# Improving Effectiveness of Honeypots: Predicting Targeted Destination Port Numbers During Attacks Using J48 Algorithm

Tanveer Gangabissoon(✉), Amaan Nathoo, Rakshay Ramhith, Bhooneshwar Gopee, and Girish Bekaroo

School of Science and Technology, Middlesex University Mauritius, Coastal Road, Uniciti, Flic-en-Flac, Mauritius
{TG368, AN861, RR678, BG337}@live.mdx.ac.uk, g.bekaroo@mdx.ac.mu

**Abstract.** During recent years, there has been an increase in cyber-crime and cybercriminal activities around the world and as countermeasures, effective attack prevention and detection mechanisms are needed. A popular tool to augment existing attack detection mechanisms is the Honeypot. It serves as a decoy for luring attackers, with the purpose to accumulate essential details about the intruder and techniques used to compromise systems. In this endeavor, such tools need to effectively listen and keep track of ports on hosts such as servers and computers within networks. This paper investigates, analyzes and predicts destination port numbers targeted by attackers in order to improve the effectiveness of honeypots. To achieve the purpose of this paper, the J48 decision tree classifier was applied on a database containing information on cyber-attacks. Results revealed insightful information on key destination port numbers targeted by attackers, in addition to how these targeted ports vary within different regions around the world.

**Keywords:** Destination port · Honeypot · Prediction · J48 algorithm · Decision tree

## 1 Introduction

During the previous decade, cybercrime and cybercriminal activities have escalated significantly and this ranges from infected end-user computers to compromised web-servers that surreptitiously infect unsuspecting visitors [1]. Statistics showed that most cyber-attacks with monetary gain motive were reported in 2014 with hacktivism, cyber-espionage and cyber warfare between rival cyber-crews being the most prominent cyber-crimes since the past decade [2]. With the continuously growing number of Internet users, cyber-attacks are expected to increase as cyber-crime and cyber-security is estimated to cost the world $6 trillion annually by 2021 [3]. As such, it becomes important to reduce the treats globally through effective attack prevention and detection mechanisms. A popular tool to augment existing attack detection mechanisms is the honeypot and using such systems, new attacks could be unveiled, assault patterns could be uncovered, and the precise thought processes of the intruder could be studied [4, 5].


P. Fleming et al. (Eds.): ELECOM 2018, LNEE 561, pp. 225–234, 2019.
https://doi.org/10.1007/978-3-030-18240-3_21

Honeypots are traps designed to detect attempts of unauthorized infiltration and use of an information system. The main purpose of a honeypot is to improve cyber security by not only detecting and preventing attacks but also by keeping track of the perpetrator's activities, understand methodologies used, to eventually develop counterattacks and save forensic information about attackers for prosecution [5]. Along with ensuring a secure network, the information gathered could be used for law enforcement. Furthermore, compared to the traditional network security techniques like firewalls, intrusion detection systems and encryption, the use of honeypots is considered a more proactive, cost effective and promising approach to detect and battle against network security threats [6].

For the correct operation of honeypots and to correctly trace back the attacker, such systems need to effectively listen to ports on hosts such as servers and computers within networks [7]. A port refers to a part of a network address, which identifies a specific process/service in a computer and messages can be transmitted through the network to communicate with the process on a port number. These ports utilize certain protocols like Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) to arrange for data to be transferred. Ports are divided into three different ranges, namely, well-known ports, registered ports and dynamic/private ports. It is important to keep track of port numbers to determine which process or service (e.g. email, world-wide web or remote access services) is utilizing a particular port and what type of protocol is being used. This provides information on where an issue occurs. In reference to honeypots and taking into account mainly destination ports, organizations are able to find the most targeted ports, hence find what processes are deemed vulnerable and what attackers look for in the system. It can be said that predicting port numbers is crucial to understanding where the next most likely attack will occur thereby enabling organizations to prioritize security and take actions to prevent or deflect any security threats in time [7].

