Meta-heuristics and Machine Learning for Detecting Cyber Attacks Literature Review

Christos Christou-Mavropoulos, Durham University, Durham, UK

### Abstract

**The exponential advancement of computer capability has allowed more complex applications to develop with an increasing number of features. However, this growth has allowed the creation of highly complex algorithms and methods which goals are to expose/attack systems. Today, more than ever, the multitude of attacks requires a system to be able to efficiently detect and counter these various attacks. A Network Intrusion Detection System (NIDS) monitors network traffic flow and packets to detect attacks. In order to train these systems to detect anomalies, a novel dataset can be used, however, actions must be taken beforehand to make the computations simpler. Such actions include dimensionality reduction and categorical data encoding and are going to be examined in this paper.**

**Terms:** dimensionality reduction, anomaly detection, curse of dimensionality

### 1. Introduction

Since 1999 there has been a significant lack of testing and training data of normal and abnormal behaviour in internet traffic used for intrusion detection. This has led to the reuse of the same datasets that are documented to have faults such as their representation of reality and methods used to gather the data (Amjad, 2018). In 2015, a new dataset was generated UNSW-NB15 produced a dataset that simulates both normal and different types of abnormal behaviour. The dataset consists of approximately 87.35% normal behaviour which means that the data is heavily imbalanced. The rest of the data contains the abnormal behaviour that can be divided into 9 sub-categories of attacks. (Moustafa and Slay, 2015) This poses a problem as some models require or can provide higher performance when the data-classes are balanced.

Furthermore, each data point collected contains 49 different dimensions (features) such as source IP/port, destination IP/port, timings and others. High dimensional data are prone to issues often categorized as curse of dimensionality. This term refers to various phenomena that arise when operating on high-dimensional data. Thus, it is common practice to reduce the number of dimensions while retaining as much of the original information as possible. This procedure is known as dimensionality reduction and can be achieved in many ways. (Zimek, Schubert and Kriegel, 2012)

Subsequently, a system should detect abnormal behaviour using meta-heuristics and should trained on the new dataset. Anomalies are patterns in data that do not conform to a defined notion of normal behaviour (Chandola, Banerjee, & Kumar, 2009). The process of detecting data points that do not conform to normal behaviour is referred as anomaly detection (Kotu and Deshpande, 2019). As mentioned before, the different classes are massively imbalanced which makes using supervised learning methods not favourable for detection. This is because the system will assume normal behaviour and get high accuracy number as a result of the imbalance and thus unsupervised methods should be used instead for the algorithm to not learn the imbalance as a feature. Another step that can be taken to avoid that is the oversampling of the minority class to balance the dataset. However, this process might add noise and bias to the results.

Moreover, some of the features in the data set are categorical and need to be encoded in a numeric format as they can cause further problems. Specifically, some techniques can only operate or produce more desirable results with quantitative data. Moreover, categorical variables have too many levels and can significantly impact an algorithms performance. Lastly, dealing with integers instead of strings will improve performance of the subsequent algorithms.

### 2. Method

In order to solve the above problems, firstly some pre-processing steps need to be applied. One of them is the conversion of categorical data to numerical format. Later, dimensionality reduction is going to be applied to the processed data to reduce the number of features. This is done in hopes of improving the performance of the anomaly detection algorithms and in addition, avoid the curse of dimensionality. Subsequently, a variety of anomaly detection algorithms and techniques will be used and compared. However, these last two processes should be repeated with the use of different combinations of the techniques discussed below in order to achieve higher prediction accuracy.

### 2.1 Categorical Data Conversion

As a pre-processing step, the features that are not in numeric or continuous format should be encoded into one for the reasons discussed above. To achieve that, one of the most popular methods is going to be used: one-hot encoding. With this method, one column is created for each value to compare against all other values. For each new column, a row gets a one if the row contained that columns’ value and a zero if it did not. For the IP addresses feature, one can use character encoding as there is an inherit structure to the feature’s digits. (Pasumarthy, 2019) Other encoding approaches that can be used are Target, LeaveOneOut, Binary, Hashing and Label/Ordinal encoding. Different methods can be used for different features to add the least amount of bias and increase efficiency. These should be used for features with a high cardinality. (Roy, 2019)

