2015

Analysis into developing accurate and efficient intrusion detection approaches

Priya Rabadia
Security Research Institute, Edith Cowan University, prabadia@ecu.edu.au

Craig Valli
Security Research Institute, Edith Cowan University, c.valli@ecu.edu.au

DOI: 10.4225/75/57b3f841fb88b
This paper was originally presented at The Proceedings of [the] 13th Australian Digital Forensics Conference, held from the 30 November – 2 December, 2015 (pp. 70-76), Edith Cowan University Joondalup Campus, Perth, Western Australia.

ANALYSIS INTO DEVELOPING ACCURATE AND EFFICIENT INTRUSION DETECTION APPROACHES

Priya Rabadia, Craig Valli
Security Research Institute, Edith Cowan University, Perth, Australia
prabadia@ecu.edu.au, c.valli@ecu.edu.au

Abstract
Cyber-security has become more prevalent as more organisations are relying on cyber-enabled infrastructures to conduct their daily actives. Subsequently cybercrime and cyber-attacks are increasing. An Intrusion Detection System (IDS) is a cyber-security tool that is used to mitigate cyber-attacks. An IDS is a system deployed to monitor network traffic and trigger an alert when unauthorised activity has been detected. It is important for IDSs to accurately identify cyber-attacks against assets on cyber-enabled infrastructures, while also being efficient at processing current and predicted network traffic flows. The purpose of the paper is to outline the importance of developing an accurate and effective intrusion detection approach that can be deployed on an IDS. Further research aims to develop a hybrid data mining intrusion detection approach that uses Decision Tree classifications and Association Rules to extract rules using the classified data.

Keywords:
Association Rules, Cyber Security, Data Mining, Decision Trees, Intrusion Detection System

INTRODUCTION

With cyber-enabled infrastructure increasingly utilised by the global community to conduct their daily activities, securing these assets becomes a priority. These assets include: data, networks and systems that could be vulnerable to cyber-attacks (Al-Ahmad, 2013). Cyber-security strategies are used to mitigate cyber-attacks. An Intrusion Detection System (IDS) is a cyber-security tool commonly deployed as part of a defence in depth cyber-security strategy (Madbouly, Gody, & Barakat, 2014).

IDSs monitor network traffic and trigger an alert when unauthorised activity is detected (Rajasekaran & Nirmala, 2012). The concept of intrusion detection was first introduced by Anderson (1980). IDSs can be either software based or hardware based, depending on the requirements and resources of the individual or organisation deploying them. There are two common types of IDSs: Host based Intrusion Detection System (HIDS) and Network Based Intrusion Detection System (NIDS). HIDS operate on an individual host located on the network and monitor inbound and outbound packets for that particular host, whereas NIDS are placed strategically on a network to monitor the network traffic (Spathoulas & Katsikas, 2010).

In order to detect intrusions on a network, IDSs commonly employ a signature or misuse-based detection approaches or either an anomaly-based detection approaches. The signature or misuse-based detection approaches involves comparing the signatures of current network traffic to a databases of known signatures of bad network traffic. If a reasonable match is found, an alert is triggered indicating a possible intrusion on the network (García-Teodoro, Díaz-Verdejo, Maciá-Fernández, & Vázquez, 2009). Whereas anomaly-based detection approaches involve generating a baseline of normal traffic exhibited on a system or network then comparing it to current network traffic. Any nonconformities to the baseline are considered to be an anomaly, triggering an alert for a possible intrusion on the network (Portokalidis & Bos, 2007).
The detection approaches classify IDS network traffic into four distinct groups:

- True Negative (TN): unauthorised traffic is classified as an intrusion.
- False Negative (FN): authorised traffic is misclassified as an intrusion.
- True Positive (TP): normal traffic is classified as authorised traffic.
- False Positive (FP): unauthorised traffic is misclassified as authorised traffic.

The percentage of network traffic classed into these groups show the accuracy of an IDS. The lower the rate of FNs and FPs in conjunction with a high rate of TNs and TPs, show that the implemented IDS detection approach is accurate at mitigating cyber-attacks (Elngar, Mohamed, & Ghaleb, 2012; Madbouly et al., 2014; Spathoulas & Katsikas, 2010).

In addition to IDS detection approaches being more accurate, they also need to be efficient at analysing network traffic. As predictions have shown with the passing of each year, traffic flowing through a network is also set to increase. Figure 1, is an adaptation from Cisco's 2015 white paper forecasting the trends and predictions of global internet traffic from 1992 until 2019 (Cisco, 2015).

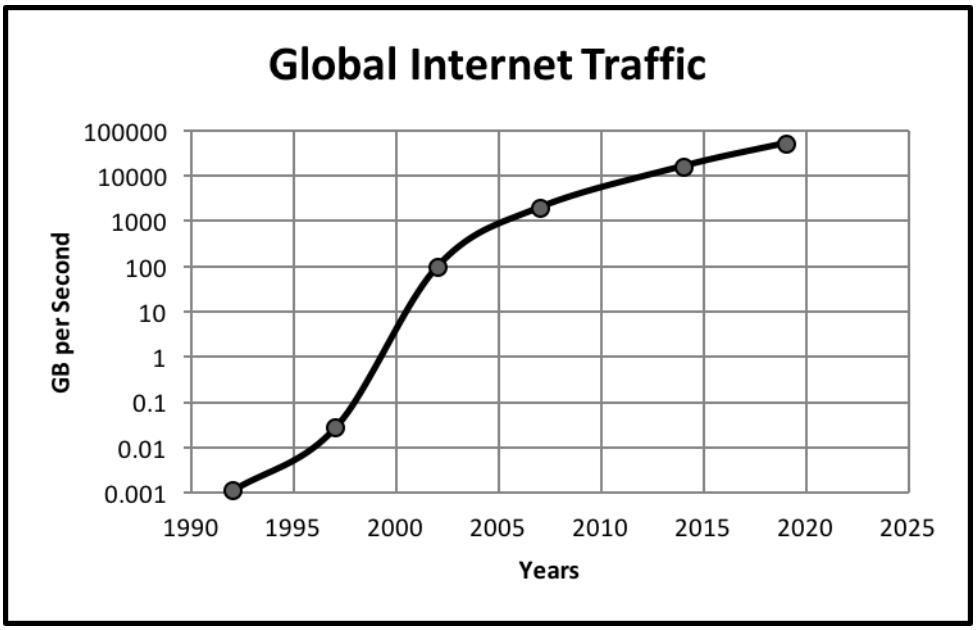

Figure 1: A representation of the Cisco Visual Network Index (VNI) forecast – Historical internet context (Cisco, 2015), showing the increase of internet traffic from 1992 – 2019

Although Figure 1 only shows the increase in global Internet traffic, the current trend is being forecasted for all networks using cyber-enabled infrastructures. Figure 1, shows global Internet traffic is predicted to increase to 51,794 GB per second in 2019 from 16,144 GB per second in 2014. As global network traffic is forecast to increase, traditional IDSs need to be able to detect intrusions accurately and efficiently. Other techniques of intrusion detection need to be researched aside from the commonly deployed signature or misused-based detection and anomaly-based detection approaches. Studies have been conducted using data mining to increase the accuracy of IDSs by analysing the percentages of FNs and FPs (Elngar et al., 2012; Madbouly et al., 2014; Spathoulas & Katsikas, 2010). However, research needs to be conducted on applying data mining algorithms and techniques to increase both the accuracy and efficiency of an IDS. The purpose of the paper is to highlight the need for an accurate and efficient intrusion detection approach using data mining algorithms and techniques.

The structure of the paper is as follows: explore the literature on the accuracy and efficiency of intrusion detection approaches using data mining algorithms and techniques, discuss the proposed hybrid data mining solution followed by a conclusion and suggested future work to be conducted.

ACCURACY AND EFFICIENCY OF INTRUSION DETECTION APPROACHES

IDSs are cyber-security tools with the purpose of detecting intrusions on a network or system. IDSs need to be able to accurately detect intrusions in order to mitigate cyber-attacks and to prevent exposure of assets. For an IDS to be accurate at detecting intrusions, the rate of false negatives (FN) and false positives (FP) detected within network traffic must be low. In turn the rate of true negatives (TN) and true positives (TP) detected within network traffic must be high.

