ANOMALY DETECTION IN CLOUD COMPUTING

Tapas Kumar Mishra

ANOMALY DETECTION IN CLOUD COMPUTING

Thesis submitted to the
Indian Institute of Information Technology Guwahati
for award of the degree

of

Master of Technology

by

Tapas Kumar Mishra

under the supervision of

Dr Angshuman Jana

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY GUWAHATI

April 2020

CERTIFICATE

This is to certify that the thesis entitled **“Anomaly detection in cloud computing”**, *submitted by* **name** *to the institute name, for the award of the degree of Master of Technology, is a record of bona fide research work carried out by him under my supervision and guidance. The thesis, in my opinion, is worthy of consideration for the award of the degree of Master of Technology in accordance with the regulations of the Institute. To the best of my/our knowledge, the results embodied in the thesis have not been submitted to any other university or institute for the award of any other degree or diploma.*

Dr guide name,
Assistant Professor,
Department of Computer Science and Engineering
institute name

Date:

DECLARATION

I certify that

a. The work contained in this thesis is original and has been done by me under the general supervision of my supervisor(s).

b. The work has not been submitted to any other Institute for any degree or diploma.

c. I have followed the guidelines provided by the Institute in writing the thesis.

d. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

e. Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.

f. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

(Candidate name)

ACKNOWLEDGMENTS

Anomaly detection has crucial significance in the wide variety of domains as it provides critical and actionable information. For example, an anomaly in MRI image scan could be an indication of the malignant tumor or anomalous reading from production plant sensor may indicate faulty component.This thesis studies in detail various kind of anomaly detection techniques and compares them based on their effectiveness. Algorithms have been implemented in Python mostly.

I would like to thank my supervisor ¡¿ for helping me to carry out this thesis under his guidance. I am also thankful to our Head of Department Dr Rakesh Matam Sir for his valuable guidance from time to time.I would also like to thank our thesis panel based on their reivews and suggestions I improved the the work carried out.His guidance helped me in researching and writing the thesis.

ABSTRACT

In today's world of changing technologies due to advancement of technology,more and more organizations are shifting to usage off cloud computing services worldwide. With cloud computing service provider companies like Google, Amazon, AliBaba, Microsoft ,IBM investing more and more in cloud computing.It is becoming target of internet threats, such as malware or virus, technical vulnerability and negligent behaviours. Hence anomaly detection in cloud has been an emerging area of research for researchers like us. This thesis addresses the main security and privacy issues in Cloud Computing. It is essential to have an Anomaly Detection System (ADS) to detect anomalies with a high detection accuracy in cloud environment. With the evolving technological setup this work becomes complex day by day. This work proposes an anomaly detection system at the administrator level to improve the accuracy of the detection system. The proposed system is implemented and it uses classifiers and data mining techniques to detect anomalies in network. The DARPA's KDD cup data-set 1999 is used for experiments. Using simple data analytics techniques the system is able to find out anomalies in cloud. Machine Learning has various applications: classification, predicting next value, anomaly detection, and discovering structure. In this thesis we have studied,how anomaly detection detects anomalies from security perspective on cloud service providers. Anomaly detection has a wide range of applications such as fraud detection, surveillance, diagnosis, data cleanup, and predictive maintenance.

List of Abbreviations

IIITG	Indian Institute of Information Technology
CSE	Computer Science and Engineering
PhD	Doctor of Philosophy
VM	Virtual Machine
GCP	Google Cloud Platform
AWS	Amazon Web Services

List of Symbols

Symbol	**Description**
β	intensity factor
α	threshold value

Contents

List of Figures

Chapter 1

Introduction

Anomaly detection refers to identification of rare items,events,observations which are suspicious in their occurrences. Any deviation from normal beavior in a working system is an anomaly. Different kind of anomalies differ differently. Typically when a system is deviating from its normal behavior then it could be under some kind of attack which might lead to more problems from a security point of view. These challenges have been keeping the researchers hooked on to find anomalous behavior in systems.Once we have established the anomalous behaviour of system then we can apply several algorithms and approaches to detect and study it.That is the purpose of this work.Anomaly Detection Anomaly detection is the problem of identifying data points that don't conform to expected (normal) behaviour. Unexpected data points are also known as outliers and exceptions etc. Anomaly detection has crucial significance in the wide variety of domains as it provides critical and actionable information. For example, an anomaly in MRI image scan could be an indication of the malignant tumor or anomalous reading from production plant sensor may indicate faulty component.This thesis studies in detail various kind of anomaly detection techniques and compares them based on their effectiveness. Algorithms have been implemented in Python mostly.

What is Anomaly Detection? Anomalies or outliers come in three types. Point Anomalies. If an individual data instance can be considered as anomalous with respect to the rest of the data (e.g. purchase with large transaction value) Contextual Anomalies, If a data instance is anomalous in a specific context, but not otherwise (anomaly if occur at a certain time or a certain region. e.g. large spike at the middle of the night) Collective Anomalies. If a collection of related data instances is anomalous with respect to the entire dataset, but not individual values. They have two variations. Events in unexpected order (ordered. e.g. breaking rhythm in ECG) Unexpected value combinations (unordered. e.g. buying a large number of expensive items) In the next section, we will discuss in detail how to handle the point and collective anomalies. Contextual anomalies are calculated by focusing on segments of data (e.g. spatial area, graphs, sequences, customer segment) and applying collective anomaly techniques within each segment independently.

When there is no Training Data If you do not have training data, still it is possible to do anomaly detection using unsupervised learning and semi-supervised learning. However, after building the model, you will have no idea how well it is doing as you have nothing to test it against. Hence, the results of those methods need to be tested in the field before placing them in the critical path.

1.1 Motivation

In today's ever changing and dynamic world where technology keeps changing daily and a new attack surfaces every fortnight it has become a challenging task for users and administrators of any infrastructure to keep a track of anomalous behaviour of system.Which should be detected at initial level so that it could be predicted before an attack happens. A lot of cloud computing services have emerged in 21st century like Amazon's AWS,Microsoft Azure,Google Cloud Platform,Alibaba Cloud, Huawei Cloud,Oracle Cloud.Not only software developers but organizations,researchers,companies ,enterprises many govt organizations are using these on demand computing services.The computational power which can be offered via cloud service provider is much more than a stand alone system or an infrastructure set up by a small individual or organization can offer. But with more power more challenges have propped up in terms of security of these services.

1.1.1 First contribution

A detailed investigation and analysis of various attacks have been carried out for finding the cause of problems associated with various anomalies in detecting intrusive activities. Attack classification and mapping of the attack features is provided corresponding to each attack. Issues which are related to detecting low-frequency attacks using network attack dataset are also discussed and viable methods are suggested for improvement. Anomaly detection techniques have been studied and compared in terms of their detection capability for detecting the various category of attacks.

1.1.2 Second contribution

In anomaly detection the challenge is to identify a virtual machine or server that is running with malicious code. The cloud service provider has to provide just the underlying infrastructure. The customer has the responsibility to maintain the integrity of system running on his virtual machine. The challenge is that anomalies in data translate to important actionable information. This is a new area of research which is emerging and it has potential to further develop.

1.2 Objective of the thesis

This thesis explores what is anomaly detection, different anomaly detection techniques, discusses the key idea behind those techniques, and wraps up with a discussion on how to make use of those results.
To do anomaly detection following three conditions are needed:-
You have labeled training data Anomalous and normal classes are balanced Data should not be autocorrelated. That means one data point does not depend on earlier data points. This often breaks in time series data). If all of above is true, we do not need an anomaly detection techniques and we can use an algorithm like Random Forests or Support Vector Machines (SVM). However, often it is very hard to find training data, and even when you can find them, most anomalies are 1:1000 to 1:10$^{6}eventswhereclassesarenotbalanced.Moreover, themostdata, suchasda$

–**i** To be able to understand the anomaly detection techniques by existing service providers.
–**ii** To be able to implement an approach which can detect anomalies in a given data set.

1.3 Contribution of the thesis

In the current work we have tried to implement a clustering technique which analyzes the data set and try to find out anomaly in the given data set.

In summary, the areas of contribution of this thesis are anomaly detection in cloud environment.

1.4 Organization of the thesis

The rest of the thesis is organized as follows.
In **Chapter 2**, We reviewed literature for intrusion detection and various techniques which are used in anomaly detection.
In **Chapter 3**, In Chapter three we review the existing security practises by existing cloud computing vendors. We explored their implementation for anomaly detection.
In **Chapter 4**, In this Chapter we have described the approach used to detect anomalies in given data set.
Finally, we conclude in **Chapter 5**.

Chapter 2

Literature Review

Anomaly detection covers numerous things in cloud computing from security perspective. In order to understand them, the underlying concepts that might identify the source of vulnerabilities and threats must be understood. This section analyses those concepts, starting with an explanation on virtualization elements and then on multi-tenancy. Cloud services are also discussed, followed by the discussion of the concept of service providing in cloud and the section ends with a discussion on anomaly detection.

Four main models used for deployment of Cloud Computing are as follows [**?**]:

Private Cloud: It is set up by a single organization for its own use or for customers.

Community Cloud: The infrastructure is provided for use by organizations which can have
multiple customers.

Public Cloud: In this kind of deployment model general public is given access to resources in cloud on internet.

Hybrid Cloud: A hybrid cloud is a computing environment that combines a public cloud and a private cloud by allowing data and applications to be shared between them.

Security Issues with respect to Cloud Computing [**?**] have been widely discussed both in academics and industry several international conferences have focused on this subject :

Confidentiality
Virtualization Level Issues
Multi Tenancy Issues
VM Isolation Issues
Virtual Network Issues
Virtual Machine Introspection Issues
VM Management Issues
Application Level Issues
Isolation Issues
Synchronization Mechanism Issues
Data Storage Level Issues
Outsourcing Issues
Data Deletion Issues
Network Level Issues

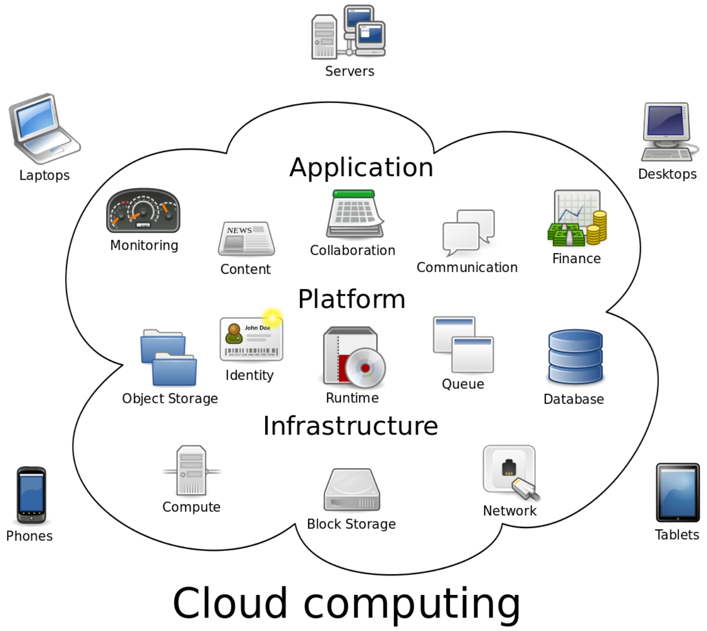


Figure 2.1: Cloud Computing [?]

DoS attacks, and DDOS attacks affect cloud computing infrastructure and services. These type of attacks are major attacks which affect the available services. The Cloud Infrastructure
provided by vendor could be a victim of such attacks but apart from that it could also be participating in such attacks. Botnets , botClouds can be deployed in Cloud Environment [?] to launch such attacks. Hence from security perspective it is important to identify such an issue that may exist in underlying cloud infrastructure.
Three main type of service models used in Cloud Computing are as following :
Infrastructure as a Service (IaaS)
Platform as a Service (PaaS)
Software as a Service (SaaS)
In [?] author discusses that with the advent of cloud computing a lot of data has been generated how this data has been impacting the world and there is a lot of growth in the data of devices that have been connected. So, with the improvements in bandwidth and availability. In the context of the Internet of Things, the trouble with the cloud is that data needs to be sent back from the sensors gathering info, such as a Nest thermostat or a Fitbit wristband, to a database in a remote public cloud. The time that it takes for the data to be transferred from the device or sensor to the remote public cloud, that is the latency, is often too great to meet the requirements of the IoT system. The cloud complicates this process even more.We're focused on centralized computing, thus there will be latency. Now, instead of sending the data back to the data centre on the other side of the factory, we send it to a remote cloud server that can be thousands of miles away. To make things worse, we send it over the open Internet.
In [?] discussion about data sharing has been made. As how participants must share data.

Protocols have played an important role. Protocols have played an important role in transfer of data in case of cloud computing. Storage has become a hot topic in today's cloud computing world. We prefer to store all types of data in cloud servers, which is also a good option for companies and organizations to avoid the overhead of deploying and maintaining equipment when data are stored locally. In cryptography, a key agreement protocol is a protocol in which two or more parties can agree on a key in such a way that both influence the outcome. This kind of protocol has widespread application in technology of internet and cloud computing.
In [?] mobile edge computing and emerging models in fog computing have been discussed. Relation between them is evident. The approach in the paper is to examine and underpin the models that are existing in cloud computing. Characteristics of cloud like support of ubiquitous connectivity, elasticity, scalable resources and ease of deployment have played an important role in development of existing cloud computing infrastructure. Research community has proposed new technologies namely fog and cloud. These technologies have been labelled in the paper as extended cloud they allow computing needs to be performed closer to source of data. This results in improvement in quality of services provided since this results in reduction of delay in conveying data between end nodes and cloud. Such technologies have enabled support for new application and services example Google now, foursquare and both are location aware applications for mobile platforms. Further this can be extended to services like autonomous vehicles robotics, public safety and augmented reality.

The acceptable level of service depends upon user expectations. Now these days users require rapid access to service like always on always available. So a new term has popped up known as resilience [?]. Resilience is concerned with availability of services and maintaining confidentiality and integrity of information in face of challenges. Resilience has become a fundamental property of cloud service provisioning platforms. With the advancement in wireless related technologies security and resiliency have become key issues when considering Mobile Edge Computing Services. With regards to edge model there are few threats also which have evolved. For example, infrastructure related threats, virtualization related threats, privacy related threats. Fog computing model was originally conceived by Cisco as an extension of cloud. The term fog was originally coined by Cisco as there is need to enable a platform that can cope up with the requirements posed by challenges put forward by Internet of Things. Another requirement in fog computing is privacy of data. In the paper detection and resilience mechanism have been discussed. The area that has been challenging to researchers is anomaly detection. In September 2016 [?] website of computer security consultant Brian Krebs was hit with 620 Gbps traffic. At the same time a bigger DDoS attack using Mirai malware was done on French web hosting and cloud service provider OVH. Mirai's source code was release by its creator soon after wards. Hackers offered Mirai's botnets for rent with as many as 400,000 connected devices. More attacks happened in October 2016 using Mirai they took down hundreds of websites like Twitter,Netflix,Reddit, Github for several hours. Mirai spreads by infecting devices as web cams,DVRs, routers, then it finds out administrative controls of those devices by a brute force attack which relies on a dictionary of potential usernames and passwords.

Chapter 3

First Contribution

We have studied various anomaly detection techniques deployed by cloud service providers for their services. The study from our side for various cloud service providers for anomaly detection is our first contribution for the project. In this study we took three major service providers Microsoft,Amazon and Google.

3.1 How Microsoft detects anomaly in azure

Microsoft uses a technology called security centre [?] in its cloud environments.To detect threats and reduce false positives Security Centre collects data from Azure resources and network and then applies a lot of machine learning and big data algorithms to detect threat.These techniques as more advance than normal signature based anomaly detection techniques.It is impossible to manually identify the attack and predict the attack when it might happen.So the Microsoft
Security Centre uses following technologies

Intelligent threat intelligence
It looks for bad actors by using the information obtained via Microsoft Products and services,Microsoft Digital Crimes Unit and Microsoft Security Response Centre along with external feeds.Microsoft has an immense amount of global threat intelligence. Telemetry flows in from multiple sources, such as Azure, Office 365, Microsoft CRM online, Microsoft Dynamics AX, outlook.com, MSN.com, the Microsoft Digital Crimes Unit (DCU), and Microsoft Security Response Center (MSRC). Researchers also receive threat intelligence information that is shared among major cloud service providers and feeds from other third parties. Azure Security Center can use this information to alert you to threats from known bad actors.

Behavioral analytics
Behavioral analytics is a technique that analyzes and compares data to a collection of known patterns. However, these patterns are not simple signatures. They are determined through complex machine learning algorithms that are applied to massive datasets. They

are also determined through careful analysis of malicious behaviors by expert analysts. Azure Security Center can use behavioral analytics to identify compromised resources based on analysis of virtual machine logs, virtual network device logs, fabric logs, crash dumps, and other sources. In addition, there's correlation with other signals to check for supporting evidence of a widespread campaign. This correlation helps to identify events that are consistent with established indicators of compromise.

Anomaly detection
It uses statistical techniques to build a historical data which is based on usage patterns and any deviation from normal alerts those deviations and it creates a baseline which if confirms to a potential attack vector then the particular usage is detected as an anomaly and in turn thus could be a security event.

Security Centre in Azure also works with connected partner solutions, like firewall and endpoint protection solutions. Microsoft uses **Fusion Analytics** [**?**] as the backbone of Security Centre's anomaly detection system.Fusion works by looking at various kind of alerts generated in Microsoft Azure ecosystem and then it tries to find pattern which could reveal attack progression indicating what should be next course of action.

3.2 How Amazon finds anomalies in AWS

Amazon uses a technology called guard duty [**?**].It is a threat detection service which continuously monitors for bad behavior and protects aws accounts and workloads.In the cloud collection and aggregation of network activities is simplified, but it can be time consuming for security teams to continuously analyze event log data for potential threats. With the help of GuardDuty now you have an intelligent and cost effective solution to for threat detection in AWS cloud. This technology uses machine learning, anomaly detection, and integrated threat intelligence to identify and prioritize potential threats. GuardDuty analyzes tens of billions of events across multiple AWS data sources, such as AWS CloudTrail, Amazon VPC Flow Logs, and DNS logs. With a few clicks in the AWS Management Console, GuardDuty can be enabled with no software or hardware to deploy or maintain. By integrating with AWS CloudWatch Events, GuardDuty alerts are actionable, easy to aggregate across multiple accounts, and straightforward to push into existing event management and workflow systems.

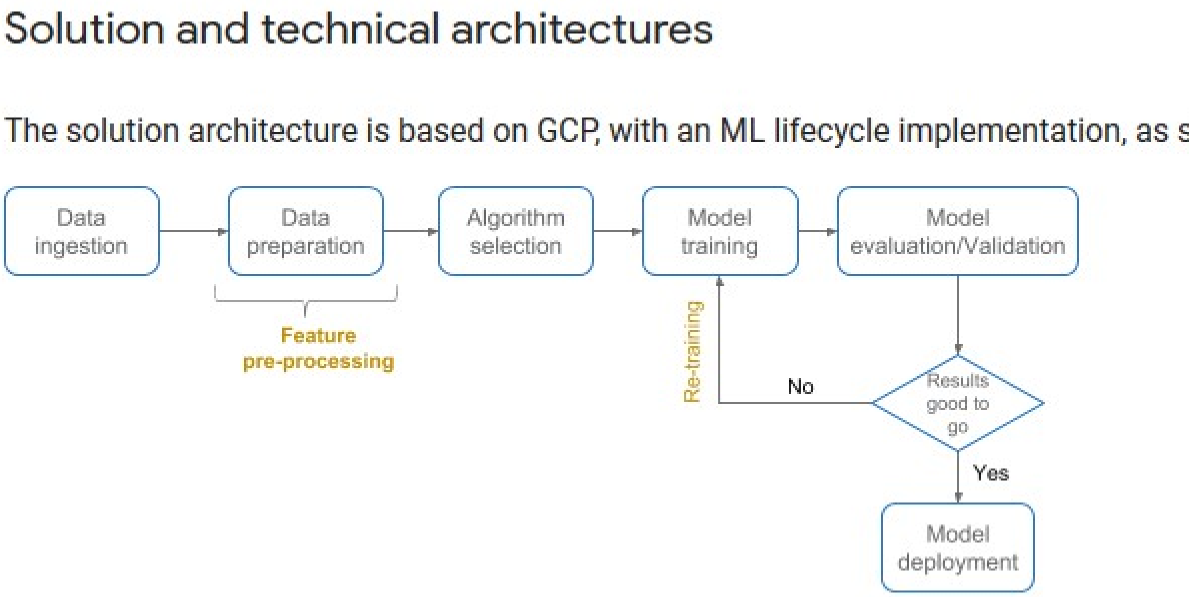
3.3 How Google finds anomalies in GCP cloud

Google uses an open source library Forseti [**?**] which can

- Detect unusual firewall behaviors between snapshots.
- Alert users to any unusual behaviors and provide a comparison with expected behaviors.

- Provide potential remediation steps.

The key elements for this technology are firewall rules.Firewall rules can be inbound or outbound.A firewall rule can either allow traffic based on IP address or ports.Firewall rules are applied to those instances which are associated with the user who has created his cloud in GCP setup. The figure below shows the technical architecture as how this has been implemented in GCP



Solution and technical architectures

The solution architecture is based on GCP, with an ML lifecycle implementation, as s

Figure 3.1: GCP Anomaly Detection

Chapter 4

Second Contribution

Software to find anomalies works at some level in the system to protect an electronic network from unauthorized users, which maybe insiders as well as outsiders. The task in anomaly detection is to build an algorithm or model which is capable of differentiating between *bad connections*, known as intrusions or attacks, and *good traditional connections*.
A connection may be a sequence of communications protocol packets beginning and ending at some well outlined times, between that information flows to and from a specified address to a target destination address beneath some well-defined protocols. Every connection is labelled as either traditional, or as an attack, with specifically one specific attack kind. Every connection record consists of concerning a hundred bytes.

Attacks make up four main categories:
- DOS: denial-of-service, e.g. syn flood;
- R2L: unauthorized access from a far off machine, e.g. guesswork password;
- U2R: unauthorized access to native superuser (root) privileges, e.g., numerous “buffer overflow” attacks;
- searching: law agencies investigation and different probing, e.g., port scanning.

Almost all novel attacks are square measure of variants of notable attacks and therefore the ”signature” of notable attacks may be ample to catch novel variants. supporting this concept, we are going to experiment with a few approaches. We will begin by acting on a reduced data set (the ten percent data set provided).

We will do some exploration of data using Panda library in python. