CMP417 Engineering Resilient Systems: Machine Learning

Mairi MacLeod
School of Design and Informatics
Abertay University
DUNDEE, DD1 1HG, UK

ABSTRACT

The organisation has asked the author to research a machine learning algorithm to act as an Intrusion Detection System (IDS) on their network. This paper discusses network security, machine learning as a whole as well as how these algorithms sort the data they are given.

The two algorithms discussed are K-means clustering, which is unsupervised and sorts the data into groups, and Decision Tree learning where data is recursively sorted until it is organised into very specific groups. In order to design an algorithm for this purpose the author suggested the following stages; deciding how data is input, how input is interpreted, what happens to the data in the algorithm and finally how the results are displayed. The evaluation metrics discussed are; the confusion matrix, AUC-ROC curve and the use of standardised testing data for intrusion detection systems.

1. INTRODUCTION

This paper is written to provide some network security recommendations to the company, who from here on will be referred to as FooBar Incorporated. The recommendations are specifically for a machine learning algorithm that will be used to classify malicious packets being sent in the network.

This is especially important to the company as they have recently been threatened with a cyber attack by a hacktivist organisation. Therefore any security that can be strengthened must be in order to prevent their customers or their own important data being stolen or destroyed.

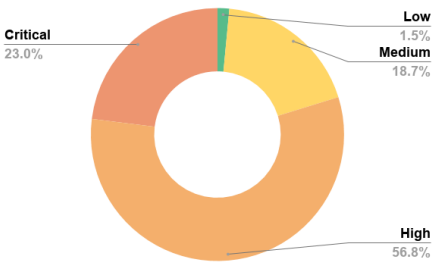

Figure 1: Percentage of Network Attacks by Severity (Mammen, 2020)

According to research done by PaloAlto in 2020 around 79.8% of network attacks worldwide can be categorised as either High or Critical *(Mammen, 2020).* These can consist of Man in the Middle (MitM), Denial of Service (DoS) or Trojan packets being used to deliver harmful malware to the organisations network. These attacks can be devastating to an organisation. They can result in customers' personal data being stolen, services such as web applications being DoS-ed and servers being shut down or wiped.

Due to the volume of network traffic FooBar inc. may receive on an hourly or daily basis having a human attempt to sift through the traffic and highlight any malicious packets is infeasible. An algorithm can provide constant analysis of network traffic, can be significantly more effective at flagging harmful data and will work considerably faster than simply using an employee.

2. BACKGROUND

NETWORK SECURITY

Network or system security is crucial to an organisation and can be defined as; the actions taken to prevent malicious attackers from gaining access or exploiting vulnerabilities on a network. These security measures are implemented by the system administrators of an organisation and can include a combination of physical and logical security.

These security types refer to preventing the type of attack that they attempt to prevent, physical security does not necessarily refer to a physical real-world security feature. Some of these techniques can be having a topology that allows for redundancies and isolates important devices behind a firewall or Demilitarised Zone (DMZ) to prevent ease of access by attackers *(Malik, 2003).*

MACHINE LEARNING

Machine Learning (ML) refers to software that can be trained to 'learn' from data and improve it's accuracy in identifying specifics in data sets without being specifically told to do so. These exist in neural networks which use a series of inputs and outputs that are weighted to determine patterns in data. The simplest of these are perceptrons,

a 'neuron' that determine whether an input can be categorised in specific classes.

These networks can also be called an Artificial Neural Network (ANN). By using perceptrons to generate layers of inputs and outputs values can be generated that correspond to the weighting of data in reference to how likely it is to belong to the data the algorithm is trying to match. An example would be determining if an image contains a face by combining the weighting of found features and calculating the score, which if it is high enough the algorithm can determine that there is indeed a face. The reason it is called machine 'learning' is due to the fact that the algorithm constantly adapts the weighting of data in order to make it more accurate with every run. In a sense it is learning how to determine patterns more accurately and why it is sometimes referred to as 'artificial intelligence'.

Deep Neural Network (DNN)s are a multi-layer ANN and can use multiple algorithms and formulae to generate an output. A Convolutional Neural Network (CNN) is a form of DNN that has a 'convolutional layer' that consists of a series of filters which is applied to the input as a whole and the filtered data is passed onto the next node until a recognised pattern is detected *(Figure 1).* The hidden layer contains the nodes that perform the calculations, which are all interconnected and communicate with each other *(Xiao et al, 2019).*

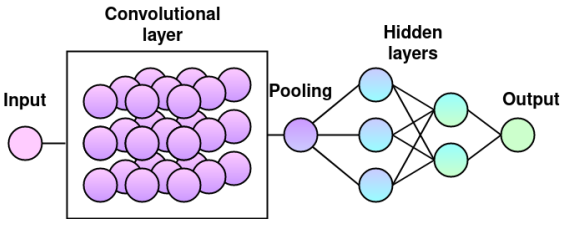

Figure 2: Diagram of a Convolutional Neural Network

CLASSIFICATION OF DATA

To determine whether data is relevant to what the algorithm is trying to match it must be classified. There are a number of ways that this can be done and they are; Binary, Multi-class, Multi-label and Imbalanced.

Binary classification is the most basic way that data can be organised by a ML algorithm. It sorts the data into two predefined categories, a practical example of this is in disease detection where it can either be present or not. Due to the simplicity of this method it is prone to false positives and negatives. As well as not being suitable for more complex problems, such as intrusion detection in a network.

Multi-class or multinomial classification is similar to binary but it has three or more categories to sort data into. Popular examples of this are in facial recognition algorithms such as Face++ or Amazon's Rekognition software. Multi-labelled classification is similar to multi-class but each piece of data can be assigned to multiple categories.

Imbalanced classification is where the data used to train the algorithm is not equally balanced across all of the categories. Resulting in weighting that may lean more towards certain classes of data.

K-MEANS CLUSTERING

One algorithm that can be used for creating an IDS in a network is K-means clustering. This is an unsupervised algorithm which means that it decides the classification of its inputs without the assistance of a 'response variable' that can be used to train the algorithm.

Clustering is when data is organised into groups, in this case the k clusters or groups are predetermined by a user *(Ahmad et al, 2020).* The most commonly used algorithm to use with this clustering technique is one developed in 1979 by a researcher called Hartigan-Wong, shown below.

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

This algorithm clusters data by calculating the data's distance from the centroids, the middle of the cluster. For an IDS it has been suggested that this algorithm can be used to create a multi-level system that passes data down to each level in order to precisely identify it *(Ahmad et al, 2020).*

A benefit of using this algorithm is that it is considered simple to implement and can easly be scaled up. This can be very important to FooBar inc. as in the future their network may increase in size and so they require an Intrusion Detection System (IDS) that will be able to upscale along with their requirements.

Some limitations of this algorithm are that it assumes that all clusters are perfectly spherical, meaning it cannot work with any that are differently shaped. It also gives priority weighting to larger clusters, deeming them more important which may lead to false positives or negatives due to certain features of the data being deemed more important.

DECISION TREE

This algorithm uses recursive binary splitting, where on each recursion a piece of data is classified in a more specific group. A maximum depth is determined prior to the algorithms use in order to prevent infinite recursion and to ensure that data is only classified as specifically as needed. Some decision trees use a technique called

'pruning' where any branches of data deemed unimportant or irrelevant are removed and will not continue the recursion. One pruning technique is reduced cost pruning, which is one of the most popular ones.Reduced cost pruning is where the least popular nodes of a 'branch' are removed and this change is kept only if the accuracy of the algorithm is unaffected.

A benefit of using this algorithm is that there is relatively little pre-prossessing required, which is good for this organisation which may not have the resources to train an algorithm for a long period of time. As with k-means classification decision trees scale well and do not require retraining for this.

A limitation of this algorithm is it does not handle changing data very well. This is a big disadvantage for a network security system as the exact way a malicious packet is going to manifest itself cannot always be predicted, especially if new vulnerabilities are discovered after the algorithm has been trained.

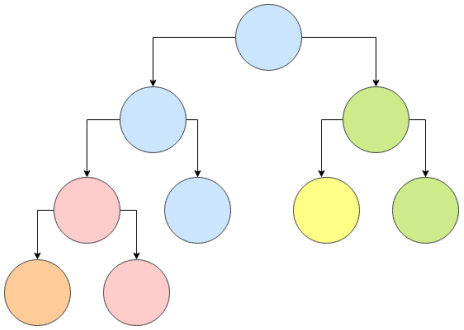

Figure 3: Basic decision tree diagram

3. DESIGN

For FooBar inc. the author will discuss the steps required to create a fictitious ML algorithm that will act as an IDS to detect any malicious packets on the network.

