SJIF Impact Factor: 7.001| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 5 | Issue: 12 | December 2020 - Peer Reviewed Journal

A STUDY ON ADVANCED BOTNETS DETECTION IN VARIOUS COMPUTING SYSTEMS USING MACHINE LEARNING TECHNIQUES

K. Vamshi Krishna
Assistant Professor, Department of Computer Science and Engineering,
Narasaraopet Engineering College, Narasaraopet, Guntur, India

Article DOI: https://doi.org/10.36713/epra5902

ABSTRACT

Due to the rapid growth and use of Emerging technologies such as Artificial Intelligence, Machine Learning and Internet of Things, Information industry became so popular, meanwhile these Emerging technologies have brought lot of impact on human lives and internet network equipment has increased. This increment of internet network equipment may bring some serious security issues. A botnet is a number of Internet-connected devices, each of which is running one or more bots.The main aim of botnet is to infect connected devices and use their resource for automated tasks and generally they remain hidden. Botnets can be used to perform Distributed Denial-of-Service (DDoS) attacks, steal data, send spam, and allow the attacker to access the device and its connection. In this paper we are going to address the advanced Botnet detection techniques using Machine Learning. Traditional botnet detection uses manual analysis and blacklist, and the efficiency is very low. Applying machine learning to batch automatic detection of botnets can greatly improve the efficiency of detection. Using machine learning to detect botnets, we need to collect network traffic and extract traffic characteristics, and then use X-Means, SVM algorithm to detect botnets. According to the difference of detection features, botnet detection based on machine learning technology is divided into network traffic analysis and correlation analysis-based detection technology.

KEYWORDS: *Botnet, Study, Security, Internet-network, Machine Learning, Techniques.*

I. INTRODUCTION

A Botnet is a number of devices connected to internet each of which runs one or more bots [8]. A botnet could be a logical collection of Internet-connected devices like computers, smartphones or IoT devices whose security are broken and management ceded to a 3rd party. every compromised device, referred to as a "bot", is made once a tool is penetrated by software package from a malware (malicious software) distribution. The controller of a botnet is ready to direct the activities of those compromised computers through communication channels fashioned by standards-based network protocols, like IRC and machine-readable text Transfer Protocol (HTTP). Botnets are progressively rented out by cyber criminals as commodities for a spread of functions. To accomplish malicious activities over internet connected devices Botnet can be used very effectively. The main reason behind of getting many security problems over internet is Botnets [9].The most common attack created by Botnets is Denial of service(DoS) [1].Botnets are very emerging Malwares compared with other computer malwares and they may generate most serious problems at the same time acts like a threat to cyber security[10]. Among the various varieties of malware, botnets are rising because the most serious threat against cyber-security as they supply a distributed platform for many illegal activities like launching distributed denial of service attacks against essential targets, malware dissemination, phishing, and click on fraud. The shaping characteristic of botnets is that the use of command and management channels through that they will be updated and directed. Recently, botnet detection has been a motivating analysis topic associated with cyber-threat and cyber-crime difficulty. Botnets are most dangerous network attacks [11]. The study of computer algorithms is called as Machine Learning (ML), Machine Learning

SJIF Impact Factor: 7.001| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 5 | Issue: 12 | December 2020 - Peer Reviewed Journal

is the sub branch in the field of Artificial Intelligence whereas Artificial Intelligence is the sub branch in the field of Computer Science [12]. Machine Learning algorithms creates a model with the help of training data, sometime it is referred to as sample data to make prediction and to perform decision making [13]. To do anything with computers it is often required programming but by introducing a model created by machine learning algorithms would avoid programming and do what programming can able to do without the help of programming. Machine Learning algorithms can be used widely in various real-world applications such as Email filtering, Computer networks, Natural language processing, Search Engines, Telecommunications, Internet fraud detection and DNA sequence classification Three types of Learning algorithms are present Supervised, Unsupervised and reinforcement. Machine Learning ML is widely used multidisciplinary field which uses various training models and algorithms to predict, Classify and analyse any statistical data by the uses of computer science algorithms [14]. In this paper we are going to present Machine Learning Techniques, Application and Research Issues towards Botnet detection. We are going to use two algorithms one is X-Means algorithm and the other one is Support Vector Method (SVM) for achieving the above said.

