literature Survey

Incremental Anomaly-based Intrusion Detection System Using Limited Labeled Data
Parisa Alaei, Fakhroddin Noorbehbahani –

The proposed model, called Network Anomaly Detection using Active Learning (NADAL) involves an offline and an online step. The selected dataset is preprocessed in an offline fashion. The NSL-KDD dataset contains instances labeled with the attack type. During the preprocessing step, the attacks are divided into four categories: DoS, Probe, R2L, and U2R. Furthermore, there are four classifiers at the respective layers of attacks. Thus, the preprocessing carried out using Weka selects the appropriate features for each classifier. The selected features are then given to the feature filtering module in NADAL. In the roposed online method, at each time, each instance is processed at most once to improve the model. The instance is then discarded. Initially, instance having label passes through the feature filtering module and the appropriate features for each classifier are considered. At each layer, the naive Bayesian module incrementally predicts the probability that the instance belongs to the class. Thereafter, the selected active learning strategy (i.e. uncertainty with randomization) is called. The output of the strategy determines whether the label for the instance must be inquired. A logical OR gate is used to aggregate the results from different active learning modules. The classifiers are updated using the instance if the gate outputs 1. Otherwise, the aggregate output module predicts the label according to the maximum certainty calculated by the classifiers. In this case, represents the actual label for instance.

1) An evaluation of machine learning algorithms To detect attacks in scada network –
Hicham Belhadaoui, Mahmoud Almostafa Rabbah, Sara Tamy, -

A Naïve Bayes could be a greatly simplified Bayesian likelihood model. The naïve Bayes classifier is predicated on a powerful assumption of independence. It means the likelihood of AN attribute doesn't have an effect on the likelihood of another [12]. The Naive Bayes classifier provides typically correct results, and it's proved effective in several sensible applications, specifically diagnosis, text classification, and systems performance management [13]. SVM was initial detected in 1992, introduced by Boser, Guyon, and Vapnik [14]. it's a strong methodology for resolution classification issues [15], The SVM could be a set of connected supervised learning strategies used for classification and regression . it's a classification and regression prediction methodology, that uses machine learning theory to optimize prognosticative accuracy whereas avoiding over-fitting to the information. The SVM use hypothesis house of a linear functions in a very high dimensional feature house, trained with a learning algorithmic rule from improvement theory that applies a learning bias derived from applied math learning theory. The SVM was at first fashionable the NIPS community and now's an energetic a part of the machine learning analysis round the world. we are able to additionally use it for varied applications, including, face analysis, hand writing analysis, and so on, notably for classification and regression applications [14].

Ross Quinlan developed C4.5 algorithmic rule that generates a choice Tree. J48 classifier could be a easy C4.5 call tree for classification. Open supply Java implementation of C4.5 unharness in rail data processing tool [15]. this can be a regular call Tree algorithmic rule. one in every of the classification algorithms in data processing is call Tree Induction. The Classification algorithmic rule is inductively learned to construct a model from the reclassified knowledge set. every knowledge item is outlined by values of

The options. Classification will be thought of as mapping from a group of options to a specific category. the choice tree approach is helpful for classification downside. With this method, a tree is made to model the classification method. Once the tree is made, it's applied to every tuple in

2) Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection - Kazi Abu Taher, Billal Mohammed Yasin Jisan, Md. Mahbubur Rahman.

To implement and analysis the system we've got used wide used open supply machine learning code suite referred to as rail. together with machine learning algorithmic rule enforced, rail conjointly has many algorithmic rule and search technique enforced to perform feature choice. within the ANN model, we have a tendency to experimented with totally different range of hidden layer and located that the detection success rate varies with the amount of hidden layer. when many trial and error ways, we have a tendency to found best detection rate with three hidden layers and zero.1 learning rate. within the wrapper feature choice methodology, we have a tendency to conjointly used SVM algorithmic rule as classifier. The model enforced in rail has been run on a computing platform having sixty four bit a pair of.6 Gc Intel core i5 processor with eight GB RAM on Windows seven setting with restricted network traffic instances. Implementing the answer on giant scale network would force further infrastructure with some higher capability server platform.