Although it is essential to track and forecast port numbers utilized by honeypots, limited research has been undertaken in this direction. As related works, a previous study presented the design and real-world evaluation of an innovative social-honeypot based approach to social spam detection [8]. In the same work, machine learning based classifiers were developed in order to identify previously unknown spammers with high precision and a low rate of false positives. Another study modelled the interaction between honeypots and bot-masters by a Markov Decision Process in order to determine the honeypots optimal policy for responding to the commands of bot-masters [9]. Another paper investigated the use of an automated state machine in conjunction with a client honeypot towards providing a powerful framework to organize monitoring of malware activity and record the results [10]. As such, limited work has been conducted regarding analysis or predicting port numbers utilized by attackers so that effectiveness of honeypots could be improved.

Taking cognizance of this limitation, this paper investigates, analyzes and predicts destination port numbers targeted by attackers in order to improve the effective of honeypots. This work is intended to help network administrators in different countries understand targeted port numbers during attacks to eventually implement network security measures against cyber-attacks.

This paper is organized in five key sections. After the introductory section in the first part, the theoretical background is provided, which describes the techniques and algorithms used for prediction. The third section describes the methodology for achieving the purpose of this paper and the results are presented in the fourth section. Finally, Sect. 5 presents the concluding remarks in addition to future research directions.

## 2 Theoretical Background

Amongst data mining tasks, classification and prediction are popular ones for knowledge discovery and help in decision-making [11]. The classification technique in data mining classifies data according to their classes by putting data in single group that belongs to a common class [12]. Amongst the different classifiers, decision trees or classification trees are commonly used for classifying instances or objects into a set of classes with assigned values or types based on their labels/attributes [12]. The internal nodes of a decision tree represent different attributes; the branches among the nodes describe possible values that these attributes can have in the given samples, while the terminal/last nodes give the final value/classification of the dependent variable.

Amongst the classifiers, the J48 algorithm is a popular and powerful one due to its high accuracy in decision-making [13]. It is an open source Java implementation of the C4.5 algorithm. The J48 decision tree classifier classifies items based on the attribute values of a supplied training set [14]. This algorithm works in a way that when it comes across a set of items, it finds what attributes discriminate the numerous cases clearly. It can produce both decision trees and result-sets in order to improve prediction accuracy [15]. Furthermore, the resulting classification rules generated by this algorithm is human readable and easy to understand thereby simplifying interpretation [11]. This classifier has been used in various studies including landslide susceptibility mapping [16] and network packet classification for use by network-based intrusion detection systems [17], amongst others.

In terms of operation, this algorithm creates a decision tree by using the divide-and-conquer algorithm where if all cases within a set belong to the same class or the set is small, then the tree is a leaf labelled with the most frequent class [11]. Within the same set, a test is chosen on single attribute with two or more outcomes and is made the root of the tree with one branch for each outcome of the test, before partitioning the set into different subsets according to the outcome for each case. The same procedure is then applied recursively to each subset. As such, this algorithm generates a decision tree where each node splits classes based on information gain and that the attribute with highest normalized information gain is utilized as splitting criteria [16].

## 3 Methodology

In order to achieve the purpose of this paper and to predict destination port numbers targeted by attackers by using J48 algorithm, an analysis on the Amazon Web Services (AWS) honeypot data [17] was performed. It is an open-source database containing

information on cyber-attacks/attempts and was chosen due to its relevance to the purpose of the paper, while other relevant open-source datasets were unavailable. In order to prepare dataset for analysis, the preprocessing stage consisted of firstly analyzing the attributes in order to determine their usefulness. In this process, a few attributes were removed to optimize the data and these included latitude and longitude of the attack. Following this clean-up, the attributes listed in Table 1 were left.

**Table 1.** Description of attributes.

| Number | Attribute | Description |
|---|---|---|
| 1 | Host | The region the computer connected to a network |
| 2 | Source (src) | The IP address of the origin |
| 3 | Proto | The protocol used e.g. TCP, UDP |
| 4 | Source port (spt) | The origin port number |
| 5 | Destination port (dpt) | The destination port number |
| 6 | Srcstr | The source string shows the source number of the source user |
| 7 | Country | The country involved |
| 8 | Country code (cc) | The 2 letter code to represent the corresponding country |
| 9 | Locale | Locale is the location/region in the particular country e.g. USA is the country and Texas is the locale |

The next stage involved preparing the data for training and evaluation in Weka. The Waikato Environment for Knowledge Analysis or Weka is a suite of machine learning software written in Java and is commonly used for data mining [18]. This tool was chosen for data analysis since it is free and that it has been used in different similar studies. Preparing the data for Weka environment started by converting the data into the ARFF format. Moreover, due to the fact that J48 did not support the default data types assigned to the attributes, changes had to be made and all of the attributes were assigned nominal values. For this, records in the dataset were then modified through a conversion software where every attribute was specialized into nominal data types. Furthermore, records containing null values were removed in order to further optimize the dataset. Following optimization, 20,000 data points were available for training and evaluation. For training the J48 algorithm on Weka 90% of the records were utilized in order to ensure enough data was utilized in order to train the algorithm since the J48 algorithm works better with a larger training set [21]. In the training process, all the selected attributes defined in Table 1 were utilized and the J48 classifier produced analysis of the training dataset and classification rules. Furthermore, during the training process, the percentage split feature was applied, to split the dataset into two parts each dependent on the percentage specified from the user. In addition, only top 10 ports were targeted thereby reducing the number of instances in order to improve the effectiveness of prediction. Focusing only on top 10 ports also meant removal of records related to uncommon port numbers which were used less than 5 times during attacks, so as to

obtain a better structured classification tree as outcome. Finally, the remaining 10% of the records were used for evaluation and interpretation of the classification rules. The aim of this evaluation process is to determine the accuracy of classification rules for prediction and to identify the important attributes and rules [11].

During the analysis process, different challenges were faced where the major one was during the preprocessing stage especially for treating the null values present within the dataset. For this, a software had to be written in order to filter the dataset line by line in order to remove these lines. Another challenge encountered was massive amount of data in the dataset caused stutters in WEKA and the training/evaluation processes were thus lengthy.

## 4 Results and Discussions

Using the previously defined methodology, evaluation was conducted on 2044 records (10% of the dataset) and the extracted summary of the analysis for the J48 algorithm from Weka is given in Fig. 1. From these 2044 records, 68.1% were correctly classified as compared to 31.9% instances, which were incorrectly classified. This high percentage for the correctly classified instances also implies that the values are accurate enough to perform the prediction. On the other hand, the incorrectly classified instances were particularly due to some attacks that originated from countries with reduced number of attacks within the dataset.

| | | |
|---|---|---|
| Correctly Classified Instances | 1392 | 68.1018 % |
| Incorrectly Classified Instances | 652 | 31.8982 % |
| Kappa statistic | 0.6234 | |
| Mean absolute error | 0.0028 | |
| Root mean squared error | 0.0386 | |
| Relative absolute error | 43.3377 % | |
| Root relative squared error | 68.3484 % | |
| Total Number of Instances | 2044 | |

**Fig. 1.** Extracted summary by the J48 classifier

The reliability of results obtained was further statistically assessed and extract of results are given in Fig. 2. In the same figure, true-positive (TP) represents a case where the condition detected to be true is actually true. On the other hand, false positive (FP) is a case where the condition detected to be true is actually false. In addition, precision is the number of instances that are actually true, compared to total number of instances classified. Finally, class represents the destination port number, which is the most significant attribute analyzed in this study. Findings in Fig. 2 show a significantly higher true positive as compared to false positive for the top ten common port numbers, with some good precision and f-measure values. As such, the accuracy of prediction by the J48 classifier could be considered as reliable.

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---|---|---|---|---|---|---|---|---|---|
| | 0.587 | 0.002 | 0.968 | 0.587 | 0.731 | 0.740 | 0.930 | 0.700 | 22 |
| | 0.057 | 0.007 | 0.250 | 0.057 | 0.092 | 0.103 | 0.843 | 0.155 | 23 |
| | 0.143 | 0.000 | 1.000 | 0.143 | 0.250 | 0.372 | 0.882 | 0.269 | 53 |
| | 0.602 | 0.009 | 0.831 | 0.602 | 0.698 | 0.689 | 0.946 | 0.657 | 80 |
| | 0.500 | 0.003 | 0.833 | 0.500 | 0.625 | 0.638 | 0.921 | 0.593 | 135 |
| | 0.829 | 0.007 | 0.944 | 0.829 | 0.883 | 0.871 | 0.968 | 0.880 | 445 |
| | 0.962 | 0.464 | 0.566 | 0.962 | 0.713 | 0.511 | 0.894 | 0.789 | 1433 |
| | 0.132 | 0.002 | 0.857 | 0.132 | 0.229 | 0.323 | 0.824 | 0.310 | 3306 |
| | 0.444 | 0.019 | 0.747 | 0.444 | 0.557 | 0.538 | 0.855 | 0.566 | 3389 |
| | 0.446 | 0.004 | 0.891 | 0.446 | 0.594 | 0.614 | 0.884 | 0.573 | 8080 |
| Weighted Avg. | 0.668 | 0.184 | 0.733 | 0.668 | 0.633 | 0.564 | 0.898 | 0.664 | |

**Fig. 2.** Reliability of data

Following reliability tests, analysis was conducted on the commonly used destination port numbers by attackers on the same evaluation data. Results are given in Table 2 where out of the 282 different port numbers that formed part of the 2044 records, the top 10 commonly ones are given. Amongst, port number 1433 was found to be the most targeted one by attackers and this port is the default one for SQL Server. This also shows that attackers are particularly interested in attacking database servers so as to obtain various pieces of meaningful information such as credit card numbers, credentials, transaction details and personal details of clients, amongst others. Similarly, the default port number of MySQL, notably 3306 was found to be amongst the leading destination port numbers targeted by attackers for the same reasons mentioned. On the second position, the port number 3389, which is the default port number for Microsoft WBT Server was found. This server is used for Windows Remote Desktop and remote assistance connections through which attackers can potentially connect to other computers within the network in order to extract meaningful information. Likewise, port numbers 22 (SSH) and 23 (Telnet) were also found amongst the most targeted ones and are used for the same purpose of connecting to computers within the same network. In addition, ports 8080 and 80 were found amongst the top 5 targeted ports principally used for the web. Port 80 is reserved for HTTP and attackers target this port in order to gain administrative access to a website or to the web-server hosting it. In the same way, many web servers run on port 8080 and attackers target this port in order to gain administrative access. The remaining most common ports from the list included port 445 for Server Message Blocks over the Internet Protocol, 135 utilized for Remote Procedure Call) and 53 used by Domain Name System (DNS) servers, as listed in Table 2.

**Table 2.** Top 10 targeted destination port numbers

| Rank | Destination port number | Count |
|---|---|---|
| 1 | 1433 | 526 |
| 2 | 3389 | 123 |
| 3 | 8080 | 97 |
| 4 | 3306 | 94 |
| 5 | 80 | 90 |
| 6 | 445 | 72 |
| 7 | 22 | 71 |
| 8 | 23 | 53 |
| 9 | 135 | 24 |
| 10 | 53 | 7 |

Finally, the classification tree generated in Weka is depicted in Fig. 3 to show how the targeted destination port numbers vary across different regions. In the same figure, the leaves represented by the rectangular boxes in the final level represent the destination port numbers targeted by attackers whilst the ovals identify labels given. The branches at the first level show host of the honeypot, and the branch at the second level shows the protocol used. The most common protocols involved included the Transmission Control Protocol (TCP), User Datagram Protocol (UDP) and Internet Control Message Protocol (ICMP). The leaves displaying the ports also show the count, which represent how many times it has been targeted.

Findings reveal that the port 1433 is the most common one being targeted in most regions. However, this is not the case for Europe, Australia and East of US. In Europe and Australia, port 3388 for Microsoft WBT Server was found to be the mostly targeted one also highlighting the noticeable target for remote desktop by attackers. On the other hand, port 80 is targeted within the Eastern region of US, as shown in the generated tree, particularly to obtain administrative access to a website or to the web-server hosting it, as mentioned earlier. As such, in order to improve effectiveness of honeypots and to better lure attackers, network administrators could configure honeypots to listen to port numbers revealed as findings of this study.

The study is however undermined by the limitations of the J48 algorithm where its run-time complexity only matches to the tree depth, which in turn cannot exceed the number of attributes [12]. In addition, some part of data from the dataset was removed so as to optimize the training and evaluation set. In this process, some essential information could have been lost while also removing important port numbers.

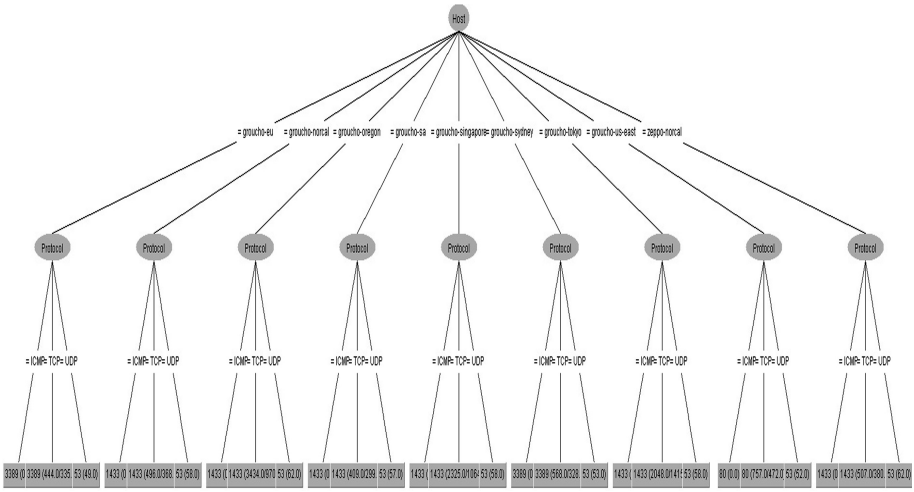


**Fig. 3.** Tree showing the most targeted port numbers

## 5 Concluding Remarks

This paper investigated, analyzed and predicted destination port numbers targeted by attackers by applying the J48 decision tree classifier on an open-source database containing information on cyber-attacks. The algorithm was trained on Weka by using 90% of the dataset since the algorithm needs a large training set. The remaining 2044 data-points were used to evaluate the decision tree, out of which 68.1% records were correctly classified as compared to 31.9% instances incorrectly classified. This high percentage for the correctly classified instances in addition to reliability tests conducted showed that the values are accurate enough to perform the prediction. Results showed that database related ports, notably 1433 for SQL Server and 3306 for MySQL were amongst the most targeted ones by attackers, who are particularly interested in obtaining meaningful information from compromised database servers. Similarly, ports for remote desktop connection, secure shell and telnet were among the mostly targeted destination port numbers. In the decision tree generated, findings reveal that the targeted port numbers vary slightly across different regions, although destination port number 1433 remain the dominant one targeted. In order to improve effectiveness of their honeypots, companies can better perform configurations to target the common destination port numbers investigated in this study. In other words, this could potentially help honeypots to be better prepared to detect the potential ports being targeted and therefore secure those ports more effectively from attackers.

As future work, the same data set could be further analyzed by varying the percentage of records for the training set and evaluation to assess associated effects on the decision tree. Furthermore, the attributes removed for optimization could be re-integrated to assess any change in the resulting decision tree since small variation in data can lead to different decision trees. Moreover, further work is also needed to better address the scattered port numbers in the dataset.