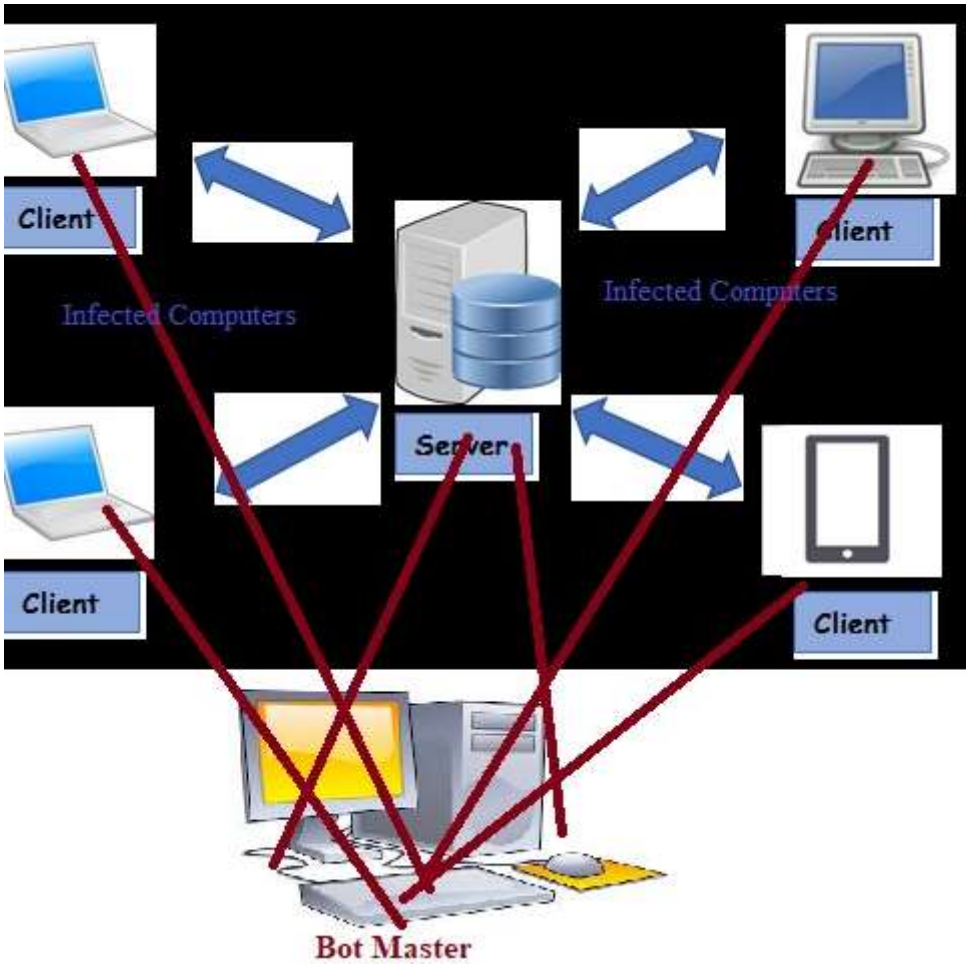

Fig-1 Botnet Scenario

SJIF Impact Factor: 7.001| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 5 | Issue: 12 | December 2020 - Peer Reviewed Journal

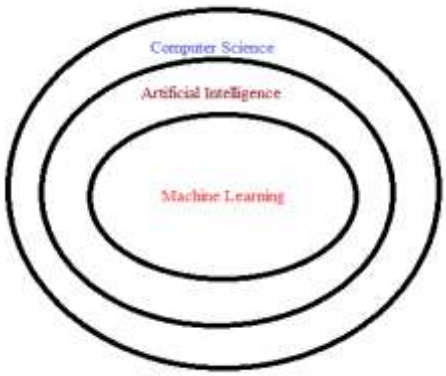

Fig-2 Machine Learning Vs AI and CS

II. OBJECTIVE OF THE PAPER

The main aim of this paper is to detect Botnets using machine learning algorithms in various connected devices. Machine Learning approach uses X-Means and SVM algorithms to identify Botnet in a Network of connected devices-Means is and advanced approach than K-Means in data mining at the same time on the other hand SVM-Support Vector Method is a supervised machine learning algorithm which can be used for classification and regression problem.