Introduction

With the wide spreading usages of web and will increase in access to on-line contents, law-breaking is additionally happening at AN increasing rate [1-2]. Intrusion detection is that the beginning to stop security attack. thus the protection solutions like Firewall, Intrusion Detection System (IDS), Unified Threat Modeling (UTM) and Intrusion bar System (IPS) have gotten a lot of attention in studies. IDS detects attacks from a range of systems and network sources by grouping data and so analyzes the data for attainable security breaches [3]. The network primarily based IDS analyzes the info packets that travel over a network and this analysis square measure meted out in 2 ways in which. until nowadays anomaly {based|based mostly|primarily primarily based} detection is way behind than the detection that works supported signature and thus anomaly based detection still remains a significant space for analysis [4-5]. The challenges with anomaly primarily based intrusion detection rectangular measure that it has to have an effect on novel attack that there's no previous records to spot the ambiguity. Hence the system one way or the other has to have the intelligence to segregate that traffic is innocent and that one is malicious or odd and for that machine gaining knowledge of strategies rectangular degree being explored via the researchers over the previous few years [6]. IDS but isn't always a approach to any or all security connected issues. as an instance, IDS cannot compensate vulnerable identification and authentication mechanisms or if there's a weak point in the community protocols.learning the sector of intrusion detection initial started in 1980 and therefore the initial such model was printed in 1987 [7]. For the previous couple of decades, although large industrial investments and substantial analysis were done, intrusion detection technology remains immature and thus not effective [7]. while community IDS that works supported signature have visible business success and good sized adoption through the era primarily based employer in the course of the world, anomaly based network IDS have not received achievement in the same scale. way to that motive within the field of IDS, currently anomaly primarily based detection may be a main recognition space of analysis and improvement [8]. And before approximately to any huge scale education of anomaly based intrusion detection device, key troubles stay to be resolved [8]. however the literature these days is constrained once it entails evaluate on but intrusion detection plays as soon asvictimisation supervised system getting to know techniques [9]. To shield target systems and networks against malicious activities anomaly-based totally community IDS can be a treasured generation. no matter the variety of anomaly based network intrusion detection techniques delineated inside the literature in current years [eight], anomaly detection functionalities enabled security tools rectangular measure sincerely beginning to seem, and some crucial troubles live to be resolved. many anomaly based strategies are deliberate in addition to simple regression, guide Vector Machines (SVM), Genetic algorithmic application, Gaussian mixture model, knearest neighbor algorithmic application, Naive Thomas Bayes classifier, name Tree [three,five]. among them the most wide used studying algorithmic application is SVM as it has already installed itself on differing styles of disadvantage [10]. One predominant trouble on anomaly primarily based observeion is even though of these deliberate techniques will stumble on novel assaults but they all go through a high cautioncharge typically. The purpose behind is that the complexness of generating profiles of realistic conventional conduct by using studying from the coaching facts units [11]. today synthetic Neural network (ANN) square measure typically trained via the rear propagation algorithmic program, that had been around given that 1970 because the reverse mode of automatic differentiation [12]. the primary demanding situations in evaluating performance of community IDS is that the inconvenience of a complete community based totallyinformation set [13]. maximum of the planned anomaly based totally strategies observed in the literature had been evaluated victimisation KDD CUP ninety nine dataset [14]. at some stage in this paper we have a tendency toused SVM and ANN –two system getting to know strategies, on NSLKDD [15] that is a stylish benchmark dataset for network intrusion. The promise and therefore the contribution device learning did untilnowadays square degree captivating. There square measure numerous truth packages we will be predisposed to rectangular measure victimisation nowadays presented via system getting to know. It looks that gadget studying can rule the globe in returning days.hence we have a propensity to got here out into a hypothesis that the project of function new attacks or zero day assaults dealing with through the generation enabled groups these days are often triumph over victimisation machine gaining knowledge of strategies. right here we will be predisposed to developed a supervised system learning model so one can classify unseen community traffic supported what's learnt from the seen traffic. we've got a tendency to used every SVM and ANN mastering algorithmic application to search out the only classifier with better accuracy and fulfillment charge. layout and Implementation Constraints The machine must be particularly reliable.

Design and Implementation Constraints

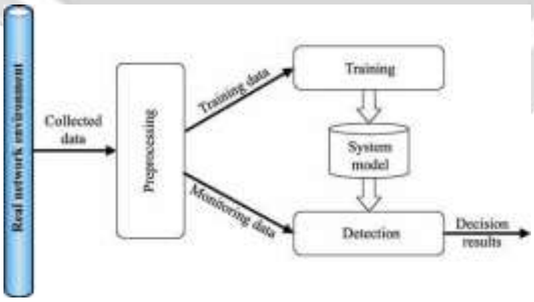
- The system should be highly reliable.
- The system should be secure.
- To get the system maintenance cost very less when compared to existing systems.

