Data Mining in Cybersecurity - Intrusion Detection System

Yoong Pei Xian 100334770,
Internet and Multimedia Techniques,
University of East Anglia

1 Chosen Topic and Rationale

In an article published by Deloitte on the Impact of COVID-19 on Cybersecurity, the coronavirus pandemic has resulted in a spike of cybersecurity due to the default remote working setting as employees do not have access to the same level of inherent protection at home (Nabe n.d.). About 47% of the employees are the victim of phishing fraud while working at home. The recent case of the Colonial Pipeline attack is one of the cases that resulted from the relaxed security of the remote working setting (Dudley & Golden 2021).

As data breaches resulting from remote working can cost as much as $137,000, there is a strong need to enhance the network design, where Intrusion Detection System (IDS) is highly recommended to be placed at each edge computing site. IDS is used to detect malicious or anomalous network traffic that cannot be detected by a conventional security system. It is divided into signature-based and anomaly-based.

Signature-based uses predefined models to compare the events and look for specific patterns of events to identify a new attack. It requires constant defines and updates of the signatures into the system, which is more labor-intensive and time-consuming. Its major drawback is that it cannot capture unlabelled data.

Anomaly-based adapts to capture new attacks without any prior knowledge, meaning that it can potentially detect an attack the first time it is used as the alarm is generated if the activity deviates from the normal activity (Buczak & Guven 2016).

In the design of IDS, data mining techniques are increasingly being used to gain insightful knowledge of intrusion prevention mechanisms (Salo et al. 2018). This is achieved by finding patterns in a big dataset to identify potential attacks. The most used techniques are classification and clustering.

A combination of classification and clustering that forms a hybrid learning algorithm is often adopted to address the weaknesses in each, and to produce a high detection rate and low false alarm rate (Buczak & Guven 2016).

2 Objective

To examine if hybrid models are more effective than individual models in detecting anomalies in IDS and their presented limitations.

3 Methods

This paper provides insights into the different research results of the impacts on IDS using classification, clustering, and hybrid learning algorithms.

Dataset: KDD 1999, a dataset extensively used by many researchers to detect and classify the anomaly in the computer network, particularly IDS KDD. (n.d.). The dataset consists of 4 categories of attack: Denial of Service (DOS), Remote to Local (R2L), User to Root (U2R), and Surveillance.

Pre-processing: The majority of the research papers modify the dataset by taking only $10-20\%$ of the subset as the dataset is too huge for a conventional computer to handle. The dataset is modified to include different amounts of malicious activity used for training and testing while retaining the characteristics of the data.

4 Findings

4.1 Classification

Classification is a supervised learning algorithm where it classifies network traffic into either normal or malicious activity based on the classifiers (Peddabachigari et al. 2004). It handles only labeled data, resulting in a weaker performance when it comes to unlabelled data. Several most popular classification algorithms include decision trees, artificial neural networks, and support vector machines.

4.1.1 Decision Tree (DT)

A Decision Tree is a tree-structured classifier that consists of nodes that represent different attributes that break the dataset into smaller subsets (Peddabachigari et al. 2004). It is easy to understand but the main issue of the DT is that it is challenging to locate the most ideal attributes to split the data into their corresponding classes. To overcome this, Quinlan (1996) has come up with the DT C4.5 model that builds the DT using the information gain.

Information gain, which is a measurement taken using the entropy value to measure the utility of each attribute is adopted to select the attributes. Larger information gain usually symbolizes a more useful attribute (Peddabachigari et al. 2004). The generalization accuracy of the decision tree allows it to capture new attacks, thus used in anomaly-based IDS (Peddabachigari et al. 2004).

Research conducted by Peddabachigari et al. (2004) and Lee et al. (2008) demonstrates a similar result of which the DT can present a high detection rate only on the more frequently occurring classes, such as DoS and Probe as rare classes do not present any sequential patterns (Lee et al. 2008). To improve the performance, Gudadhe et al. (2010) proposed an ensemble learning method – boosted DT, which combines several decision trees to take the weighted majority vote of each tree, which demonstrates a better detection rate, but it is more complex and time-consuming to perform.

Balogun (2015) has proposed a hybrid of Decision Tree and KNN learning approach showing that the hybrid method performs slightly better than the individual DT and KNN algorithms, and it can capture the new attacks that do not previously exist in the system effectively. Nevertheless, it is still unable to achieve a 100% detection rate across all classes. Kim et al. (2014) has also proposed a hybrid model of DT and SVM and the integration has shown an improvement for the anomaly-based system to capture the new attacks at a higher detection rate and faster speed.

4.1.2 Support Vector Machine (SVM)

SVM finds a hyperplane to divide the training data set into two classes and predict which class a new sample falls into (Li et al. 2012). It demonstrates better performance for text classification when compared to other algorithms such as Naïve Bayes and KNN classifiers (Li et al. 2012).

