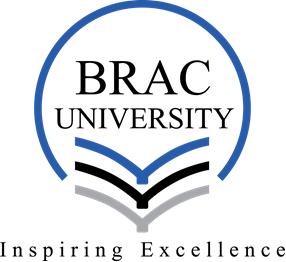
**Pre-thesis -I Report**



A Model for Anomalies Detection in Internet of Things (IoT) Using Inverse Weight Clustering and

Decision Tree

Name and ID:

**Supervisor:**

**Co-supervisor:**

**Date of Submission:**

**Department of Computer Science and Engineering**

**BRAC University**

A Model for Anomalies Detection in Internet of Things (IoT) Using Inverse Weight Clustering and Decision Tree

**Abstract**

Internet of Things (IoT) is one of the fast growing technologies today. It is a technology by which billions of smart objects or devices known as “Things” can use several types of sensors to collect various types of data about themselves and/or the surrounding environment. They can then share this with authorized parties to serve several purposes such as controlling and monitoring industrial facilities or improving business service or functions. There are currently 3 billion devices connected to the Internet. The number will increase to 20 billion by 2020. While these devices make our life easier, safer and healthier, they are expanding the number of attack targets available to hackers. Therefore, protecting these devices from adversaries and unauthorized access and modification is very important. This research proposes a novel approach for anomalies detection in IoT systems based on a combination of two robust machine learning algorithms; inverse weight clustering (IWC) and C4.5 decision tree algorithm. IWC is an enhanced version of k-means algorithm that can be used to effectively cluster data into groups based on similarities between this data. C4.5 is a decision tree algorithm that can be used to build decision tree for classifying data. The proposed model was tested and evaluated, and the result demonstrates that the model is very accurate in detecting anomalies in IoT data.

1. **Introduction**

Since the 1960s, the Internet, as a massive network of Local Area Networks (LAN) has been playing a vital role in connecting people, businesses and organizations together. The Internet has broken down geographical barriers between people and has given them a robust, efficient and cost effective means of communications.

Now, it seems that things are about to change in the world of the Internet as a result of the appearance of smart objects that have the capability of generating and communicating data over the Internet in a similar way to humans. Internet of Things (IoT) is the latest technology and systems that holds the potential to change our way of life. One can think of IoT as a technology made of two components; “Things” and “Internet”. The “Things” refers to any object or device that has the ability to perceive or collect data about itself or the surrounding environment. Depending on what type, smartness and capabilities of this object, the object may be able to analyze and act smartly with other objects using “Internet” as a network for communications.

The communication over the Internet in IoT is not limited to communication between IoT objects only. It goes beyond this and approaches humans in a way that can make life easier, healthier, and much better. For example, there are so much research around how IoT can help improve humans’ health through monitoring their health remotely thus eliminating the need to visit the hospital quite often.

One of these projects on IoT for health is being conducted by the University of Edinburgh in Scotland. The university has created the Centre for Speckled Computing to investigate how IoT can be utilized to improve the health of people in Scotland. They have created tiny computing devices about the size of a thumb. These devices can be attached to people on their chests, monitor and collect respiratory data and then transmit it wirelessly to doctors who are following their cases remotely [1].

IoT is pervasive and used in almost all aspects of our life. IoT is used by governments around the world to collect data from different sectors and to provide better services in health, transposition, security and development. IoT is used by businesses in order to provide better services to the customers’ or to enhance safety and security to workers in the workplace.

IoT is also used by people to better enjoy and manage their life as in the case with Amazon Echo; a smart IoT devices that has the capability of interacting with people using voice. Echo can be asked to provide advice with regards to the weather, schedule alarms, play music, or obtain new feeds from various resources from the internet.

1. **Research Problem**

As IoT becomes more and more pervasive every day, attacks against it increase. According to [2], the number of devices connected to the internet by the end of 2016 will be 3 billion. It will increase by 2020 to hit 20 billion. Having such a massive increase in the number of devices by 2020, underlines the question of what sort of sensory data these devices will be able to communicate. Such data can be medical, financial, social, environmental, or any other type of data that has various structures and importance. With this in mind, it is very important to think of how to secure such data and devices. According to [3], information security is a component of computer science that is concerned with protecting information systems from adversaries who aim to take down the availability, integrity and confidentiality of the data and services provided by an information system. Security of an information system can be maintained using physical and logical controls. The physical controls are used to prevent adversaries from physically reaching the hardware devices that are used to run a given information system. Physical controls can be guards, CCTV, electronic budgets, finger print scanners and iris scanners.

Logical controls are software safeguards that can be used to ensure that data and information systems are available only to authorized entities – individuals and process, according to business policies. Logical controls including passwords, access control and intrusion detection and prevention systems.

Intrusion detection and prevention systems are systems that can be used to detect known attacks or anomalies in information systems. The key difference between an intrusion detection system (IDS) and intrusion prevention system (IPS) is that IPS is a proactive system. Therefore, upon detecting or matching attack signature or anomalies, the IPS takes predefined and systematic steps to stop the attack or anomalies. The IDS, on the other hand, is limited to detecting and sending alerts in case of detecting attacks or anomalies [4].

IDS and IPS can be network-based or host-based systems. They can be used to protect a network or end-user device. The network could be an IoT network and end-user device can be an IoT device such as sensing devices. Furthermore, IDS and IPS can be hardware and software in one bundle or software [4].

Additionally, there are two methods of detections; signature based detection and anomaly based detection. Signature based detection methods are effective in detecting known attacks by inspecting network traffic or data in systems memory for specific patterns. Anomaly based detection is used in detecting unknown attacks by monitoring the behavior of the whole system, objects or traffic and compare it against predefined assumed to be normal behavior. Any division from normal behavior is treated as a potential attack [5].

According to [5], IDS and IPS used both types of detections to stop known and unknown attacks. However, anomaly based detection IDS and IPS have become more popular due to the constant increase in the complexity of attacks against assets. Attackers have become capable of producing malwares that have the ability to change their structure (i.e. polymorphism) on the fly to evade signature based IPS and IDS. Furthermore, once an attack is detected, it takes time to create a signature and apply it to IDS or IPS.

Anomaly based IDS and IPS relies on artificial intelligent (AI) and machine learning (ML) to detect anomalies [6]. The idea behind AI and ML is to make a machine capable of learning by itself and distinguish between normal and abnormal behavior on the system. The process of teaching a machine takes different forms; supervised, unsupervised and reinforcement learning.

Regardless of the means of teaching, a machine needs to be trained in order to be able to predict. Several ML algorithms have been used in intrusion detections. K-means is one of the most popular algorithms for clustering data into groups. Decision tree algorithms such as C4.5 and C5.0 are among the most popular classification algorithms used to distinguish between normal and abnormal behavior or patterns in data. Most of the intrusion detection systems use such a combination of algorithms to cluster sample data into groups, label them, and then use a classifier to train the intrusion detection systems to distinguish between these groups [7].

According to [7], anomaly based intrusion detection systems suffer from false positive, which is the case when IDS or IPS identify normal activity as abnormal. Additionally, anomalies based on intrusion detection systems are said to be computing intrusive system. It requires a lot of processing power and memory to work fast specially if the IDS or IPS is a real time intrusion detection system.

The [8] examine the security in IoT. Using anomaly based detection in IoT is challenging and harder than using it with non-IoT for several reasons. Given the large number of IoT devices, attackers have more targets to attack than any time before. Furthermore, these targets are relatively easier to attack than traditional computers since their hardware capabilities in terms of processing and memory is very limited in a way that can render it difficult to use host-based intrusion detection systems. Additionally, IoT devices produce data of different structures and formats and communicate it over various types of networks including, the Internet, wireless sensory network (WSN), radio frequency identification (RFID), Bluetooth and many more.

With this in mind, the question posed could ask whether anomaly based detection techniques be used in detecting anomalies in IoT. The answer to this is yes and a lot of research has been conducted in this area. In [9], [10], used several ML learning techniques to detect anomalies in IoT networks. Most of these techniques are designed to detect anomalies of network attacks such as fake IoT nodes, malicious routing in IoT. Minimum attention has been paid to detecting anomalies in data generated by IoT devices themselves.

Therefore, the question that this research is trying to answer is:

***How effective is a combination of inverse weight clustering (IWC) and***

***decision tree (DT) classification algorithms in detecting anomalies in IoT?***

As mentioned earlier, DT algorithms are generally used for classification in IDS and IPS. It is, however, used with K-means algorithm to produce IDS and IPS models. According to [11], K-means has limitations; one of which is the initial selection of data before starting clustering. The [11] proposed inverse weight clustering (IWC) which is an enhanced version of K-means.

This research will answer the above question through exploring a IWC and DT combination instead of K-means and DT combination.

1. **Research Objectives**

This research aims to develop an intrusion detection system for detecting application layer attacks against IoT device using a combination of inverse weight clustering (IWC) and decision tree (DT) algorithms. Usually an IoT device collects sensing data and sends it for further processing via gateways to a central system. Anomalies in data sent by IoT devices might be detected at an application layer in central systems. The objectives of this research are:

1. To deeply understand IoT, and how it works.

2. To deeply understand anomalies detection techniques; IWC and DT.

3. To develop a model for intrusion detection based on IWC and DT and apply it to IoT.

4. To evaluate the model.

5. To offer recommendations on improving the model.

1. **Literature Review**

The Internet as we know it is changing rapidly as a result of the introduction of smart devices that have the ability to communicate with humans and with each other’s use over the Internet. In a recent published research by Gartner – the technology research giant, the number of “Things” connected to the Internet by 2016 is expected to be 6.4 Billion [2]. According to the same source too, this number is expected to reach 20.8 billion by 2020.

One question immediately posed would be where such a number could come from. The answer is simple; an end user may have more than one device connected to the Internet. The list of devices includes but not limited to iPhone, iPad, iWatch, Smart TVs and many more. Each device has the ability to generate data and share it over the Internet, and this is what constitutes the Internet of Things (IoT).

* 1. **Internet of Things (IoT)**

IoT is an umbrella term that covers technologies, design principles, and systems associated with the ever-growing phenomenon of Internet-connected devices – “Things”. According to [3], IoT as a phrase is not new. It appeared for the first time in 1999 at Massachusetts Institute of Technology (MIT) Auto-ID Centre, and was used to refer to building an Internet based network that cover all things in the world to realize automatic identification of things through information sharing. The “Things” use related technologies such as Radio Frequency Identification (RFID) to enable communication and realization.

**2.1.1 IoT Architecture**

According to [4], ITU proposed a layered architecture of IoT. There are four layers in this architecture; device layer, network layer, service and application support layer, and application layer – the layers are ordered in a bottom-up approach. In addition to these layers, ITU proposed two additional layers; management capabilities and security capabilities.

Figure … shows all the layers in the proposed architecture for IoT by ITU. The main four layers stacked horizontally are functional layers. In other words, these layers are used to collect, transmit, and process data in IoT networks. The other two layers depicted vertically take a part of functionality of each layer for the purpose of management and security.

* 1. **Related Works**

This part aims to critically review previous relevant work in the field of Intrusion Detection

Systems in the context of Internet and specifically in the context of Internet of Things. We

analyses the different techniques used for the main results achieved and we show how intrusion detection for Internet of Things has its own specific challenges due to the heterogeneity of the devices, the limited computational capabilities and the massive number of connected devices that makes the embedding of intrusion detection systems harder.

Intrusion Detection Systems (IDS) are used to protect data that is being exchanged between end points and processes within an information system, from unauthorized access and modification. According to [14] , an IDS is designed to distinguish between normal behaviours from abnormal behaviours based on effective classification model built inside that IDS.

The problem of building effective classification models for an IDS is generally approached using several data clustering and classification algorithms. A recent work [14] represents a classical hybrid approach closer to what we implement in this thesis work. Goldman suggests a hybrid intrusion detection system based on Support Vector Machine (SVM) and C5.0. The author claims that using such a combination of algorithms would improve the accuracy of intrusion detection when compared to using SVM and C5.0 separately.

The research work [15] proposes a novel intrusion detection model. The model uses MapReduce to process large structured and unstructured data set into key/value pairs. The way MapReduce produces key/value pairs for intrusion detection relies on using a combination of Fuzzy CMeans

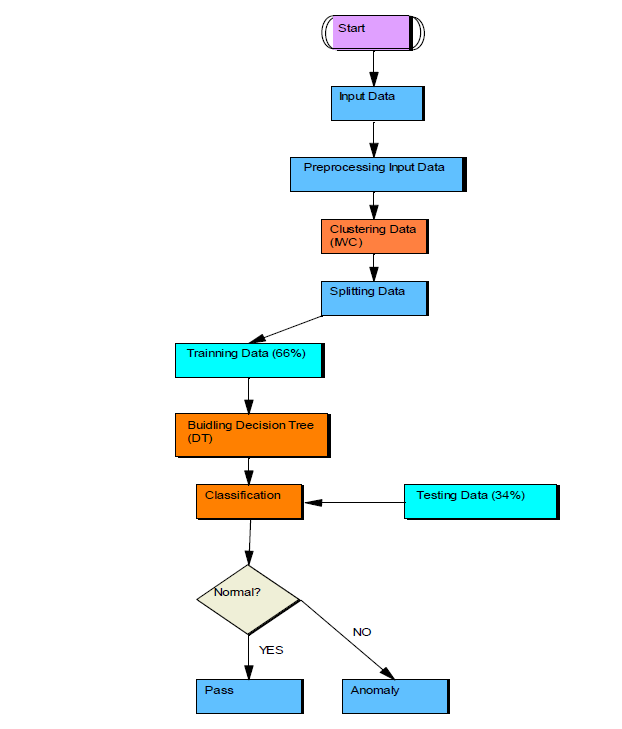
(FCM) clustering and SVM for classification. FCM uses features to cluster data into groups. The advantage of FCM over non-fuzzy CM is that FCM allow data points to belong to multiple clusters while non-fuzzy CM does not. In non-fuzzy CM, the data point can only belong to one cluster. The [15] has not published the results of testing the proposed model, therefore, no results can be given with regard to the accuracy of such model.

The research work [16] propose a hybrid intrusion detection system using two data mining techniques; K-means and Naïve Bayes for clustering and classifying data. The attacks that the [16] model is designed to detect are; DoS, Probe, U2R and R2L. These classes refer to the types of attacks that exist in KDD cup 99 data set used by [16] to evaluate their work.

From the above discussion, it is observed that most of the researches design anomaly detection schemes for IoT using similar approach as in designing anomaly detection schemes for other devices. However, one has to consider the unique challenges that IoT has. For instance, intrusive detection schemes that require significantly high computing power or memory consumption do not suit IoT. Furthermore, the IoT devices are heterogeneous and generate heterogeneous data that mount to big data. Thus the approaches to detect anomalies in IoT should not be limited to detecting network attacks, instead, they should go deeper and detect anomalies in data or the payloads carried in network traffic.

1. **Work Plan**

The purpose of the proposed anomalies detection model in IoT is to detect anomalies in data that is being exchanged by IoT devices at the application layer of IoT architected as described in the previous chapter. In order to do so, the model requires designing a process that takes data from IoT devices as an input, systematically process input data, and produce predictions of two folds; “normal” or “abnormal”. The Figure 1 provides a high level view of the model design.



**Figure 1**: *The flow chart of the proposed intrusion detection model*

Anomalies Detection process (ADP) is the process that is responsible for clustering, classification, and producing predictions. It consists of three major stages:

1. Input data preprocessing: this stage is concerned with formatting the input data in a way that makes it easy for ADP to process it.

2. Processing: this stage is concerned with clustering input data into groups using IWC and building DT for classification.

3. Predictions: this stage is concerned with using the already built DT for predictions.

The input data goes into the preprocessing stage and then used by IWC to build two clusters; one cluster combines normal data and the other one combines abnormal data. Clustering takes place based on similarities between data attributes.

After clustering, the preprocessed input data is split into two groups; one group is used for training and building the decision tree, and the other group is used for testing the accuracy of the decision tree in distinguishing between normal and abnormal data.

1. **Conclusion**

Security solutions such as intrusion detection are used to minimize the risk to having any component or data in an information system being compromised. This means that despite so much research and security solutions that are being utilized every day to protect IoT systems, the chance of having an IoT object being compromised and sending abnormal data to be processed at application level still stands. There has been a shortage of research that tries to find anomalies in IoT data at application level. Thus this research represents an attempt to fill this gap by adding another layer of protection to the already existing layers or the layers that are being developed for IoT systems.

**References**

[1] Nerney, C. (2012) The tiny (yet powerful) world of speckled computing. Retrieved

from: http://www.itworld.com/article/2721483/consumer-tech-science/the-tiny--yetpowerful--

world-of-speckled-computing.html.

[2] Van Der Meulen, R. (2015). Gartner Says 6.4 Billion Connected "Things" Will Be in

Use in 2016, Up 30 Percent from 2015. Retrieved from:

http://www.gartner.com/newsroom/id/3165317.

[3] Alexander, D. & Finch, A. (2014). Information Security Management Principles (2nd

Eds.). BCS, The Chartered Institute for IT.

[4] Gordon, A. (2015). Official (ISC)2 Guide to the CISSP CBK (4th Eds.). Auerbach

Publications.

[5] Prabha, K. & Sree, S. S. (2016). A Survey on IPS Methods and Techniques. In

International Journal of Computer Science Issues, Vol. 13, Issue 2, pp. 38-43. March

2016. DOI:10.20943/01201602.3843.

[6] Buczak, A. L. & Guven, E. (2015). A Survey of Data Mining and Machine Learning

Methods for Cyber Security Intrusion Detection. In IEEE Communications Surveys &

Tutorials, vol. 18, issue 2, pp. 1153-1176. Second quarter 2016. DOI:

10.1109/COMST.2015.2494502.

[7] Haq, N., Onik, A., Hridoy, A., Rafni, M., Shah, F., & Farid, D. (2015). Application of

Machine Learning Approaches in Intrusion Detection System: A Survey. In

International Journal of Advanced Research in Artificial Intelligence, vol. 4, issue 3, pp.

9-18.

[8] Zhao, K. & Ge, L. (2016). A Survey on the Internet of Things Security. In 9th

International Conference on Computational Intelligence and Security (CIS). DOI:

10.1109/CIS.2013.145.

[9] Pajouh, H. H., Javidan, R., Khayami, R. and Ali, D. (2016) A Two-layer Dimension

Reduction and Two-tier Classification Model for Anomaly-Based Intrusion Detection

in IoT Backbone Networks. In IEEE Transactions on Emerging Topics in Computing,

vol. PP, Issue 99, pp. 1-11. DOI: 10.1109/TETC.2016.2633228.

[10] Hodo, E. et al. (2016) Threat analysis of IoT networks using artificial neural network

intrusion detection system. In International Symposium on Networks, Computers and

Communications (ISNCC), pp. 1-6.

[11] Ashour, W. & Fyfe, C. (2007) Inverse Weighted Clustering Algorithm. The University

of Paisley.

[12] Ning, H. (2013). Unit and Ubiquitous Internet of Things. CRC Press Inc.

[13] Vermesan, O., & Friess, P. (Eds.). (2014). Internet of Things Applications - From

Research and Innovation to Market Deployment. River Publishers.

[14] Golman, V. (2014). An Efficient Hybrid Intrusion Detection System based on C5.0 and

SVM. In International Journal of Database Theory and Application, vol. 7, No. 2 (2014),

pp. 59-70.

[15] Hajare, S. A. (2016) Detection of Network Attacks Using Big Data Analysis. In

International Journal on Recent and Innovation Trends in Computing and

Communication, vol. 4, issue 5, pp. 86-88.

[16] Tanpure S. S. et al. (2016) Intrusion Detection System in Data Mining using Hybrid

Approach. In International Journal of Computer Applications, pp. 0975-8887.