- To get instantaneous results with high accuracy.
- Retrieving of data from image and text database.
- Comparing of similar input using data mining algorithms.
- To calculate the most accurate prediction probability using text processing.

System Features

- Its use helps in prediction of attack having similar network pattern
- Ensure accurate attack prediction.
- Require less manpower
- Ensures low cost
- Easily accessible
- Ensure fast attack prediction

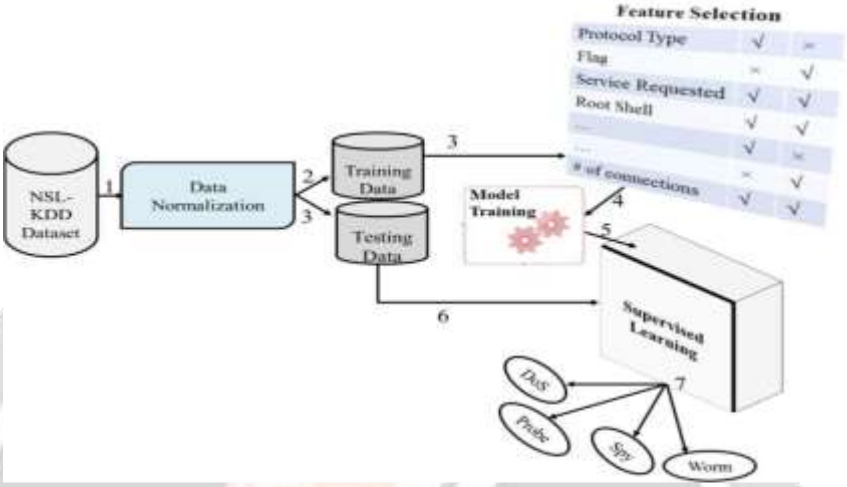
State Transition Diagram:



State Transition Diagram

System Architecture:

System Architecture



PREPROCESSING:

1. **Naive Bayes algorithm.**

It's miles a classification method based totally on Bayes' Theorem with an assumption of Independence among predictors. In simple phrases, a Naive Bayes classifier assumes that the presence of a selected feature in a class is unrelated to the presence of some other characteristic. Naive Bayes version is easy to construct and in particular useful for very massive statistics sets. at the side of simplicity, Naive Bayes is thought to outperform even tremendously sophisticated class strategies.

Working:

Step 1: Convert the data set right into a frequency table set.

Step 2: Create likehood table via finding the probabilities.

Step 3: Now, use Naive Bayesian equation to calculate the posterior opportunity for every elegance. The magnificence with the best posterior possibility is the outcome of prediction.

CLASSIFICATION:

2. **CNN AlGORITHM**

STEP 1: THE CONVOLUTION OPERATION

The convolution function is used to generate a feature map. In this process, each pixel, or section, of the image is reviewed

STEP 2: POOLING

We have our feature map, we need to get this into manageable data set.The algorithm will move across feature map, a few sections at a time and pool the data there. There are several ways that we can choose for your data to be pooled. For example, deciding to keep only the maximum value from the section the algorithm is looking at. Another option is to take the mean of all the data points in the section. The main thing is that once you have completed pooling the data

STEP 3: FLATTENING

This step is relatively self-explanatory. For your ANN to be able to take in your dataas an input, it needs to be flat. In this step you will take your pooled, and minimizeddataset, transforming it into a 1D vector. Here the algorithm you use for flattening will take each row of your ‘image pixel data and stack them on top of one another.

STEP 4: CONNECTION

The final step is where you connect the flattened out feature vector and feed it into your ANN as an input. This is important for running the algorithm efficiently.

Applications

Proposed system is used to detect network intrusion by using machine learning algorithm with better accuracy. Proposed system considering various factors like

- Accuracy
- Time
- Cost

To implement proposed system, we are using dataset from kaggle website

System Advantages

- Its use helps in prediction of network attack having similar pattern
- Ensure accurate attack prediction.
- Require less manpower
- Ensures lowcost
- Easily accessible
- Ensure fast attack prediction.

