International Journal of Research in Advent Technology, Vol.6, No.9, September 2018
E-ISSN: 2321-9637
Available online at www.ijrat.org

Intrusions Detection by Using Automated Honey-pot Machine Learning Technology in an unstructured Big-Data

Rashid Hussain[1], Shivani Joon[2], Dr. Rajesh Kumar Tyagi[3]
Department of Mathematics and Computer Science, Sule Lamido University, Kafin Hausa, Jigawa State, Nigeria[1]
Research Scholar, Department of Computer Science, Starex University, Gurugram, India[2]
Director, Sudiksha Institute of Machine Learning and Artificial Intelligence, New Delhi[3]
rashid65_its@yahoo.com[1], shivanijoon333@gmail.com[2], drtyagi1234@gmail.com[3]

Abstract: Unauthorized Access Is Increases Day By Day As Per Use Of Digital Activity. So The Purpose Of This Research Paper Is To detect the unauthorized activity and provide data Confidentiality, Integrity, Authentication, QoS (Quality of Service), relevance, Privacy and Trust etc. A new method has been evolved using machine learning which contributed to efficient and cost effective implementation of Automated Honey-pot(IDS). This technique is providing security to complex digital data and reducing the probability of unauthorized access from the network architecture. In this paper, moving unstructured data has been analyzed and made some clusters with help of K-mean algorithm and after that a Naive Bayes classification has been applied for predicting the malign nodes. Contemporary Methods suffers from high computational complexity and our aim was to propose a method for reducing it and embed the innovative machine learning tools for detecting the unknown attacks within a peer networks. Due to this system, Confidentiality, Data integrity, QoS, Authentication, relevance, Privacy and Trust etc. increases manifold of unstructured big data within the networks. In this work, Firstly, we analyzed the well-known KDD CUP99 dataset for intrusion detection. In next Step, after learning intrusions automated system again transfer traffic back to the load balancer and then transfer it to the processor for checking. If IDS found some traffic anomalies, then transfer these anomalies to the Honey-pot server for advertising alarm among all nodes of the systems.This new proposed system is very accurate and give promising results as compared with previous techniques.

Keywords-*Big-Data, Machine-learning, K-Means, Naive-Bayes Classification, Honey-pot, IDS (Intrusion Detection System*
).

1. INTRODUCTION

At present over the digital world data is playing major role. Data is categorized into two categories; one is structure data and second is unstructured data. Structure data follows RDBMS rules and unstructured does not follow RDBMS rules. Structure data has pre- defined data model. Unstructured data is information that cannot be easily defined and it has no pre-defined data model. It is data such as videos, images, application log files or any data that does not easily fit into the traditional model. In digital world 20 per cent structured data and 80 per cent unstructured data are created. The reason of creating a huge amount of unstructured data is IoT; this will be one of the largest generators of unstructured data [1-3].

Unstructured data is non-relational data, which is growing heterogeneous sources around the digital world i.e. sensors, social sites, calls, bank transactions etc., which results a number of risks such as protection of data from unauthorized access, a disclosure of complex data and how to achieve high level security of data. Due to a huge amount and distributed nature of unstructured data; secure access, authenticity, integrity, consistency is an essential security challenges of unstructured data. Data is growing at very high speed and it is making difficult to manage unstructured data from unauthorized access because it includes complex information [4-5].

Lots of data is created in an unstructured way more than analysis in every second, so in future, it will more difficult to manage the security of complex data by the 40 to 50 per cent data increasing rate per year. International Data Corporation (IDC) predicts that volume of data is increasing very fast by the rate of 50 per cent per year and by 2020, data will have reached up to 45 Zettabytes. Figure 1 shows the data is growing at a rate of 50 per cent compound annual rate, reaching nearly 45 Zettabytes by 2020.

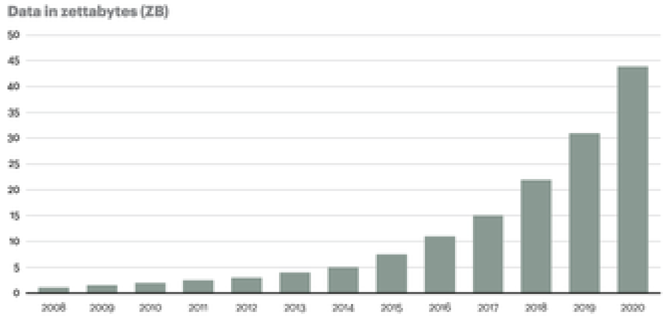

Figure 1: Worldwide Corporate Data Growth [2]

Data is increasing more than analysis every day from heterogeneous sources, which arise some major security issues and challenges related to data analysis, process and social control etc.

