International Journal on Wireless, Networking & Mobile Communication Innovations
http://eurekajournals.com/wireless.html
ISSN: 2581-5113

A Study on AI and Unsupervised Learning Approaches for Clustered Analysis of Attacker Activities in IoT Data

Vicky Singh[1], **Renuka Mahajan**[1]

[1]KPR Institute of Engineering and Technology

Abstract

The realm of Internet of Things (IoT) has expanded its influence across various applications, spanning from compact wearable devices to large-scale industrial systems, delivering substantial benefits to humanity. With an escalating number of devices equipped with sensing and processing units, the increased interconnectedness poses a heightened risk of data breaches. Security and privacy concerns loom large over diverse applications utilizing IoT technology. The susceptibility of IoT data to malicious activities by attackers, who may compromise the integrity of the data by hacking into IoT devices, is a significant issue.

This paper introduces distinct clustering techniques aimed at identifying potential attacker activities within IoT data. Leveraging AI and Machine Learning techniques, the proposed approach clusters both attacker-modified data and authentic data. The utilization of Bayesian algorithm, Chi-square algorithm, and convergence algorithm contributes to training and validating a model designed to recognize such attacks on IoT data.

Keywords: Artificial Intelligence, Machine Learning, IoT, Bayesian bandit algorithm, Convergence algorithm

Introduction

The evolution of technology has significantly eased communication between devices and individuals in recent years. Projections suggest that approximately 50-70 billion devices will be interconnected by the Internet by 2025. The Internet of Things (IoT) facilitates the connection of numerous sensors with smart devices, enabling seamless communication between various objects. These connected entities interact and exchange data through software [2], supporting a diverse range of devices, including smartphones, laptops, desktops, and tablets. Such devices are designed to communicate, interact, and be remotely monitored and controlled via the Internet. The IoT encompasses a broad spectrum of applications, such as smart homes, smart traffic, smart cities, weather forecasting, healthcare, and more, generating extensive volumes of data characterized by varying dimensions like volume, variety, and velocity.

Given the prevalence of attacks, particularly in data-intensive sectors like healthcare, where vast amounts of confidential information are stored, there is a heightened risk of unauthorized modification by attackers. Hence, it becomes imperative to distinguish between actual data and potentially compromised (unsecured) data [7]. To address this challenge, clustering techniques within machine learning algorithms are employed to categorize IoT data into secured or unsecured clusters.

Contribution

- A framework to model and cluster attacker activities is designed using machine learning techniques.
- We identify and differentiate the attacker activity by analysing the IoT data by indicating 1 as attackermodified data and 0 as attacker unmodified (original) data.
- Then, graphing is done to identify the activity pattern and the normal data.

Literature Survey

Various machine learning algorithms are applied to IoT data. The selection of algorithm depends upon the IoT application, characteristics of IoT data and the type of data processing task. K-means algorithm is used to identify the cluster centres and to assign data points and the Principal Component Analysis (PCA) is used to project data points onto L dimensional linear subspace[1].Clustering is an attempt to group the observations which have similar characteristic to find out valuable comprehensive information in each group. When the data are binary values, then hierarchical clustering is recommended [2]. In smart healthcare that deals with massive data that are processed in either edge or cloud computing data stream clustering techniques are used to cluster data abbreviations [3]. Generally many authors concentrate on the interoperability and integration of health data for analysing informative patterns and hardly concerned with the security. But recent researches also concentrate on security and privacy of patients' data. Activity monitoring, End point validation, two factor authentication, Granular access control , data masking and homomorphic encryptions are incorporated in healthcare frameworks to enable data security and privacy [4].