System Disadvantages

- The system must provide data integrity to insure data remains consistent andupdated.
- The system should provide means for protecting and
- securing the input provided by the user
- The system should provide the most accurate prediction of network afterclassi- fying and comparing it.

Future Scope

- Audio data can be analyzed
- Network security
- Video data can be analyzed
- More security related work.(Packet Analysis)
- Bank Sector to detect intrusion in transaction

Mathematical Modelling

Let ‘S’ be the system

- Where,
  - S= I, O, P, Fs, Ss Where,
  - I = Set of input Set of output
  - P = Set of technical processes
  - Fs = Set of Failure state
  - Ss = Set of Success state
- Identify the input data I1, I2
  In I = (Input Data (Text), Dataset (KDD))
- Identifytheoutputapplicationsas O1, O2,,On
  - Atrack Prediction

Identify the Process as P
P = (Data pre-processing, Data Processing, segmentation, featureextraction, classification, show result)
Identify the Failure state as Fs
Fs = (If data set not loaded, If not predicted, if more time required to predict
Identify the Success state as Ss P = (Correct prediction within time)

Conclusion –

Traditional knowledge packets square measure inherently static. In distinction, streaming knowledge square measure ceaselessly created; they can not be stored; and should by analyzed as one unit. during this paper, a completely unique network anomaly detection framework was planned to boost potency in classifying knowledge in a web fashion. moreover, active learning was wont to scale back labeling prices.

The planned system was evaluated exploitation the quality. during this paper, we've given totally {different|completely different}|completely different} machine learning models exploitation different machine learning algorithms and different feature choice strategies to seek out a best model.

The analysis of the result shows that the model designed exploitation ANN and wrapper feature choice outperformed all alternative models in classifying network traffic properly with detection rate of ninety four.02%. we have a tendency to believe that these findings can contribute to analysis more within the domain of building a observeion system that may detect famed attacks yet as novel attacks. The intrusion observeion system exist these days will solely detect famed attacks. detective work new attacks or zero day attack still remains a pursuit topic thanks to the high false positive rate of the prevailing systems.

Appendix A (Problem Statement Feasibility Assessment)

NP Hard(non-deterministic polynomial-time hard):

A problem is NP-hard if an algorithm for solving it can be translated into one for solving any NP problem. NP-hard therefore means” at least as hard as anyNP-problem,” although it might, in fact, be harder. Creating the system considering all the diseases makes this problem NP Hard. To do this we need to combine all the diseases. Also we need datasets consistingof all symptoms that are occurring to various humans. Another image dataset of almost all diseases will be very large. Therefore, for this situation the problem is NP Hard.

NP COMPLETE:

To make the problem NP Complete we need to reduce it to some extent. So,if we think that we will do this for all the diseases with similar symptoms then the problem will be NP Complete. But this is also not feasible to do and as we know wecannot solve NP problems.

P:

P-type problems are the problems that are solvable in polynomial time. To make NP Complete problem solvable we need to reduce it to P-type. So to reduce our problem to P-type, we can solve it we will consider only 2 diseases i.e malaria and dengue.

References

[1] H. Song, M. J. Lynch, and J. K. Cochran, "A macro-social exploratory analysis of the rate of interstate cyber-victimization," American Journal of Criminal Justice, vol. 41, no. 3, pp. 583–601, 2019.

[2] P. Alaei and F. Noorbehbahani, "Incremental anomaly-based intrusion detection system using limited labeled data," in Web Research (ICWR), 2017 3th International Conference on, 2019, pp. 178–184.

[3] M. Saber, S. Chadli, M. Emharraf, and I. El Farissi, "Modeling an implementation approach to evaluate the intrusion detection system," in International Conferenceon Networked Systems, 2018, pp. 513–517.

[4] M. Zamani and M. Movahedi, "Machine learning techniques for intrusion detection," arXiv preprint arXiv:1312.2177, 2018.

[5] N. Chakraborty, "Intrusion detection system and intrusion prevention system: A comparative study," International Journal of Computing and Business Research (IJCBR) ISSN (Online), pp. 2229–6166, 2018.

[6] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," computers & security, vol. 28, no. 1–2, pp. 18–28, 2018.

[7] M. C. Belavagi and B. Muniyal, "Performance evaluation of supervised machine learning algorithms for intrusion detection," Procedia Computer Science, vol. 89, pp. 117–123, 2018.