## References


1. Jhaveri, M., Cetin, O., Gañán, C., Moore, T., Eeten, M.: Abuse reporting and the fight against cybercrime. ACM Comput. Surv. (CSUR) **49**(4), 68 (2017)
2. The Windows Club, "What are Honeypots and how can they secure computer systems" (2018). http://www.thewindowsclub.com/what-are-honeypots. Accessed 11 Apr 2014
3. Harrison, J.: Honeypots: The sweet spot in network security (2018). https://www.computerworld.com/article/2573345/security0/honeypots–the-sweet-spot-in-network-security.html. Accessed 28 Apr 2018
4. Yang, Y., Yang, H., Mi, J.: Design of distributed honeypot system based on intrusion tracking. In: 2011 IEEE 3rd International Conference on Communication Software and Networks (ICCSN) (2011)
5. Zakari, A., Lawan, A., Bekaroo, G.: Towards improving the security of low-interaction honeypots: insights from a comparative analysis. In: International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (2016)
6. Duong, B.: Comparisons of attacks on honeypots with those on real networks, Naval Postgraduate School, Monterey, California (2006)
7. Kreibich, C., Crowcroft, J.: Honeycomb: creating intrusion detection signatures using honeypots. ACM SIGCOMM Comput. Commun. Rev. **34**(1), 51–56 (2004)
8. Lee, K., Caverlee, J., Webb, S.: Uncovering social spammers: social honeypots + machine learning. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (2010)
9. Hayatle, O., Otrok, H., Youssef, A.: A markov decision process model for high interaction honeypots. Inf. Secur. J.: Glob. Perspect. **22**(4), 159–170 (2013)
10. Alosefer, Y., Rana, O.: Automated state machines applied in client honeypots. In: 2010 5th International Conference on Future Information Technology (FutureTech) (2010)
11. Jantan, H., Hamdan, A., Othman, Z.: Human talent prediction in HRM using C4. 5 classification algorithm. Int. J. Comput. Sci. Eng. **2**(8), 2526–2534 (2010)
12. Neeraj, B., Girja, S., Ritu, D., Manisha, M.: Decision tree analysis on j48 algorithm for data mining. Int. J. Adv. Res. Comput. Sci. Softw. Eng. (JARCSSE) **3**(6), 1114–1119 (2013)
13. Amin, R., Sibaroni, Y.: Implementation of decision tree using C4. 5 algorithm in decision making of loan application by debtor (Case study: Bank pasar of Yogyakarta Special Region). In: 2015 3rd International Conference on Information and Communication Technology (ICoICT) (2015)
14. Patil, T., Sherekar, S.: Performance analysis of Naive Bayes and J48 classification algorithm for data classification. Int. J. Comput. Sci. Appl. **6**(2), 256–261 (2013)
15. Delen, D., Walker, G., Kadam, A.: Predicting breast cancer survivability: a comparison of three data mining methods. Artif. Intell. Med. **34**(2), 113–127 (2005)
16. Bui, D., Ho, T., Revhaug, I., Pradhan, B., Nguyen, D.: Landslide susceptibility mapping along the national road 32 of Vietnam using GIS-based J48 decision tree classifier and its ensembles. In: Cartography from Pole to Pole. Springer, Heidelberg (2014)
17. Sahu, S., Mehtre, B.: Network intrusion detection system using J48 decision tree. In: 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (2015)
18. Wu, X., Kumar, V., Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G., Ng, A., Liu, B., Philip, S., Zhou, Z.: Top 10 algorithms in data mining. Knowl. Inf. Syst. **14**(1), 1–37 (2008)

19. Jacobs, J., Rudis, B.: Kaggle (2018). https://www.kaggle.com/casimian2000/aws-honeypot-attack-data. Accessed 10 Apr 2018
20. Holmes, G., Donkin, A., Witten, I.: Weka: a machine learning workbench. In: Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems (1994)
21. Salzberg, S.: C4. 5: programs for machine learning by j. ross quinlan. Mach. Learn. **16**(3), 235–240 (1994)