There are many studies within literature that have conducted researched into improving the accuracy of IDSs. The application of applying filters to IDS detection approaches in order to improve the accuracy has been conducted by various researchers. A study conducted by Spathoulas and Katsikas (2010) suggested implementing post-processing filters to identify FPs in processed SNORT IDS network traffic. In this study the 1999 DARPA intrusion detection dataset (1999 DARPA) was applied to the SNORT IDS. Then three filters where applied to the sorted traffic one after another.

1. Neighbouring Related Alerts – NRA: Based on the assumption that attackers produce a burst of attacks in the early stage when looking for vulnerabilities and their victims.
2. High Alter Frequency – HAF: Based on the assumption that attacks usually produce anomalies.
3. Usual False Positives – UFP: Exploits the signature subgroup in order to detect false positives.

Each filter was applied separately to the sorted SNORT network traffic to evaluate the probability of each record being a TP. All the evaluations were then combined and a final decision was made on whether the record is a FP or a TP. The post-processing filters are suggested to reduce the percentage of FP by 75%. However, the article also makes reference to the 1999 DARPA dataset being problematic and out of date, but was chosen as the testing dataset. Spathoulas and Katsikas conducted another study in 2013 (Spathoulas & Katsikas, 2013) following on from the 2010 study. The experiment was conducted on the 2000 DARPA dataset and generated 'real' network traffic. Although results from the 2013 study appear to support the initial findings from 2010, there is no discussion into the reduction of the FP percentage.

Studies have been conducted in applying data mining algorithms and techniques to improve the accuracy of IDS detection approaches. Research conducted by Elgar et al. (2012) proposed increasing classification accuracy as well as decreasing the build time speed of an IDS by improving the feature selection of the C4.5 Decision Tree by using Particle Swarm Optimisation (PSO). The proposed method is referred to as PSO- Decision Tree IDS. Records where selected randomly from the NSL-KDD dataset, to test the proposed PSO- Decision Tree IDS classification. (The NSL-KDD dataset is a subset of the KDD 99 dataset, which a subset of the DARPA intrusion detection dataset.)

The PSO- Decision Tree DS has three phases: Pre-processing, feature selection based on PSO and classification using C4.5 Decision Tree. Three feature selection methods where tested and compared in order to evaluate the performance of the PSO- Decision Tree IDS. The experiments where conducted on:

1. C4.5 Decision Tree (Standard)– The standard Decision Tree with the no feature selection. All 41 features where tested.
2. GA - Decision Tree – Genetic Algorithm used for feature selection. Selecting 12 out of 41 features where then applied to a Decision Tree.
3. PSO- Decision Tree IDS – This is the proposed detection model using, PSO feature selection selected 11 out of the 41 features where then applied to a Decision Tree.

The finding from Elgar et al. (2012) study suggests there is a correlation between the selection of features and build time. The higher the number of features the lower the accuracy and building time required. The PSO-Decision Tree IDS reduced 41 features to 11, with 99.17% accuracy with 11.65 sec building time. Although the study does conclude a highly accurate detection approach is achievable, there was no testing process conducted to verify whether the accuracy is the same when applied to an IDS, with no mention of further research to be conducted. The research conducted Elgar et al. (2012) compared the PSO-Decision Tree IDS to two other feature selection methods: C4.5 Decision Tree (standards) and GA-Decision Tree.
Both studies have shown by enhancing traditional intrusion detection approaches, the accuracy of an IDS can be improved. Spathoulas and Katsikas (2010) reduced the rate of FPs by 75% using post-processing filters. While the study conducted by Elgar et al. (2012) used PSO to reduce the number of features deployed on a C4.5 DT to achieve an accuracy rating of 99.1% and a build time of 11.65 seconds. Other studies have been conducted in improving accuracy and efficiency in building time of intrusion detection approaches.

Research into improving the accuracy and efficiency of build time of IDS detection approaches has been conducted by Derhab and Bouras (2015). The purpose of the study was to analyse and evaluate the proposed IDS detection approach on network traffic attacks. The proposed IDS approach is Normal Behavioural Graph (NBG) based multivariate correlation analysis. Multivariate correlation analysis is used for feature extraction and then is used in the develop the NBG. The experiment was conducting using the KDD 99 and the NSL-KDD dataset, where proposed proposed detection approach was compared to multiple existing detection approaches including: statistical-based, data mining and machine learning algorithms. The study concluded the NBG outperformed its competitors with 99.76% accuracy on the KDD 99 dataset and 84.63% on the NSL-KDD dataset, with an acceptable building time of 25.14 seconds. Although studies have shown achieving accuracy and efficient build time is possible, little research has been conducted in improving efficiency in network traffic processing.

The study conducted by Sadeghi and Bahrami (2013) proposes improving the Negative Selection (NS) feature traditionally used in IDSs to increase the efficiency. The study proposed using an Artificial Immune System (AIS) applied to the NS algorithm. AIS is based on the ability of the human body to detect intruding pathogens and cells. The AIS uses various feature to detect intruder activities within network traffic, based on pattern comparison similar to anomaly-based detection and clustering. The experiment consisted of testing the traditional NS and NS using AIS on the KDD 99 dataset. The results suggested the improved NS using AIS is 50.45% more efficient at processing network traffic data then the traditional NS feature selection. Although the results from the study show there is an improvement in the efficiency of the IDS by 50.45%, there is no comparison of the accuracy rate of the traditional NS and the NS using AIS detection algorithm in the study. While research has been conducted into improving aspects of intrusion detection approaches such as the accuracy and building time of intrusion detection approaches. There is a need for research to be conducted in improving efficiently processing network traffic as well as accuracy of classifying network traffic.

Discussion

Analysis in the domain of improving the accuracy of IDS approaches has shown that applying feature selections to data mining algorithms improves the accuracy of intrusion detection approaches. Many of the studies that have been conducted use the outdated DARPA dataset and its variations. Research needs to be conducted in applying proposed approaches to 'real' network traffic and comparing then to existing intrusion detection approaches. The gap in knowledge is in improving the accuracy and efficiency of IDS detection approaches. Studies have been conducted in the area however they do not address the problem. A study by Albayati and Issac (2015) was conducted to test a hybrid data mining classification approach could improve accuracy and efficiency on an IDS. The experiment on the prototype was tested using a subset of NSL-KDD dataset. The researchers expressed the limitations of their research being the use of a subset of an outdate dataset and the prototype is yet to be properly trained and tested. However, early experiments are promising. The difficulty is developing a detection approach that is accurate at detecting intruder activities and is also efficient processing 'real-time' network traffic.

PROPOSED SOLUTION

Research needs to be conducted in the domain of intrusion detection approaches improving the accuracy of network traffic classifications, while also being efficient at processing current and predicted network traffic. The proposed solution is to develop a hybrid data mining intrusion detection approach, that combines Decision Trees and Association Rules.

Decision Tree

Decision Trees are used to model particular outcomes, based on statistical probability, using a dataset of past events. Decision Trees are a common data mining model as data can be classified and then rules can be extracted. Decision trees are used to classify data into sets and subsets and illustrate the statistical probability of an event occurring. The aim of a Decision Trees is to create subsets of events or variables until a pure subset has been identified.

Decision Trees allow data to be classified and visually represented. A study conducted by Li, Zhang & Ogihara (2004) showed that the more classes the multiclass has to handle, the more the accuracy of the classified data will decrease. Research has been conducted to overcome this issue by Guh and Shiue (2008). Their work showed that a univariate decision tree was less accurate and efficient when compared to using an Artificial Neural Network (ANN). Nevertheless, ANNs are time consuming to setup, maintain and the results can be difficult to interpret. Guh and Shiue suggested using a single Decision Tree multiclass classifier to identify multiple class instances. The results from their experiment showed that their approach outperformed an ANN. However He, Wang, Zhang, & Cook (2013) argues that as the dataset gets more complex, the number of classes will increase exponentially and the concept of a single decision tree multiclass classifier would not be accurate or effective. He suggested using a new MSPC (Multiple Statistical Process Control.) The new model proposes using multiple decision tree classifications with each one handling no more than three classes. The experiment results showed this proposed model can handle a greater number of classes without having any in decrease accuracy or efficiency when compared to the Guh and Shiue model. Further research needs to be conducted on the application of the new MSPC in the cyber-security domain.

Association Rules

Association Rules are a data mining technique which apply algorithms to extract rules that identify correlations between two or more features in a dataset and assign a statistical probability of the rule occurring within the training set. Association Rules have been applied to other fields including network security (Wang & Bridges, 2000), IDSs (Ping-ping & Qiu-ping, 2002) and predicting phishing websites (Aburrous, Hossain, Dahal, & Thabtah, 2010).
The application of Association Rules as a data mining technique has the advantage of clearly conveying information such as the rules extracted from a dataset, making it easier for the system administrator to interpret and deploy. The disadvantage of using Association Rules is the lack of flexibility associated when assigning variables to the rules. Association Rules rules are intended to handle Boolean data, however, real world quantitative data is not in a Boolean format. A solution is that the data must be divided into segments and then the Boolean format can be applied. Applying the solution could result in underemphasising or overemphasising the boundary of the segments. This is known as the sharp boundary problem (Kalia, Dehuri, & Ghosh, 2013). Researchers in the domain of Association Rules have been focusing on applying fuzzy logic to Association Rules to overcome the sharp boundary problem.

The current direction of research is on applying fuzzy logic to Association Rules. Fuzzy logic allows more flexible segment boundaries, by giving the administrator control of defining the fuzzy set range associated with the boundary. Normally the fuzzy set is {Low, Medium, and High}, however the administrator has full control of the definition of the fuzzy set values. The fuzzy rules are used in conjunction with the support and confidence values (Kalia et al., 2013). The application of fuzzy Association Rules has been studied with research conducted in the cyber-security domain of phishing attacks (Aburrous et al., 2010). The purpose of the Aburrous et.al (2010) experimental study was to investigate the most effective algorithm for identifying rules in order to predict phishing websites. The algorithms used in their study were Classification and Association Rules algorithms with fuzzy logic applied. These algorithms are different from the traditional Association Rules algorithms such as Apriori. The study compared six Classification and Association Rules algorithms in order to identify which one is most effective at identifying phishing websites. The dataset used in this experiment consist of 1006 elements. Results of the experiment showed the MCAR (Multi Class Classification based on Association Rule) algorithm was the most effective with an error rating of 12.662%. The research conducted by Aburrous et.al showed it is possible to uses Association Rules with the application of classification to predict phishing websites. However, further research needs to be conducted in the application of Association Rules in order to predict malicious events in other aspects of the cybersecurity domain, such as IDSs.

In summary, Decision Trees are effective and accurate at classifying datasets. While Association Rules are effective at extracting highly accurate rules from a given dataset. The proposed solution is a hybrid data mining intrusion detection approach that is accurate and efficient at detecting intrusions on a network. Decision Trees will be used to classify the datasets and Association Rules will be used to extract rules from the classified dataset.

CONCLUSION

As cybercrime is increasingly becoming a problem for the global community (McAfee, 2014). Cyber-security is becoming the focus of many governments and organisations. Existing cyber-security strategies need to be both accurate and efficient at processing ‘real-time’ network traffic. This paper explored the need for an IDS detection approach that is accurate at classifying traffic and efficient at processing network traffic to keep up with predicted future network traffic flows. Research has been conducted into improving individual aspects of intrusion detection approaches such as accuracy, however there has been little to no research conducted on developing an intrusion detection approach that is both accurate and efficient. There is a need for an IDS detection approach that can classify network traffic accurately as well as process the network traffic efficiently.

Future Work

The next step is to develop a hybrid data mining intrusion detection approach that is both accurate and efficient at detecting intrusions on a network. The proposed research to be under taken will use a hybrid approach combining Decision Trees to classify the dataset and Association Rules to extract rules using the classified data. The dataset intended for use will be acquired from three medium interaction research based honeypots located on the same /24 IPv4 address space. The proposed hybrid approach will be trained and tested against established data mining detection approaches in order to compare the accuracy and efficiency of the proposed IDS detection approach.

REFERENCES:

Aburrous, M., Hossain, M. A., Dahal, K., & Thabtah, F. (2010, 12-14 April 2010). *Predicting Phishing Websites Using Classification Mining Techniques with Experimental Case Studies.* Paper presented at the Information Technology: New Generations (ITNG), 2010 Seventh International Conference on.ADDIN EN.REFLISTXACC, A. C. C. (2015). Organised Crime in Australia 2015.

Albayati, M., & Issac, B. (2015). Analysis of Intelligent Classifiers and Enhancing the Detection Accuracy for Intrusion Detection System. *International Journal of Computational Intelligence Systems, 8(5)*, 841-853. doi: 10.1080/18756891.2015.1084705

Al-Ahmad, W. (2013). A detailed strategy for managing corporation cyber war security. *International Journal of Cyber-Security and Digital Forensics, 2*, 1+.

Anderson, J. P. (1980). Computer security threat monitoring and surveillance (pp. 56).

Cisco. (2015). The Zettabyte Era: Trends and Analysis *Visual Networking Index - VNI* (pp. 29). http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.html

Derhab, A., & Bouras, A. (2015). Multivariate correlation analysis and geometric linear similarity for real-time intrusion detection systems. *Security and Communication Networks, 8*(7), 1193-1212. doi: 10.1002/sec.1074

Elngar, A. A., Mohamed, D. A. E. A., & Ghaleb, F. F. M. (2012). A Fast Accurate Network Intrusion Detection System. *International Journal of Computer Science and Information Security, 10*(9), 29-35.

García-Teodoro, P., Díaz-Verdejo, J., Maciá-Fernández, G., & Vázquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. *Computers & Security, 28*(1–2), 18-28. doi: http://dx.doi.org/10.1016/j.cose.2008.08.003

Guh, R.-S., & Shiue, Y.-R. (2008). An effective application of decision tree learning for on-line detection of mean shifts in multivariate control charts. *Computers & Industrial Engineering, 55*(2), 475-493. doi: http://dx.doi.org/10.1016/j.cie.2008.01.013

He, S., Wang, G. A., Zhang, M., & Cook, D. F. (2013). Multivariate process monitoring and fault identification using multiple decision tree classifiers. *International Journal of Production Research, 51*(11), 3355-3371. doi: 10.1080/00207543.2013.774474

Kalia, H., Dehuri, S., & Ghosh, A. (2013). A survey on fuzzy association rule mining. *International Journal of Data Warehousing and Mining, 9*, 1+.

Li, T., Zhang, C., & Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics, 20*(15), 2429-2437. doi: 10.1093/bioinformatics/bth267

Madbouly, A. I., Gody, A. M., & Barakat, T. M. (2014). Relevant Feature Selection Model Using Data Mining for Intrusion Detection System.

McAfee, I. S.-. (2014). Net Losses: Estimating the Global Cost of Cybercrime.

Portokalidis, G., & Bos, H. (2007). SweetBait: Zero-hour worm detection and containment using low- and high-interaction honeypots. *Computer Networks, 51*(5), 1256-1274. doi: http://dx.doi.org/10.1016/j.comnet.2006.09.005

Ping-ping, M., & Qiu-ping, Z. (2002). Association rules applied to intrusion detection. *Wuhan University Journal of Natural Sciences, 7*(4), 426-430. doi: 10.1007/BF02828242

Rajasekaran, K., & Nirmala, K. (2012). Classification and Importance of Intrusion Detection System. *International Journal of Computer Science and Information Security, 10*(8), 44-47.

Sadeghi, Z., & Bahrami, A. S. (2013, 28-30 May 2013). *Improving the speed of the network intrusion detection.* Paper presented at the Information and Knowledge Technology (IKT), 2013 5th Conference on.

Spathoulas, G. P., & Katsikas, S. K. (2010). Reducing false positives in intrusion detection systems. *Computers & Security, 29*(1), 35-44. doi: http://dx.doi.org/10.1016/j.cose.2009.07.008

Spathoulas, G. P., & Katsikas, S. K. (2013). Enhancing IDS performance through comprehensive alert post-processing. *Computers & Security, 37*, 176-196. doi: http://dx.doi.org/10.1016/j.cose.2013.03.005

Wang, W., & Bridges, S. M. (2000). *Genetic Algorithm Optimization of Membership Functions for Mining Fuzzy Association Rules* Paper presented at the International Joint Conference on Information Systems, Fuzzy Theory and Technology Conference, Atlantic City, N.J.