III. BACKGROUND/SHORT DESCRIPTION

Due to the ascension and use of rising technologies like AI, Machine Learning and net of Things, data business became thus widespread, meantime these rising technologies have brought heap of impact on human lives and net network instrumentality has magnified. This increment of net network instrumentality could bring some serious security problems. A botnet could be a range of Internet-connected devices, every of that is running one or a lot of bots. The main aim of botnet is to infect connected devices and use their resource for automatic tasks and usually they continue to be hidden. Botnets will be accustomed perform Distributed Denial-of-Service (DDoS) attacks, steal data, send spam, and permit the assailant to access the device and its association.

Table-1 Important Terms:

Sno	Term	Description
1	Botnet	A botnet is a number of Internet-connected devices, each of which is running one or more bots
2	BotMaster	A botmaster is a person who operates the command and control of botnets
3	Bot	An Internet bot, web robot, robot or simply bot, is a software application that runs automated tasks over the Internet
4	C&C Server	A command and control server (C&C server) are a computer that issues directives to digital devices that have been infected with rootkits or other types of malware, such as ransomware.
5	Peer2Peer	Peer-to-peer computing or networking is a distributed application architecture that partitions tasks or workloads between peers.
6	Internet Relay Chat IRC	Internet Relay Chat is an application layer protocol that facilitates communication in the form of text.
7	HTTP	The Hypertext Transfer Protocol is an application layer protocol for distributed, collaborative, hypermedia information systems.
8	SPAM	irrelevant or unsolicited messages sent over the internet, typically to a large number of users, for the purposes of advertising, phishing, spreading malware, etc.
9	DDoS	In computing, a denial-of-service attack is a cyber-attack in which the perpetrator seeks to make a machine or network resource unavailable to its intended users by temporarily or indefinitely disrupting services of a host connected to the Internet

SJIF Impact Factor: 7.001| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 5 | Issue: 12 | December 2020 - Peer Reviewed Journal

IV. RELATED WORK

Detecting botnets in an exceedingly network is crucial as a result of bot's impact various areas like cyber security, finance, health care, enforcement, and more. Botnets are getting a lot of refined and dangerous day-by-day, associated most of the prevailing rule based mostly and flow-based detection ways might not be capable of police investigation bug activities in an economical and effective manner. Hence, planning a strong and quick botnet detection technique is of high significance. Malicious traffic, like DDoS attack and botnet communications, refers to traffic that's generated for the aim of distressful net networks or harming bound networks, servers, or hosts. As malicious traffic has been perpetually evolving in terms of each quality and amount, there are several researches fighting against it.

K-Means Clustering

K-Means Clustering is very popular clustering generally, but it is suffering from the following drawbacks

1. Poor in Scaling
2. Search is prone to Local Minima
3. Its dependents on input supplied by user.

X-means clustering and Support Vector Machine (SVM):

In statistics and data mining, X-means clustering is a variation of k-means clustering that refines cluster assignments by repeatedly attempting subdivision, and keeping the best resulting splits, until a criterion such as the Akaike information criterion (AIC) or Bayesian information criterion (BIC) is reached.

Determining the number of clusters in a data set, a quantity often labelled k as in the k-means algorithm, is a frequent problem in data clustering, and is a distinct issue from the process of actually solving the clustering problem.

For a certain class of clustering algorithms (in particular k-means, k-medoids and expectation–maximization algorithm), there is a parameter commonly referred to as k that specifies the number of clusters to detect. Other algorithms such as DBSCAN and OPTICS algorithm do not require the specification of this parameter; hierarchical clustering avoids the problem altogether.

The correct choice of k is often ambiguous, with interpretations depending on the shape and scale of the distribution of points in a data set and the desired clustering resolution of the user. In addition, increasing k without penalty will always reduce the amount of error in the resulting clustering, to the extreme case of zero error if each data point is considered its own cluster (i.e., when k equals the number of data points, n). Intuitively then, the optimal choice of k will strike a balance between maximum compression of the data using a single cluster, and maximum accuracy by assigning each data point to its own cluster. If an appropriate value of k is not apparent from prior knowledge of the properties of the data set, it must be chosen somehow. There are several categories of methods for making this decision.

Support Vector Machine:

K-means algorithm can be used to find features in Support Vector Machine SVM, A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labelled training data for each category, they're able to categorize new text. So, you're working on a text classification problem.

SJIF Impact Factor: 7.001| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 5 | Issue: 12 | December 2020 - Peer Reviewed Journal

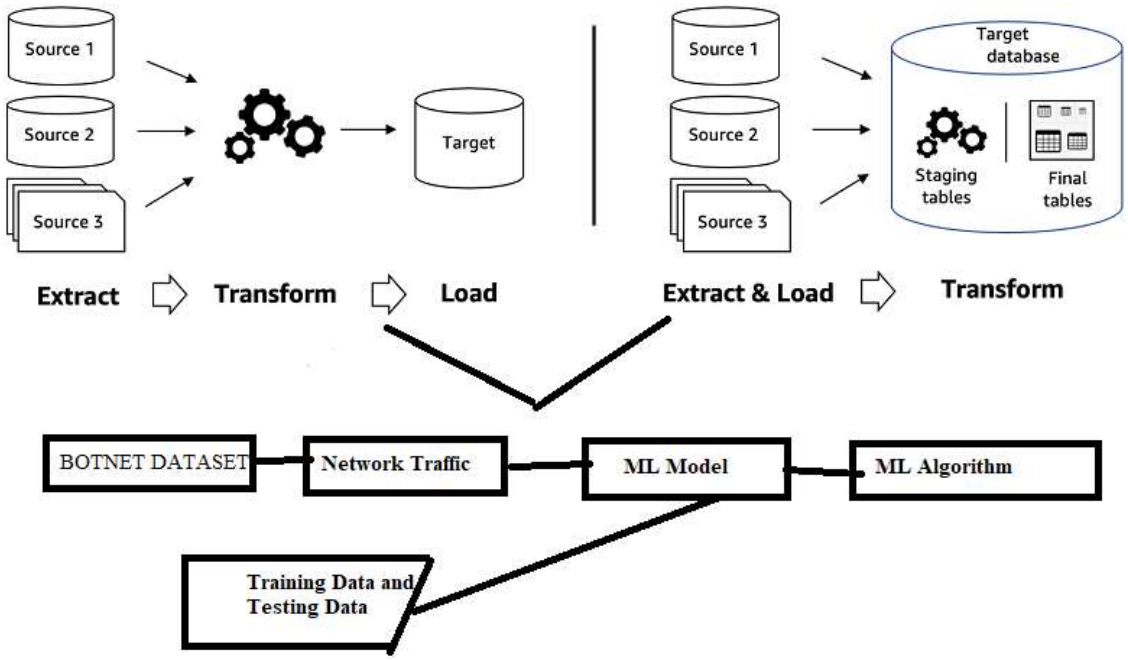

Fig-3 Botnet Analysis Using Machine Learning algorithms

Botnet Example dataset:

Id	Duration(hrs)	# Packets	#NetFlows	Size	Bot	#Bots
1	6.15	71,971,482	2,824,637	52GB	Neris	1
2	4.21	71,851,300	1,808,123	60GB	Neris	1
3	66.85	167,730,395	4,710,639	121GB	Rbot	1
4	4.21	62,089,135	1,121,077	53GB	Rbot	1
5	11.63	4,481,167	129,833	37.6GB	Virut	1
6	2.18	38,764,357	558,920	30GB	Menti	1
7	0.38	7,467,139	114,078	5.8GB	Sogou	1
8	19.5	155,207,799	2,954,231	123GB	Murlo	1
9	5.18	115,415,321	2,753,885	94GB	Neris	10
10	4.75	90,389,782	1,309,792	73GB	Rbot	10
11	0.26	6,337,202	107,252	5.2GB	Rbot	3
12	1.21	13,212,268	325,472	8.3GB	NSIS.ay	3
13	16.36	50,888,256	1,925,150	34GB	Virut	1

SJIF Impact Factor: 7.001| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 5 | Issue: 12 | December 2020 - Peer Reviewed Journal

K-Means Algorithm:

The most common algorithm uses an iterative refinement technique. Due to its ubiquity, it is often called "the *k*-means algorithm"; it is also referred to as Lloyd's algorithm, particularly in the computer science community. It is sometimes also referred to as "naive *k*-means", because there exist much faster alternatives.

Given an initial set of *k* means $m_1^{(1)},\ldots,m_k^{(1)}$ (see below), the algorithm proceeds by alternating between two steps:

Assignment step: Assign each observation to the cluster with the nearest mean: that with the least squared Euclidean distance. (Mathematically, this means partitioning the observations according to the Voronoi diagram generated by the means.)

$$S_i^{(t)} = \left\{x_p : \left\|x_p - m_i^{(t)}\right\|^2 \le \left\|x_p - m_j^{(t)}\right\|^2 \ \forall j, 1 \le j \le k\right\},$$

where each x_p is assigned to exactly one $S^{(t)}$, even if it could be assigned to two or more of them.

Update step: Recalculate means (centroids) for observations assigned to each cluster.

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \in S_i^{(t)}} x_j$$

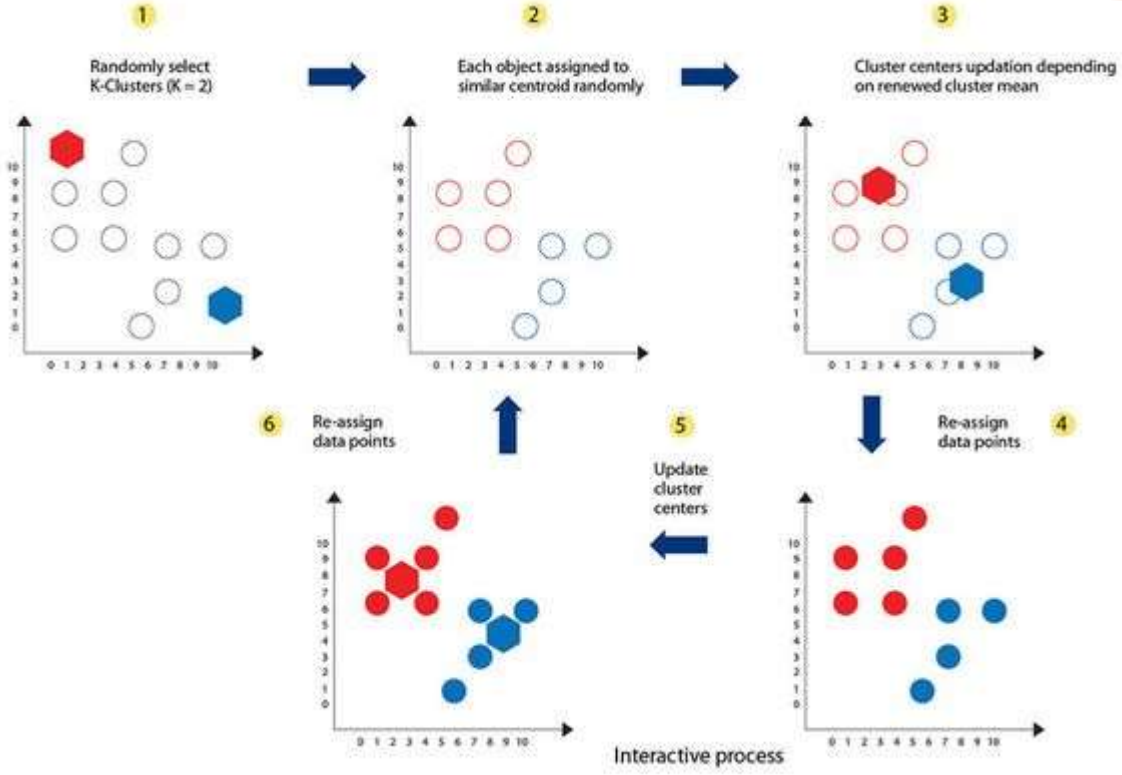

Fig-4 X-Means Clustering and Feature Extraction by SVM

SJIF Impact Factor: 7.001| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 5 | Issue: 12 | December 2020 - Peer Reviewed Journal

V. **Expected results:** The main goal of SVM is to classify data and create a classification model. This created model is used to identify Botnet Behaviours, of course the model build was done by submitting a botnet behavioural dataset to classification algorithm, to achieve clustering of data, it uses X-means clustering algorithm. Finally the system may give optimal result by the use of SVM and X-Means Algorithms.

VI. CONCLUSION AND FUTURE WORK

Botnet detection is a challenging task, since the creators of botnets continue to adopt innovative means in creating botnets. In this paper, X-Means clustering algorithm and Support Vector machine SVM is implemented in to detect bot. it is concluded that the proposed algorithms show high recall rate for all the datasets compared to k-means algorithm which means returned most of the relevant results. The future work of this study is to take any real time Botnet dataset and Identify its Behaviour and to take necessary preventive actions to safe guard the system.

REFERENCES

1. *Vogt, R., Aycock, J., & Jacobson Jr, M. J. (2007, February). Army of Botnets. In NDSS.*
2. *Bertino, E., & Islam, N. (2017). Botnets and internet of things security. Computer, 50(2), 76-79.*
3. *Ianelli, N., & Hackworth, A. (2005). Botnets as a vehicle for online crime. CERT Coordination Center, 1(1), 28.*
4. *McCarty, B. (2003). Botnets: Big and bigger. IEEE Security & Privacy, 1(4), 87-90.*
5. *Strayer, W. T., Walsh, R., Livadas, C., & Lapsley, D. (2006, November). Detecting botnets with tight command and control. In Proceedings. 2006 31st IEEE Conference on Local Computer Networks (pp. 195-202). IEEE.*
6. *Schiller, C., & Binkley, J. R. (2011). Botnets: The killer web applications. Elsevier.*
7. *Hachem, N., Mustapha, Y. B., Granadillo, G. G., & Debar, H. (2011, May). Botnets: lifecycle and taxonomy. In 2011 Conference on Network and Information Systems Security (pp. 1-8). IEEE.*
8. *[Available online]-https://en.wikipedia.org/wiki/Botnet.*
9. *Stone-Gross, B., Cova, M., Cavallaro, L., Gilbert, B., Szydlowski, M., Kemmerer, R., ... & Vigna, G. (2009, November). Your botnet is my botnet: analysis of a botnet takeover. In Proceedings of the 16th ACM conference on Computer and communications security (pp. 635-647).*
10. *Feily, M., Shahrestani, A., & Ramadass, S. (2009, June). A survey of botnet and botnet detection. In 2009 Third International Conference on Emerging Security Information, Systems and Technologies (pp. 268-273). IEEE.*
11. *Strayer, W. T., Lapsely, D., Walsh, R., & Livadas, C. (2008). Botnet detection based on network behavior. In Botnet detection (pp. 1-24). Springer, Boston, MA.*
12. *Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (1994). Machine learning. Neural and Statistical Classification, 13(1994), 1-298.*
13. *Williams, D., & Hill, J. (2005). U.S. Patent Application No. 10/939,288.*
14. *Zhang, X. D. (2020). Machine learning. In A Matrix Algebra Approach to Artificial Intelligence (pp. 223-440). Springer, Singapore.*