Owning to its decent generalization nature and the ability to deal with high dimensional data very effectively, it gains wide popularity in the IDS. However, when it comes to small training data, such as the U2R and R2L in the KDD 1999, Peddabachigari et al. (2004) concluded that

SVM has a much higher misclassification rate when compared to DT, and a longer training and testing time, in the research conducted using binary-class classification method.

Unlike DT, SVM does not cope with multi-class classification but only binary-class classification, and it is difficult to tune the regularization parameters to avoid over-fitting. Overfitting occurs when the model is too complex and it fits too closely or exactly to a particular dataset. To address this, Kim et al. (2005) have proposed to chain up the SVM with a Genetic Algorithm (GA). The result shows that the algorithm can find the optimal parameters and features sets, thus reducing the number of feature sets required and the processing time, while achieving a detection rate of over 99%. The integration of SVM with other algorithms also addresses the issues that SVM usually underperforms in detecting new attacks.

4.1.3 Artificial Neuron Network (ANN)

ANN is a mathematical model resembling how the biological nervous system of the human brain works, where it consists of many interrelated processing elements (neurons) (Al-Janabi & Saeed 2011). It possesses the strengths of working with incomplete knowledge, meaning that the data can produce output even with incomplete information after training (Mijwil 2018).

The structure of ANN consists of 3 layers - an input layer, intermediate layer, and output layer (Mijwil 2018). The learning process of ANN is to fine-tune the weights and strength between the neurons and then assign it to the connection points in the intermediate layer (Dias et al, 2017). The inherent high speed of ANN makes it more useful in real-time IDS.

ANN needs a large input pattern with decent distributed attacks in the dataset to present high accuracy during the training phase (Peddabachigari et al. 2004). Therefore, it is frequently used in a hybrid IDS with other algorithms such as DT, SVM, and GA and researchers have demonstrated hybrid algorithms can improve the accuracy but to compromise on the higher training time (Hosseini & Mohammad Hasani Zade 2020). To address this, Thaseen et al. (2021) integrates the correlation-based feature selection (CFS) technique to extract the best attributes that help reduce the dataset size and processing time.

4.2 Clustering

Clustering is an unsupervised learning algorithm used to find patterns in high-dimensional unlabelled data without the need for explicit descriptions to the system (Vasileios et al. n.d.). Data with high similarity measures are often grouped as a cluster, taking the Euclidean distance as the reference.

K-means algorithm is the most often used clustering algorithm as it is usually labor-intensive and time-consuming to manually label the data. The purpose of using K-means is to separate the normal and malicious activities that consist of similar behaviors into different clusters (Effendy et al. 2017). It comes with the limitations of assuming that the normal activities constitute a larger proportion of over 98% and the clustering width needs to be predefined, which is particularly challenging (Vasileios et al. n.d.).

Y-means algorithm, in return, resolves these issues where the clusters are automatically split and merged, but it only recognizes clusters with spherical shapes. To resolve this, Vasileios et al. (n.d.) adopts fuzzy connectedness as the similarity metric to deal with the data that does not consist of a homogenous shape. Fuzzy connectedness is often used in image segmentation where it calculates the similarities of different data by the strength of connectedness, which results in a highly effective segmentation. The research demonstrates a detection rate of over 97% and the test result improves when it involves a higher number of nearest neighbors.

On the other hand, Hendry & Yang (2008) perceived such test conducted by Vasileios et al. (n.d.) poses the weakness of being unable to capture a new attack that does not previously exist in the system. Therefore, they have proposed a Simple Logfile Clustering Tool (SLCT), which creates and updates signature creation while capturing the new attack methods in real-time. Both pieces of research deliver a detection rate of over 90% in the signature-based system, but the figure reduces to 70% - 80% when new attacks are introduced as it is more challenging for SLCT to identify which cluster exactly does the attack belong to.

5 Summary

In this paper, we examine the strengths, limitations, and effectiveness of different data mining algorithms – classification and clustering, using the KDD 1999 dataset as the benchmark. Hybrid IDS can overcome the individual limitations of each algorithm and produce a more accurate result. However, some of the hybrid models present limitations such as compromising on the training time in exchange for a higher detection rate.

In this study, we can conclude the following:

- DT works well with dominant classes but presents weakness when to identify rare classes. While boosted DT can improve the detection rate on rare classes, it compromises the complexity and speed, which can be addressed by hybrid algorithsm such as DT-KNN or DT-SVM.
- SVM presents a higher misclassification compared to DT when handling rare classes and does not handle new anomalies well. Its performance on detection rate and speed is highly improved when integrated with GA.
- ANN requires a larger output and a decent distributed dataset compared to DT and SVM to achieve a high detection rate, and the training time is longer. Using correlation-based feature selection (CFS) technique and address the issues.