### 2.2 Dimensionality Reduction

There are multiple ways to reduce the features of a given data set. Firstly, feature elimination can be used which is the process of simply eliminating features that are ranked as unimportant. However, this only leads to a loss of information so it is not a popular choice and should not be used. Secondly, feature selection is a method in which the features are ranked and either the algorithm or the user can define a threshold to ignore some of the features. It can simplify the model which will lead to shorter training times, assist in avoiding the curse of dimensionality and enhance generalisation. The problem with this method is that one needs to assume that the data contain features that are redundant or irrelevant for our purposes that can be removed without loss of information. There are three main categories of feature selection algorithms: wrappers, filter and embedded methods. All of them can rank/score features based on similarity to others, amount of information given or other statistical models and by using a threshold one can select the most important one. Still, the earlier assumption is not often the case. In a lot of applications, most if not all of the features may have some importance and can assist in our predictions.

Finally, feature extraction/projection can be used which begins with the initial set of measured values and builds derived features that can be used for the prediction. Similarly, to before, it involves reducing the resources needed to describe data points transforming higher-dimensional data to a lower dimension. Unlike the previous methods though, more information can be preserved in the form of new features generated from the existing ones. In this project the last method will mainly be used. Techniques to achieve that can be both linear and non-linear.

### 2.3 Dataset Balancing

Balancing the normal and abnormal classes is a crucial step in order to train and test some of the anomaly detection models. There are two ways to achieve this which are under-sampling and oversampling. Under-sampling refers to the reduction of the majority class to reach the desired balance. However, this technique is prone to information loss because of the elimination of data points. Thus, most commonly oversampling is used. This method increases the minority class which in this case refers to the abnormal behaviour. The problem with the latter one is that by generating or coping data points from the initial data set, stronger levels of noise and outliers could be introduced which will make the model prone to over-fitting and reduce prediction performance.

### 2.4 Anomaly Detection

After the pre-processing steps have been completed, the below models should be used to detect anomalies in the data set. This can be achieved using many different methods and techniques that can be mainly divided into: autoencoders, clustering and unsupervised learning. (Buczak and Guven, 2016)

### 2.5 Visualisation

Visualisation is a key part of machine learning as it assists researchers understand and make sense of large data sets with many convoluting features. Many of the techniques discussed below (i.e. t-SNE and UMAP) are also used for visualisation before and after any operations performed on the dataset and can be used to interpret the structure of the problem as well as the results from the algorithms used.

### 3. Techniques to be Used

### 3.1 Feature Extraction Techniques

### 3.1.1 Principal Component Analysis (PCA)

PCA is one of the most popular techniques for dimensionality reduction. It is a linear mapping of data to a lower dimension space where variance is maximized. The features/components in this new space are independent. PCA is deterministic, computationally efficient and provides substantial results with Gaussian distributed data, however, it cannot interpret complex polynomial relationships between features. There can also be some information loss when selecting the number of principal components to be used. (Brems, 2019)

### 3.1.2 Non-Negative Matrix Factorisation (NMF)

It is a linear technique that decomposes a non-negative matrix to the product of two other non-negative matrices. This technique is stable and reliable only if there are non-negative signals which is often not the case for real applications. If the assumption is correct it has the ability to provide valuable results that are easy to interpret. (Leen, 2001)

### 3.1.3 Kernel Principal Component Analysis (KPCA)

Using the kernel trick (best known in Support Vector Machines or SVM) PCA can be used in a non-linear way in order to construct non-linear mappings. The kernel trick means the use of kernel functions that operate in high-dimensional implicit feature space without computing the coordinates of the data in that space. This is favourable because there are often underlying substructures in data that can only be discovered using non-linear embedding methods. KPCA is computationally expensive like most non-linear dimensionality reduction methods but the results are harder to interpret compared to other methods. (Patrikainen, 2003)

### 3.1.4 t-Distributed Stochastic Neighbour Embedding (t-SNE)

It is a widely used non-linear method that uses a normal distribution in the higher dimension and a t-distribution in the lower one to measure similarities of the data points and the corresponding embedding of them. It aims to keep the relative similarity of the data points as close to the original space as possible. Thus, it performs well with non-linear structures among features and can also be tweaked to achieve better performance in different data sets. On the other hand, these tuneable hyperparameters must be correct to avoid overfitting and using the same ones can give different results as the technique is non-deterministic. Lastly, there is also a danger of losing large scale information during the embedding (global structure).

### 3.1.5 Autoencoders

An autoencoder neural network is an unsupervised learning algorithm that can be used for dimensionality reduction and anomaly detection. An autoencoder can learn the non-linear substructure of the data but is prone of overfitting. There are many different types of autoencoders such as sparse, deep, denoising and contractive. A disadvantage of autoencoders is that they are computationally expensive for large datasets and moreover the results that they produce are mostly uninterpretable.

### 3.1.6 Unifold Manifold Approximation and Projection (UMAP)

UMAP is a dimensionality reduction technique that can also be used for visualisation of the data similarly to t-SNE. However, it is computationally faster than t-SNE and it retains as much of the local data structure and more of the global data structure. This is because the embedding is found by searching for a low dimensional projection of the data that has the closest possible equivalent topological structure. (McInnes, Healy and Melville, 2018)

### 3.2 Anomaly Detection

### 3.2.1 Long Short-Term Memory (LSTM) Autoencoders

LSTM autoencoders belong in the family of Recurrent Neural Networks (RNNs) which contain loops in them allowing information to persist. While in theory RNNs can work with long-term dependencies, in practice they are not able to learn them. LSTM networks however solve this dependency problem and can persist information for later use. They can and should be trained using only a heavily imbalanced or single behaviour class. They can be used to recognize anomalies, but also multiple networks can be used to distinguish anomaly subclasses that exist in the data. (Colah, 2015)

### 3.2.2 K-Means Clustering

A cluster refers to a collection of data points aggregated together because of certain similarities (Garbade, 2019) K-Means Clustering is a flat-clustering technique that is suitable for large databases, is computationally cheap and forms tight clusters that are easy to interpret. On the other hand, the number of k clusters needs to be decided by the user and furthermore, it can only handle numerical data which have to be converted from non-numeric, the process of which adds bias.

### 3.2.3 Hierarchical Clustering (HCA)

This connectivity-based clustering algorithm initially treats each data-point as a separate cluster. Then it recursively identifies the two cluster that are closest together and merges them until al the clusters are merged. (Patel, 2019) In the end, it produces a dendogram that is easy to understand but is mostly accurate at the bottom. The dendogram cannot point to how many clusters should there be. There exist a wide variety of distance metric and linkage criteria that can be used to get optimal results however it is a computationally expensive greedy algorithm. (Bock, 2019)

### 3.2.4 Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

Generally referred to as just DBSCAN, it is the most popular implementation of density-based clustering. It is robust against outliers as they would be in low-density regions and is widely used for dense datasets in anomaly detection (Celik, 2011). Some forn of density-drop is expected in order for the algorithm to detect cluster borders. (Salton do Prado, 2019)

### 3.2.5 Isolation Forest (iForest)

The difference with other methods is that instead of building a normal profile, isolation forest explicitly identifies anomalies instead. It is built on the basis of decision trees where partitions are created by randomly selecting a feature and the selecting a random split value between the minimum and maximum value of the selected feature. Thus, because anomalies and outliers are less frequent that normal observations and should be identified closer to the root. The forest can be trained with or without anomalies and is computationally cheap. (Lewison, 2019) Extended Isolation Forest (EIF) improve the consistency and reliability of the anomaly score produced. (Hariri, 2018)

### 3.2.6 Self Organizing Map (SOM)

This technique like some others can be used for dimensionality reduction as well as anomaly detection. It is deterministic and can be combined with other methods (i.e. HCA) and deals in neighbourhoods instead of distances. SOM, also called Kohonen neural network is a type of unsupervised machine learning techniques based on competitive learning, where only a single node is activated in each iteration. (Kohonen, 1990). It creates a network that maintains information on the topological relationships within the training data. (Tian, 2014)

### 4. Conclusion

The problem outlined in the introduction is a complex one and requires the application of many techniques to answer it. Firstly, pre-processing steps need to be applied such as the transformation of categorical data to quantitative. Class balancing should also be considered. Subsequently, with dimensionality reduction through feature extraction the problem should be easier to visualise and become more tractable. Finally, with the use of anomaly detection methods one will be able to identify potential intruders with a high accuracy rate. For this project a combination of techniques will be used and compared. Finally, the best of these techniques will be used to create a novel intrusion detection system that will be trained on the given dataset.

### 5. References

Amjad M. Al Tobi & Ishbel Duncan (2018) KDD 1999 generationfaults: a review and analysis, Journal of Cyber Security Technology, 2:3-4, 164-200, DOI:10.1080/23742917.2018.1518061

Bock, T. (2019). *What is Hierarchical Clustering? | Displayr.com*. [online] Displayr. Available at: https://www.displayr.com/what-is-hierarchical-clustering/)

Brems, M. (2019). *A One-Stop Shop for Principal Component Analysis*. [online] Medium. Available at: https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

Buczak, A. and Guven, E. (2016). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys & Tutorials*, 18(2), pp.1153-1176.

Celik, Mete & Dadaser-Celik, Filiz & Dokuz, Ahmet. (2011). Anomaly Detection in Temperature Data Using DBSCAN Algorithm. INISTA 2011 - 2011 International Symposium on INnovations in Intelligent SysTems and Applications. 10.1109/INISTA.2011.5946052.

Chandola, V., Banerjee, A. and Kumar, V. (2009). Anomaly detection. *ACM Computing Surveys*, 41(3), pp.1-58.

Garbade, M. (2019). *Understanding K-means Clustering in Machine Learning*. [online] Medium. Available at: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

Hariri, S., J. Brunner, R., Carasso Kind, M., 2018. Extended Isolation Forest. CoRR abs/1811.02141.

T. Kohonen, "The self-organizing map," in Proceedings of the IEEE, vol. 78, no. 9, pp. 1464-1480, Sept.1990.doi: 10.1109/5.58325

Kotu, V. and Deshpande, B. (2019). *Data science*. 2nd ed. Cambridge: Morgan Kaufmann is an imprint of Elsevier, p.446.

Leen, T. (2001). *Advances in neural information processing systems*. 13th ed. Cambridge Mass.: MIT Press, pp.556–562.

Lewison, E. (2019). *Outlier Detection with Isolation Forest*. [online] Medium. Available at: https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e

McInnes, L., Healy, J. and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. [online] Available at: http://arXiv:1802.03426

Moustafa, N. and Slay, J. (2015). UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). *2015 Military Communications and Information Systems Conference (MilCIS)*.

Pasumarthy, N. (2019). *Encoding fixed length high cardinality non-numeric columns for a ML algorithm*. [online] Medium. Available at: https://medium.com/@Nithanaroy/encoding-fixed-length-high-cardinality-non-numeric-columns-for-a-ml-algorithm-b1c910cb4e6d

Patel, Sanskruti & Patel, Atul. (2019). Performance Analysis and Evaluation of Clustering Algorithms.

Patrikainen, A. (2003). *Kernel Principal Component Analysis*. [ebook] Helsinki: Helsinki University of Technology. Available at: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8&ved=2ahUKEwipjqP83MjlAhVCtnEKHbTsDzkQFjABegQIARAC&url=http%3A%2F%2Fwww.cis.hut.fi%2FOpinnot%2FT-61.183%2Fk2003%2FTalks%2Fkernelpca.ps&usg=AOvVaw2Mbpls5klVkXXv\_UoMiJJd

Roy, B. (2019). *All about Categorical Variable Encoding*. [online] Medium. Available at: https://towardsdatascience.com/all-about-categorical-variable-encoding-305f3361fd02

Salton do Prado, K. (2019). *How DBSCAN works and why should we use it?*. [online] Medium. Available at: https://towardsdatascience.com/how-dbscan-works-and-why-should-i-use-it-443b4a191c80

Tian, J., H. Azarian, M. and Pecht, M. (2014). Anomaly Detection Using Self-Organizing Maps-Based K-Nearest Neighbor Algorithm. *IJPHM*, 5.

Zimek, A., Schubert, E. and Kriegel, H. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*, 5(5), pp.363-387.