The first stage of building a machine learning algorithm is to determine how it ingests data. A very popular way is to provide the software with a csv file. These files can store a considerable amount of data in a way that the algorithm can parse through easily. Preproccessing is where the inputs are checked over to ensure that no data is missing and that it is displayed correctly. In the case of csv files all data must be separated by commas and certain unicode characters may not be readable by the algorithm. At this stage standard deviation must be reduced and the weighting of values is initially calculated.

The next step is to develop a model. A model in terms of machine learning is the result of the algorithm when it has been given data, it contains what the algorithm has learnt. This is important as the more an algorithm 'learns' the less errors it makes and the more accurately it can determine patterns. This is important for an intrusion detection system as errors can lead to network vulnerabilities being exploited by malicious users, such as the hacktivist organisation currently threatening FooBar inc.

Thirdly and perhaps one of the most important stages is to evaluate the results of the model. This can be done in multiple ways, as described below in section 4.

Any errors caught at the evaluation stage can be remedied by tuning the parameters and weighting assigned by the algorithm. This stage may have to be done multiple times in order to achieve the correct accuracy. This is important as if there are errors with the algorithm that go undetected it can render the network security inept, meaning it may be better to have no IDS than a heavily misconfigured one.

4. EVALUATION

In order to evaluate whether an algorithm works in the way is has been designed there are a number of metrics that can be used.

The confusion matrix is a two-dimensional array that is used to categorised the output of the ML algorithm. These categories are; True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). True positive or negative are outputs that correctly identify the data that has been provided whereas false positives or negatives are incorrectly classified as harmless or malicious network traffic *Alqudah and Yaseen, 2020).*

AUC-ROC curve, which stands for Area Under the Curve and Receiver Operating Characteristics respectively. ROC determines probability of false positives against true positives and plots these values on a curve. AUC as the name suggests takes the values under this curve and determines the algorithms threshold for correctly classifying data into classes.

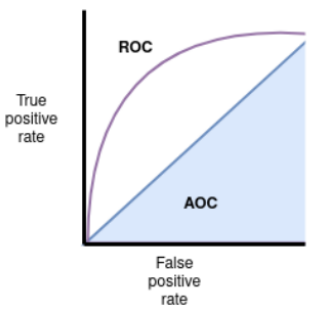

Figure 4: AUC-ROC Diagram

There exists a number of datasets that can be used to evaluate the effectiveness of an IDS algorithm. KDD Cup'99, created by Stolfo et al, is one of the most popular and contains around two million packets of data that is used to test an algorithm. It also contains data that can be used to train the IDS software *(Xiao et al, 2019)*. By using a standard dataset it is easier to establish how the algorithm is doing as the examiner can compare results to others.

5. REFERENCES

Ahmad, Z. et al. (2020). *"Network intrusion detection system: A systematic study of machine learning and deep learning approaches"*. Survey Paper, Malaysia:Universiti Malaysia Sarawak.

Alqudah, N. and Yaseen, Q. (2020). *"Machine Learning for Traffic Analysis: A Review"*. Procedia Computer Science, 170, pp.911–916

Almseidin, M. et al. (2020). *"Evaluation of Machine Learning Algorithms for Intrusion Detection System."* Research paper. Hungary:University of Miskolc.

Amrollahi, M. et al. (2020). *"Enhancing Network Security Via Machine Learning: Opportunities and Challenges"*. Handbook of Big Data Privacy. Switzerland: Springer, pp.165–189.

Malik, S. (2003). *"Chapter 3: Device Security"*. Network security principles and practices. Great Britain: Cisco, pp.36–81.

Mammen, B. et al. (2020). *Network Attack Trends: Attackers Leveraging High Severity and Critical Exploits.* [online] PaloAlto: Unit42. Available at: https://unit42.paloaltonetworks.com/network-attack-trends/ [Accessed 2 Apr. 2021].

Xiao, Y. et al. (2019) *"An Intrusion Detection Model Based on Feature Reduction and Convolutional Neural Networks"* in IEEE Access, vol. 7, pp. 42210-42219, 2019, doi: 10.1109/ACCESS.2019.2904620.