Thus, different hybrid algorithms should be the focus of future studies to address the limitations to better cope with the increasingly rampant cybersecurity issues.

The majority of the papers neglects the fact that KDD 1999 is a set of imbalanced data, which is the case of most of the real-world dataset where the data is skewed. This can be considered as one of the weaknesses of most of the learning algorithms, which do not handle skewed datasets and lean towards the dominant classes and assume that the positive and negative examples are the same. Future work should include the investigation into the imbalanced data resampling methods combined with different data mining algorithms to improve the performance.

References

Al-Janabi, S. T. F. & Saeed, H. A. (2011), A neural network based anomaly intrusion detection system, *in* ‘Conference: Developments in E-systems Engineering (DeSE)’, pp. 221–226.

Balogun, A. (2015), ‘Anomaly intrusion detection using an hybrid of decision tree and k-nearest neighbor’, **2**.

Buczak, A. L. & Guven, E. (2016), ‘A survey of data mining and machine learning methods for cyber security intrusion detection’, *IEEE Communications Surveys Tutorials* **18**(2), 1153–1176.

Dudley, R. & Golden, D. (2021), ‘The colonial pipeline ransomware hackers had a secret weapon: self-promoting cybersecurity firms’.
URL: *https://www.technologyreview.com/2021/05/24/1025195/colonial-pipeline-ransomware-bitdefender/*

Effendy, D. A., Kusrini, K. & Sudarmawan, S. (2017), Classification of intrusion detection system (ids) based on computer network, *in* '2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)', pp. 90–94.

Gudadhe, M., Prasad, P. & Kapil Wankhade, L. (2010), A new data mining based network intrusion detection model, *in* '2010 International Conference on Computer and Communication Technology (ICCCT)', pp. 731–735.

Hendry, G. & Yang, S. J. (2008), Intrusion signature creation via clustering anomalies, *in* 'SPIE Defense + Commercial Sensing'.

Hosseini, S. & Mohammad Hasani Zade, B. (2020), 'New hybrid method for attack detection using combination of evolutionary algorithms, svm, and ann', *Computer Networks* **173**, 107168.

Kim, D. S., Nguyen, H.-N. & Park, J. S. (2005), Genetic algorithm to improve svm based network intrusion detection system, *in* '19th International Conference on Advanced Information Networking and Applications (AINA'05) Volume 1 (AINA papers)', Vol. 2.

Kim, G., Lee, S. & Kim, S. (2014), 'A novel hybrid intrusion detection method integrating anomaly detection with misuse detection', *Expert Systems with Applications* **41**(4, Part 2), 1690–1700.
URL: *https://www.sciencedirect.com/science/article/pii/S0957417413006878*

Lee, J.-H., Lee, J.-H., Sohn, S.-G., Ryu, J.-H. & Chung, T.-M. (2008), Effective value of decision tree with kdd 99 intrusion detection datasets for intrusion detection system, *in* '2008 10th International Conference on Advanced Communication Technology', Vol. 2, pp. 1170–1175.

Li, Y., Xia, J., Zhang, S., Yan, J., Ai, X. & Dai, K. (2012), 'An efficient intrusion detection system based on support vector machines and gradually feature removal method', *Expert Syst. Appl.* **39**, 424–430.

Mijwil, Maad, M. (2018), 'Artificial neural networks advantages and disadvantages'.
URL: *https://www.linkedin.com/pulse/artificial-neural-networks-advantages-disadvantages-maad-m-mijwel*

Nabe, C. (n.d.), 'Impact of covid-19 on cybersecurity'.
URL: *https://www2.deloitte.com/ch/en/pages/risk/articles/impact-covid-cybersecurity.html/*

Peddabachigari, S., Abraham, A. & Thomas, J. (2004), 'Intrusion detection systems using decision trees and support vector machines', *International Journal of Applied Science and Computations* **11**.

Quinlan, J. R. (1996), 'Learning decision tree classifiers', *ACM Comput. Surv.* **28**(1), 71–72.
URL: *https://doi.org/10.1145/234313.234346*

Salo, F., Injadat, M., Nassif, A. B., Shami, A. & Essex, A. (2018), 'Data mining techniques in intrusion detection systems: A systematic literature review', *IEEE Access* **6**, 56046–56058.

Thaseen, I. S., Banu, J. S., Lavanya, K., Ghalib, M. R. & Abhishek, K. (2021), 'An integrated intrusion detection system using correlation-based attribute selection and artificial neural network', *Trans. Emerg. Telecommun. Technol.* **32**.

Vasileios, Q. W., Wang, Q. & Megalooikonomou, V. (n.d.), 'A clustering algorithm for intrusion detection'.