Identifying the type of attacker activity is considered the major task to justify if the framework is more secure. For linear models, Bayesian information capacity criterion was used and for binomial data, regression model is used [5]. For factor analysis, chi-square contingency measure gives the better result than Imax and varimax procedures with Kaiser normalization [6]. Modified BAT algorithm with modified dimension and additional inertia weight factor enhanced the convergence rate and increased accuracy when compared to the traditional BAT algorithm [7]. Internet of Things and cyber physical systems go hand in hand and clustering of attacker activities are under research. Peiyuan Sun et al. proposed Multivariate Hawkes Process to model attack pattern and graph based clustering approach to identify activity pattern [8][9]. ZuraKakushadze et al. proposed *K-means clustering algorithm to group cancer types from genome data. It can also be applied for different applications [10]. Olivier Thonnard and Marc Dacier performed clustering based on time signatures to identify intrusion detection from honeypot data [11]. IoT MicroMort was proposed by Petar Radanliev et al. to test and validate IoT risk with real data and found the current state of IoT cyber risk and the future forecasts of IoT cyber risk [12].

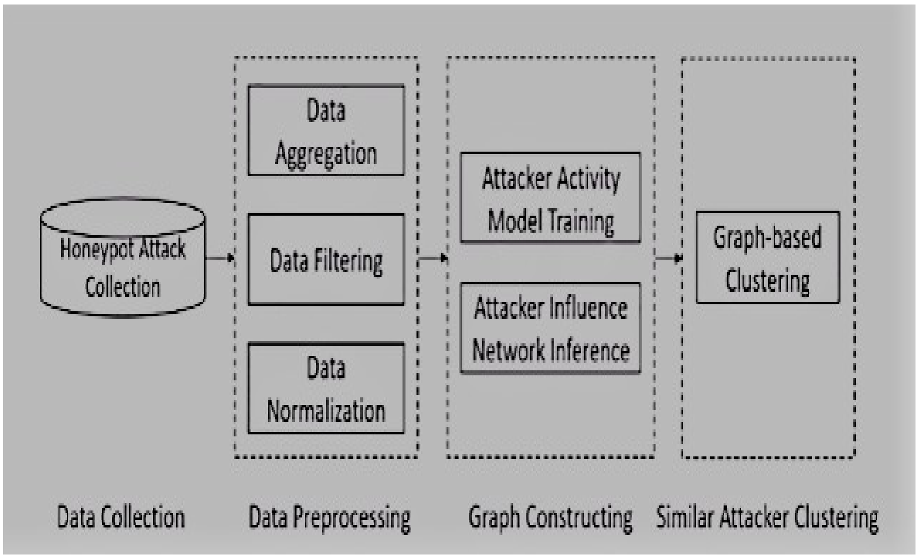
Honeypot Attack Collection
Data Aggregation
Data Filtering
Data Normalization
Attacker Activity Model Training
Attacker Influence Network Inference
Graph-based Clustering
Data Collection
Data Preprocessing
Graph Constructing
Similar Attacker Clustering

Figure 1: Work flow of the framework

Implementation Work

Data Collection and pre-processing

The dataset consists of patient information that includes patient ID, patient address, country, hacker ID and hacking action. The dataset may include out of range values, impossible information, missing data etc., where rigorous screening is required to pre-process these data before analysing. Analysing these data without pre-processing would result in incorrect results. The raw data has to be converted into clean and consistent data. However, this step takes considerable amount of time but produces analysis ready data. The independent and dependent attributes are identified. Every attribute except hacking action are identified as independent variables. The hacking action is a categorical data where 1 represent that attacker activity has occurred and 0 represents no attacker activity. Information preparation, pre-processing and filtering will consume reasonable amount of time interval. The product of information pre-processing is the final training set. The clustering algorithms are applied to the pre-processed dataset.

Bayesian-bandit algorithm

Bayesian algorithm is a statistical inference method to find the probability for a hypothesis based on the information available in prior. The probability of occurrences is computed from the likelihood function and prior probability.

$p(\theta|x) = p(x|\theta)\, p(\theta) / p(x)$ --------- (eq 1)

Where:

- p(x|θ) is that the “likelihood” i.e. the likelihood of observant the information x given our current beliefregarding the parameters theta
- p(θ) is that the “prior” – our current assumption regarding the parameters
- p(θ|x) is that the “posterior”–our updated assumption regarding the parameters once observant on theinformation x
- p(x) is a constant i.e., the likelihood of observant in any circumstances.

It will have values between zero and one that is sweet for things like click-through rate, or in alternativewords, for representing the likelihood of winning.

Once the prior may be a Beta distribution and also the likelihood is that a statistical distribution, the posterior is additionally a Beta distribution.

The update formula for deciding the posterior solely involves addition and subtraction, whereas for alternative distributions it are often a lot of difficult.

A visual illustration of theorem Bandits

Below we plot the distributions of every Beta after a specific range of trials.

Here, the green distribution, that has the best likelihood of winning, gets swindler and sharper because the range of trials will increase.

The secret is within the sampling stage. Once we sample from every distribution, the green distribution is far a lot probable to grant a number around 0.75, whereas the red and blue distributions can offer a lotof wider vary.

So once the green stealer has “proved itself” we tend to a way more probably to settle on it within the future, and that we simply don’t hassle to settle on the lower playing bandits, though there's a tiny low probability we tend to force[3].

A fat distribution suggests that a lot of “exploration”. A pointy distribution suggests that a lot of “exploitation” (if it's a relative high win rate). Note that because the technologist, you don’t opt for whether or not to explore or exploit. You sample from the distributions i.e., it israndomly set whether or not you ought to explore or exploit. Basically, the choice is random, with a bias toward bandits who have verified themselves to win a lot.

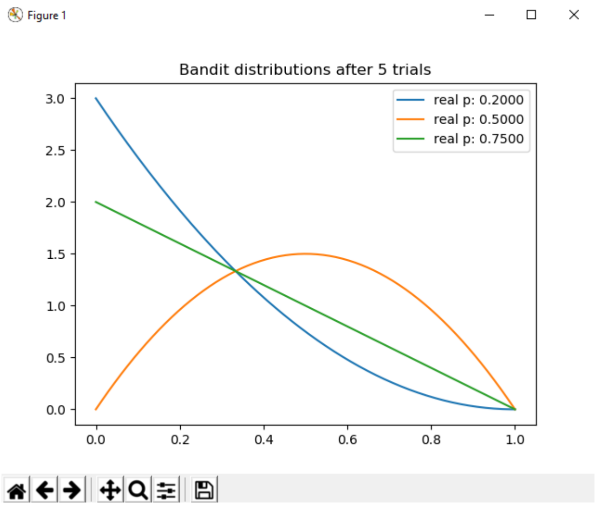

Figure Trial 1

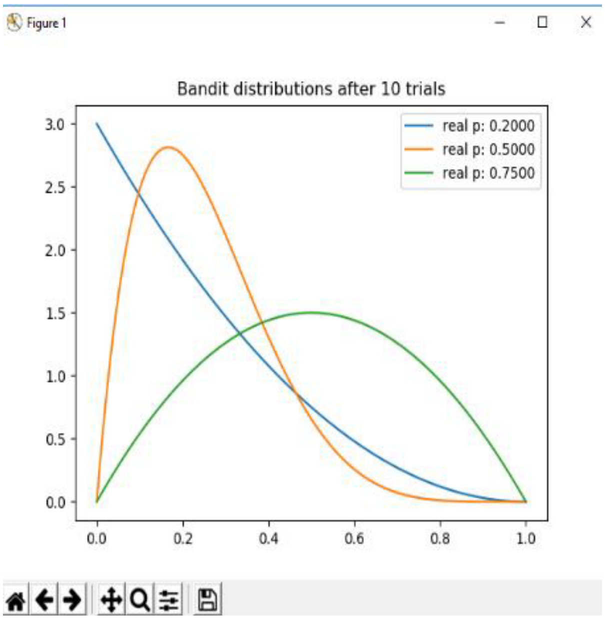

Figure Trial 2

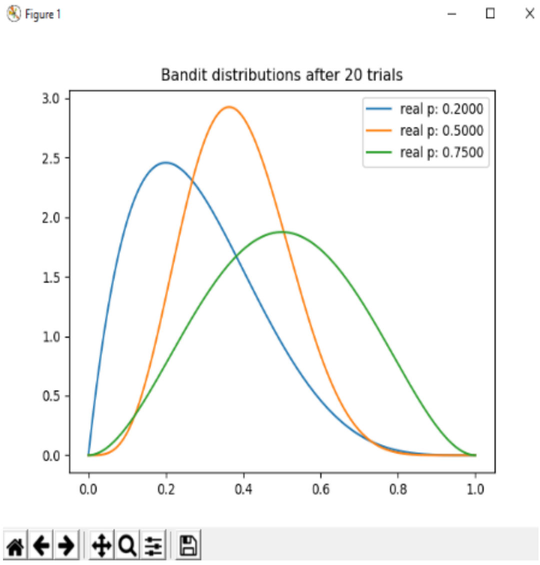

Figure Trial 3

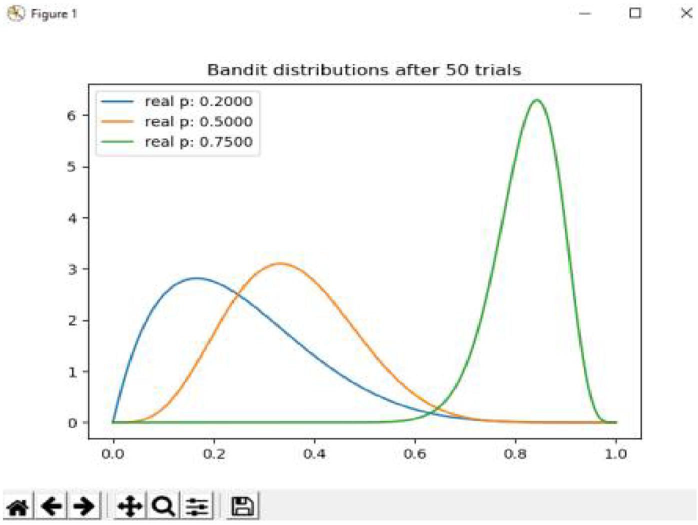

Figure Trial 4

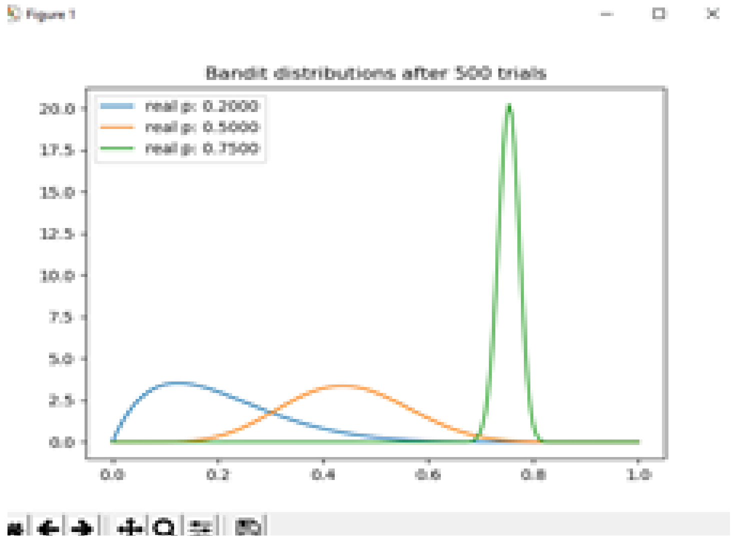

Figure Trial 5

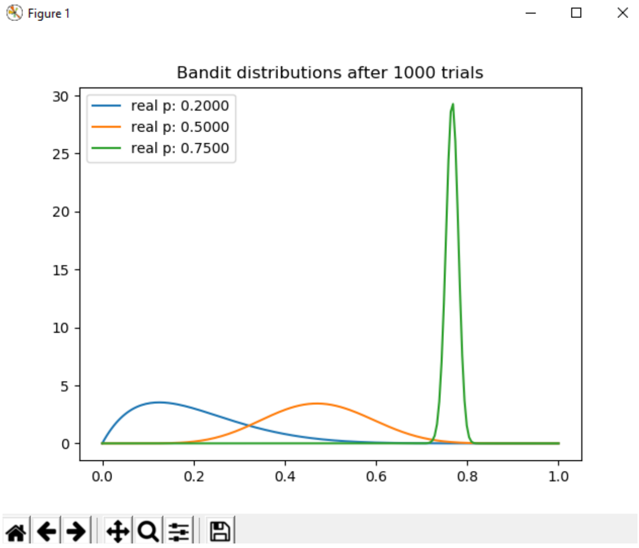

Figure Trial 6

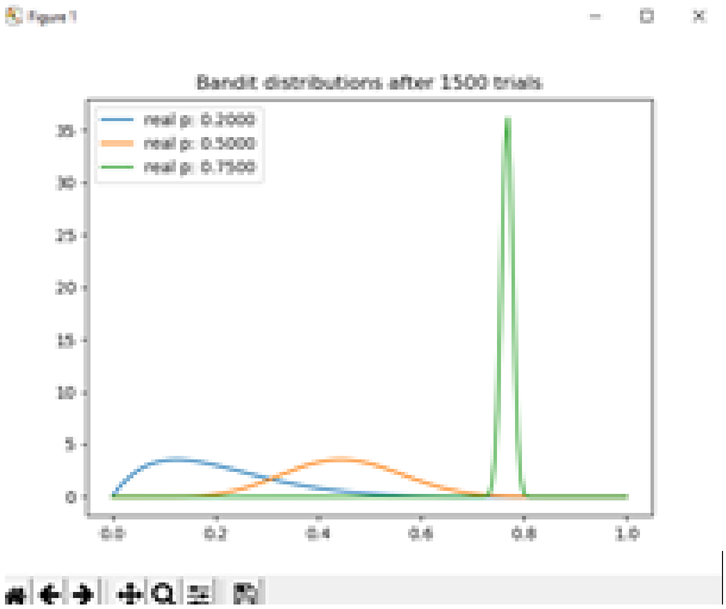

Figure Trial 7

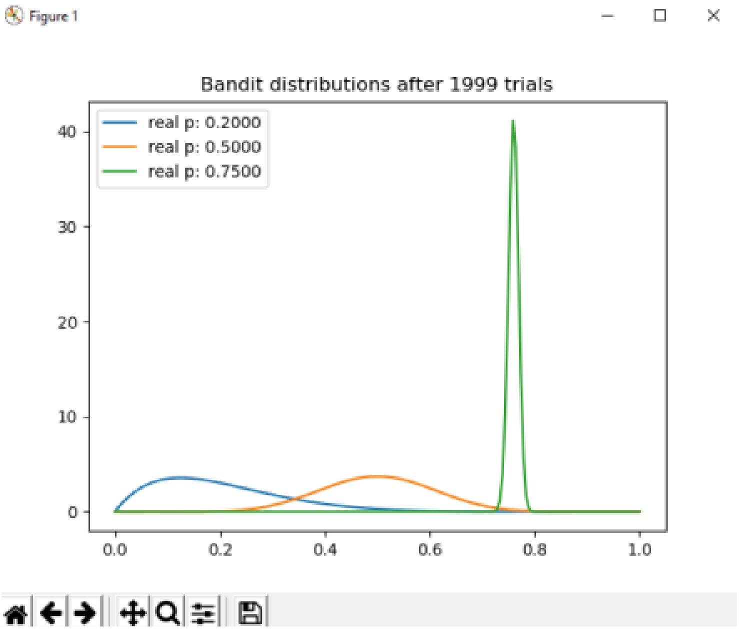

Figure Trial 8

Chi square algorithm:

A chi square (χ2) datum may be a check that measures expectations compared to actual discovered information. These information are utilized in scheming a chi square. Datum should be random, raw, reciprocally exclusive, drawn from freelance variables, and drawn from an outsized enough sample [6]. As an example, the results of moving a coin a hundred times meet these criteria.

The Formula for Chi square

χ 2c=Σ(Oi-Ei)2/Ei (eq 2)

Where:

- c are the degrees of self-determination.
- O is that the observed value(s).
- E is that the expected value(s).

Chi-squared tests are often constructed from a sum of squared errors, or through the sample variance. Test statistics that follow a chi-squared distribution arise from an assumption of independent normally distributed data, which is valid in many cases due to the central limit theorem. A chi-squared test can be used to attempt rejection of the null hypothesis that the data are independent.

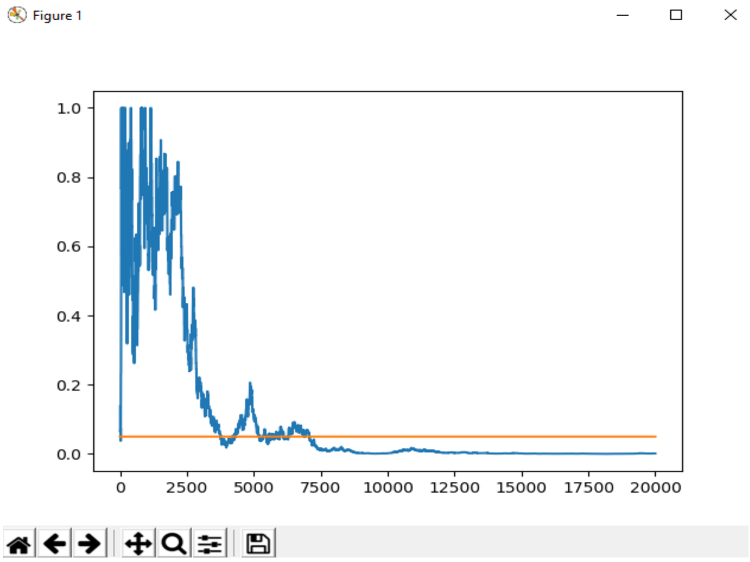

Figure Chi-square algorithm

This output shows the time at which the attacker can attack the information.

Convergence algorithm

A Convergence algorithm is usually said to converge when there is no significant improvement in the values of fitness of the population from one generation to the next. Examples of stopping criteria are generally, time limits placed on the Convergence algorithm run, generation limits, or if the algorithms find a suitably low fitness individual, lower than a specified fitness threshold (in case we are minimizing fitness).

Convergence generally means that a sequence of a certain sample quantity approaches a constant as the sample size tends to infinity. Convergence is also a property of an iterative algorithm to stabilize on some aim value.

Convergence refers to the investigation of the behavior of certain sample quantities when the sample size approaches infinity. Two important types of convergence are convergence in probability and almost sure convergence.

Convergence in probability:

A sequence of random variables X1,...,Xn converges in probability to a random variable X if

$$\lim_{x\to\infty} P(|Xn-X|\leq \varepsilon) = 0 \text{ -------(eq 3)}$$

for every ε>0. This means that at the limit as n increases to infinity almost all of the probability mass becomes concentrated around X in a small interval. This type of convergence is used in the weak law oflarge numbers.

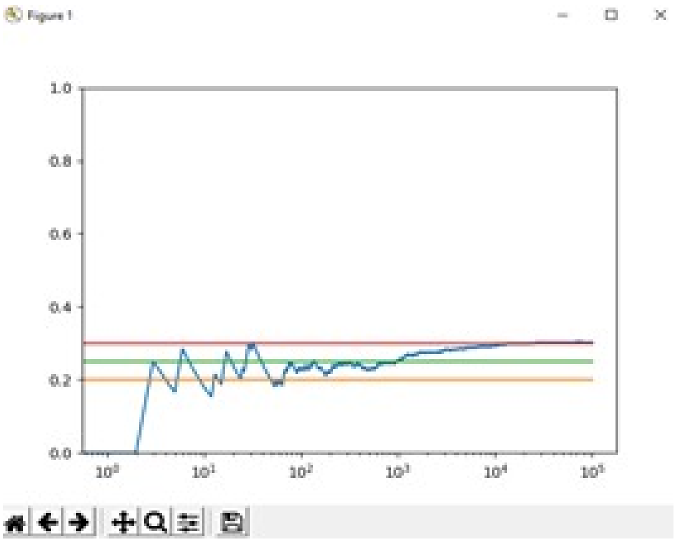

Figure Convergence algorithm showing the convergent point of attack

Almost sure convergence:

Similar to the previous statement, a sequence of random variables X1,. ....,Xn converges almost surely to a random variable X if

$$P(\lim_{x\to\infty}|Xn-X|<\varepsilon) = 1 \text{ ----(eq 4)}$$

for every ε>0. Here, compared to the previous case, the limit is achieved with probability one. Almost sure convergence is used in the strong law of large numbers and it implies convergence in probability (note that convergence in probability does not imply almost sure convergence).

The common Convergence algorithm terminating conditions are:

- When fixed number of generations is reached
- An optimal solution is obtained that satisfies the optimization criteria
- When successive Convergence algorithm iterations no longer produce better results
- Allocated budget (computational time / cost) reached

Convergence algorithmic program is alleged to converge once, because the iterations proceed, the output gets nearer and closer to a particular point.

In some circumstances, an algorithmic program won't converge; it might even diverge, wherever its output can bear larger and bigger oscillations, near approaching a helpful result.

Conclusion

IoT comprises of a huge number of devices that are linked with each other and transfer huge volumes of data and it finds its application in healthcare industry as well. The services provided by the system are enhanced and monitored by analyzing the data and preserving its integrity. Provided with many algorithms to analyze these data, it is equally important to find if the data has be attacked or not.

Thus, in this paper, clustering of IoT data is done with different machine learning technique which provides information about actual data and attacker modified data. Chi square algorithm and convergence algorithm says the time at which the attack has happened and the Bayesian Banditalgorithm gave a clear picture of attacker activity over number of trials.

References

Mohammad Saeid Mahdavinejad, Mohammadreza Rezvan, Mohammadamin Barekatain, Peyman Adibi, Payam Barnaghi, Amit P. Sheth, “Machine learning for internet of things data analysis: a survey” , Digital Communications and Networks, Volume 4, Issue 3,2018,Pages 161-175,ISSN 2352- 8648,https://doi.org/10.1016/j.dcan.2017.10.002.

KrittipatPitchayadejanant, ParinyaNakpathom, “Data mining approach for arranging and clustering the agro-tourism activities in orchard,” Kasetsart Journal of Social Sciences 39 (2018) 407 – 413

Yogita, Yogita and Durga Toshniwal. “Clustering techniques for streaming data-a survey.” 2013 3rd IEEE International Advance Computing Conference (IACC) (2013): 951-956.

Prableen Kaur, Manik Sharma, Mamta Mittal, “Big Data and Machine Learning Based Secure Healthcare Framework”, Procedia Computer Science, Volume 132, 2018, Pages 1049-1059, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.05.020.

David C. Woods, James M. McGree, Susan M. Lewis, “Bayesian information capacity designs for generalised linear models,” Computational Statistics and Data Analysis 113 (2017) 226–238

Leo Knusel, “Chisquare as a rotation criterion in factor analysis,” Computational Statistics & Data Analysis,Volume 52, Issue 9, 15 May 2008, Pages 4243-4252

M.R. Ramli, Z. Abal Abas, M.I. Desa, Z. Zainal Abidin, “Enhanced convergence of Bat Algorithm,” Computer and Information Sciences 24(2018) 316- 328

Hugh Boyes, Bil Hallaq, Joe Cunningham, Tim Watson, “The industrial internet of things (IIoT): An analysis framework,” Computers in Industry 101(2018) 1-12

Peiyuan Sun, Jianxin Li, Md Zakirul Alam Bhuiyan, Lihong Wang, Bo Li, “Modelling and clustering attacker activities in IoT through machine learning techniques," Information Sciences 479 (2019) 456–471.

Zura Kakushadze, Willie Yu, "K-means and cluster models for cancer signatures," Biomolecular Detection and Quantification 13 (2017) 7–31

Olivier Thonnarda, Marc Dacier, "A framework for attack patterns' discovery in honeynet data," digital investigation 5 (2008) S128–S139

Petar Radanliev, David Charles De Roure, "Future developments in cyber risk assessment for the internet of things" Computers in Industry 102 (2018) 14-22.

Sanaz Rahimi Moosavi, Ethiopia Nigussie, "Performance Analysis of End-to-End Security Schemes in Healthcare IoT" Procedia Computer Science 130 (2018) 432-439.