In this research work, we address major security issues and challenges of unstructured data in a network. Provide security to complex data from unauthorized access via machine learning tool such as k- means clustering algorithm.

2 SECURITY ISSUES OF UNSTRUCTURED DATA

There are many solutions are designed for scalability and performance but till now almost no more effective and efficient solutions are design for security of unstructured data. The biggest challenge in unstructured data is security. Here, we present some security concerns of unstructured data which are involve in data capture, data analysis, data storage, searching, sharing and visualization. Few major security concerns are listed below [4-7]

I. Architecture of unstructured data is highly distributed in nature with thousands of processing data nodes that runs data partitioned horizontally, replicated and distributed among various data nodes. So, there is possibility found intrusions in unstructured data architecture.

II. Data is growing at very high speed day-by-day; so, need to protect commercial information between different organizations, institutions to share their clients and users.

III. Variety of unstructured data is also varying increasingly everyday so, there is need to write various queries for fetching relevant data from the collected data stored.

IV. In present, speed of data creation is faster than speed of data analysis, so the computations of these highly distributed unstructured data must be successfully done their tasks continuously in real time without losing security, privacy and trusts of users.

V. Huge variety of unstructured data, it is difficult to move data in between different data nodes rather than their code. So, the security perspective, move the code is easier than move the data. But, increasing of data every second will become problem of sharing crucial information because of complicated encrypted code.

VI. In an unstructured data architecture, data is moving from one node to another node; due to this it is difficult to find out the exact location where data is stored among different data nodes.

VII. In an unstructured data architecture, private information of a person is not much secure over the social networking site which leads to misuse of the personal information.
VIII. Unstructured data is growing rapidly from calls, tweets and payment transactions need to control over the relevant to access the data on particular time from known location of authentic user.

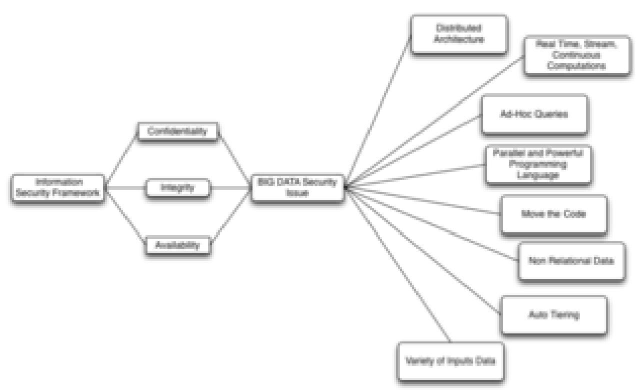

Figure2: Data Security Road map [3]

3 MOTIVATION

Now a day, Maximum human beings are using internet and sharing profiles, information related to debit/Credit card in distributed environment. In such sharing and distributed environment, the chances of active/passive attacks increase several times. Hence, it is the utmost responsibility to secure this shared data. However, this openness in the data environment creates a lot of challenges [14-15].

For providing security to complex data first we examined the issues and challenges of unstructured data then brief discussion about how to get rid of these issues and make security wall more strong than before, for this machine learning based detection methods provide insights for identifying novel attacks.

Now, it has been found that Machine-learning and Deep-Learning provides a promising results towards these challenges. Machine Learning is a branch of science in which machine can learn by itself [16]. Our first step is to detect new intrusions from the network by using K-Means clustering and after this hold requests of these intrusions in an original network by raising the alarm and divert these intrusions to honey-pot network, which is a false image of original network. At last examine the intrusions behavior by using naive Bayes and spread all over the network about intrusions.

4. PROPOSED WORK

In this research work, intrusion detection architecture and Model has been proposed and implemented. This ML Modeling has been validated on the basis of detection rate and the false alarm rate. A KDD-CUP99 datasets has been used with all features. Here, we used semi-supervised technique by applying K-Mean Clustering and Naive-Bayes classification algorithm in Honey-pot.

The Reason behind using naive Bayes in this research work is to identify the correct results of intrusion detection by using probability of occurring event. Clustering algorithm has some drawbacks which recover by using naive Bayes classification.

<u>Algorithm-</u>

The mathematical notation of Bayes theorem is given below [21]:

Likelihood
Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)\,P(c)}{P(x)}$$

Posterior Probability
Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

5 SOFTWARE ENVIRONMENT PYTHON ANACONDA

Python ver 3.6.5 has been used and it is a free and open-source Programming language and widely used in Machine Learning Projects.

Spyder (3.2.8) is the Scientific Python Development Environment has been used which is a powerful interactive development environment for the Python language with advanced editing, interactive testing, debugging and introspection features and a numerical computing environment thanks to the support of I Python (enhanced interactive Python interpreter) and popular Python libraries such as NumPy (linear algebra), SciPy (signal and image processing) or matplotlib (interactive 2D/3D plotting) [28].

6 HONEYPOT TECHNOLOGY

Honey-pot is a system which provide false image of original server. It is designed to monitor the malicious nodes. Honey-pot meaning is to detect which is offensive.

Honey-pot using a very unique characteristics and always want to force an attacker to hook attackers to attack into the-rise-of-data-anarchy, while monitoring the system activity and conduct of all, and the arrangement of these acts transcribed into the log. Due to this research, investigating the attackers mind, what type of tools he wants to use, plan of action and purpose, Honey-pot can be more impressive intrusion detection now a day, it is also help us to analyzing the behavior of real-time network for intrusion forensics.

<u>Implementation Steps</u>

Below show the steps of implementation in spyder with the help of python 3.6 [29]:

1) Installed all necessary modules such as numpy, matplotlib, pandas, Theano, Tensorflow and Keras in Ipython 3.6.5

2) Importing the dataset

3) Encoding categorical data

4) Splitting the dataset into the Training set and Test set

5) Feature Scaling

6) Initializing the ANN
   a) Adding the input layer and the first hidden layer
   b) Adding the second hidden layer
   c) Adding the output layer
7) Compiling the ANN
8) Fitting the ANN to the Training set
9) Making the predictions and evaluating the model by confusion Matrix

Pseudocode: Naïve Bayes classification [27]

1.) Given training dataset D which consists of network traffic belonging to different K clusters say cluster A, cluster B…..cluster K.
2.) Calculate the prior probability of cluster A= no. of objects of cluster A/total no. of objects.
    Prior probability of cluster B= no. of objects of cluster B/total no. of objects.
    .
    .
    .
    .
    Prior probability of cluster K= no. of objects of cluster K/total no. of objects.
3.) Find n_i the total no. of objects frequency of each cluster:
    n_a = the total no. of frequency of cluster A.
    .
    .
    .
    n_k = the total no. of frequency of cluster K.
4.) Find the conditional probability of every object occurrence in a given cluster.
    P (object 1/cluster A)= object count/n_i (A)
    P (object 1/cluster B) = object count/n_i (B)
    .
    .
    P (object 1/cluster K)= object count/n_i (K)
    .
    .
    .
    P (object n/cluster A) = object count/n_i (K)
5.) Avoid zero frequency problems by applying uniform distribution.
6.) Classify a new cluster X(anomaly) based on the probability of P(X/I_n).
    - Find P (A/I_n) = P(A)*P (object 1/cluster A) …. *P (object/cluster A).
    - Find P (B/I_n) = P (B)*P (object 1/cluster B) …. *P (object/cluster B).
      .
      .
      .
    - Find P (K/I_n) = P (K)*P (object 1/cluster K) …. *P (object/cluster K).
7.) Assign the cluster as anomaly that has higher probability.

7. MODELING AND SIMULATION

For research we used KDD CUP 99 dataset which contains 42 features. KDD CUP 1999 contains 41 attributes and last attribute is worked as a label [22-23]. In our implementation, we selected the features of the dataset. The attributes can be generalized as Normal, U2R, DoS, Probing and R2L. The short description of KDD CUP 99 used in this research shown in table 1. The performances of each method are measured according to the Accuracy, Detection Rate and False Positive Rate are following given below [25]:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN$$
$$\text{Detection Rate} = TP / (TP + FP)$$
$$\text{False Alarm} = FP / (FP + TN)$$

Where,
TP is True Positive,
TN is True Negative,
FP is False Positive,
FN is False Negative,

A Confusion Matrix is used to correspond the results, as shown in Tables 1. The Benefit of this matrix is that it tells us how many miss-classified get and as well as tells what misclassification has been extra originated. In this ANN model we get the confusion matrix are shown in Table 2, Table3, Table 4 and Table 5.

Attack Types	Training Examples	Testing Examples
Normal	97531	60692
Denial of Service	391569	237605
User to Remote	63	80
Root to User	1237	8707
Probing	4218	4277
Total Examples	495018	311361

Table 1: Shows the number of examples in 10% training and testing data of KDD99 dataset.

Actual	Predicted Normal	Predicted DoS	Predicted Probe	Predicted U2R	Predicted R2U	Accuracy %
Normal	8913	12	142	574	106	91.6
DoS	448	3696	20	1761	12	94.3
Probe	4	4	414	4	5	99.8
U2R	4	4	4	8	5	80.0
R2U	31	4	7	13	8	65.5

Table 2: Confusion Matrix for Naive Bayes Classifier Using Training Dataset.

Actual	Predicted Normal	Predicted DoS	Predicted Probe	Predicted U2R	Predicted R2U	Accuracy %
Normal	9691	7	27	9	13	99.6
DoS	7	33940	4	4	211	99.5
Probe	4	4	414	4	4	99.9
U2R	5	4	4	6	6	40.0
R2U	39	6	7	8	73	61.5

Table 3: Confusion Matrix for K-Means Clustering by Naive Bayes Classification Using Training Data Set

Actual	Predicted Normal	Predicted DoS	Predicted Probe	Predicted U2R	Predicted R2U	Accuracy %
Normal	7879	18	135	1668	47	81.0
DoS	6435	83233	421	4	4	82.5
Probe	10	16	397	4	4	95.6
U2R	5	4	4	8	4	80.0
R2U	14	4	5	4	106	90.3

Table 4: Confusion Matrix for Naive Bayes Classifier using Testing Dataset.

Actual	Predicted Normal	Predicted DoS	Predicted Probe	Predicted U2R	Predicted R2U	Accuracy %
Normal	9682	13	7	39	6	99.5
DoS	138	38988	31	4	5	99.6
Probe	4	7	408	4	4	98.3
U2R	5	4	4	8	4	80.0
R2U	8	16	4	7	98	98.3

Table 5: Confusion Matrix for K-Means Clustering via Naive Bayes Classifier using Testing Dataset.

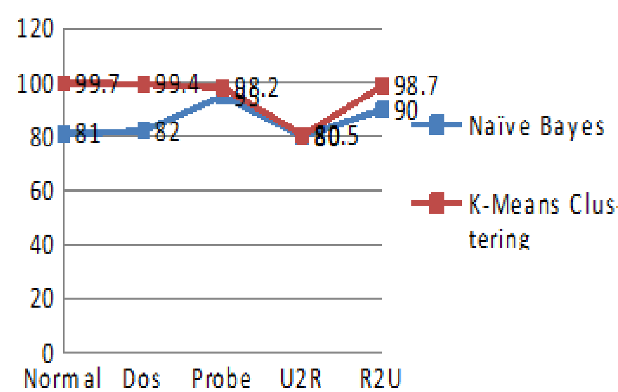

Fig 3: - Accuracy graph comparison by using testing dataset.

Above fig 3 shows the comparison of accuracy for our method and naive Bayes classification. In [17-

18] which uses Naive Bayesian Classification shows that the detection rate in detecting intrusion is 95% . However, in our case, the detection rate is 99%, with an error rate of 4%. However, in comparison to Naive Bayesian Classification, our approach generates more false positives.

Our last and final step is that revert all detected intrusions from network traffic towards honey pot server for security purpose of confidential data which may attacker could get harm within a original network. Honey-pot gives all false details of confidential data to attackers which they want to get harm of data. Hence in the end we can say that by using honey-pot technology confidentiality, integrity, relevance, availability, Qos increase and false alarm, traffic jam, denial of service, most important attacks are reduced.

8. CONCLUSION AND FUTURE SCOPE

Over all contribution towards this research work was to designed a Machine learning model by using ANN techniques. Each node is able to learn and update iteratively with its own experience. This type of Architecture is very effective for detecting the Intrusions mind-set and forward the alert among all peer nodes with the complete information about the attacker information. This approach prevents the whole networks from the complete jeopardize scenario and enhance the quality of efficiency of the network.

These models can also be enhanced by implementing SVM and Decision Tree algorithm.

REFERENCES

[1] International Data Corporation, Retrieved 23rd July 2016, http://www.idc.com
[2] Structured vs Unstructured Data, Retrieved 22nd July 2016 http://www.datasciencecentral.com/profiles/blogs/structured-vs-unstructured-data- the-rise-of-data-anarchy,
[3] M. Paryasto, A. Alamsyath, B. Raharjdo and Kuspriyanto, “Big Data Security Management Issues”, IEEE Information and Communication Technology (ICoICT), 2014 2nd International Conference on, Bandung, 2014, pp. 59-63
[4] S. Kaisler, F.Armour, J. A. Espinosa and W. Money, “Big Data: Issues and Challenges Moving Forward ”, System Sciences (HICSS), 2013 46th Hawaii International Conference on, Wailea, Maui, HI, 2013,pp. 995-1004
[5] Computer Security Division Information Technology Laboratory, Guide for conducting risk assessments. Technical report, National Institute of Standards and Technology, 2012.
[6] Yuri Demchenko, Paola Grosso, Cees de Laat, Peter Membrey, “Addressing Big Data Issues in Scientific Data Infrastructure”,IEEE 2014.
[7] Dongxio L. , Yongbo Z. , “An Intrusion Detection System Based on Honeypot Technology”, International Conference on Computer Science and Electronics Engineering, DOI 10.1109/ICCSEEE.2012.158, pp. 451-454, March 23-25, 2012
[8] Li L., “The Research and Design of Honeypot System Applied in the LAN Security”, IEEE International Conference on Software Engineering and Service Science (ICSESS), pp. 360-363, July 15-17, 2011
[9] Network Security document,22-09-2016, http://www.securitydocs.com/library/2692
[10] “Virtual Honeypots: From Botnet Tracking to Intrusion Detection” , Addison Wesley Professional, 2007, http://www.informit.com/imprint/index.aspx?st=61085
[11] Network Security from Malicious Attacks, 2016-09-22, http://www.securityfocus.com/archive/119/321957/30/0/threaded
[12] Yang, Zhen, et al. "An approach to spam detection by naive Bayes ensemble based on decision induction." Intelligent Systems Design and Applications, 2006. ISDA'06. Sixth International Conference on. Vol. 2. IEEE, 2006.
[13] Seref Sagiroglu, Duygu Sinanc “Big Data: A Review”, IEEE Collaboration Technologies and Systems (CTS), Pages 42-47, May 2013.
[14] Hue T.B.P, Thuc D.N, Thuy T.B.D, Echizen , Wohlgemuth S “Protecting Access Pattern Privacy in Database Outsourcing Service”, IEEE 27th International Conference on Advanced Information Networking and Applications Workshops, 2013.
[15] Bharti Thakur, Manish Mann, “Data Mining for Big Data: A Review”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014.
[16] Stephen Kaisler, Frank Armour, J. Alberto Espinosa, “Big Data: Issues and Challenges Moving Forward”, IEEE 46th Hawaii International Conference on System Sciences, 2012.

[17] Almeida, Fernando “The main challenges and issues of big data management" International Journal of Research Studies in Computing, Volume 2 Number 1, Pages 11-20, April 2013.

[18] B. Claise, G. Sadasivan, V. Valluri, and M. Djernaes, “Cisco Systems NetFlow Services Export Version 9,” RFC 3954 (Informational), Oct. 2004.

[19] Wafa' S.Al-Sharafat, and Reyadh Naoum “Development of Genetic-based Machine Learning for Network Intrusion Detection” World Academy of Science, Engineering and Technology 55, 2009.

[20] Ms.Nivedita Naidu, Dr.R.V.Dharaskar “An effective approach to network intrusion detection system using genetic algorithm”, International Journal of Computer Applications (0975 – 8887) Volume 1 – No. 2, 2010.

[21] Tom Mitchell. Machine Learning. Mc Graw Hill, 1997.

[22] F. V. Jensen. Introduction to Bayesien networks. UCL Press, University college, London, 1996.

[23] J. Pearl. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmman, San Francisco (California), 1988.

[24] Koniaris, I.; Papadimitriou, G.; Nicopolitidis, P.; Obaidat, M., "Honeypots deployment for the analysis and visualization of malware activity and malicious connections", IEEE International Conference on Communications (ICC), vol., no., pp.1819-1824, 10-14 June 2014.

[25] Song Li; Qian Zou; Wei Huang, "A new type of intrusion prevention system, "International Conference on Information Science, Electronics and Electrical Engineering (ISEEE), vol.1, no., pp.361-364, 26-28 April 2014.

[26] Chawda, K.; Patel, A.D., "Dynamic & hybrid honeypot model for scalable network monitoring," International Conference on Information Communication and Embedded Systems (ICICES), vol., no., pp.1-5, 27-28 Feb. 2014.

[27] Sathyadevan, Shiju, Devan M. S, and Surya Gangadharan S.. "Crime analysis and prediction using data mining", 2014 First International Conference on Networks & Soft Computing (ICNSC2014), 2014.

[28] Xiangfeng Suo; Xue Han; Yunhui Gao, "Research on the application of honeypot technology in intrusion detection system," IEEE Workshop on Advanced Research and Technology in Industry Applications (WARTIA), vol., no., pp.1030-1032, 29-30 Sept. 2014.

[29] https://pythonhosted.org/spyder/ dated 30 June 2018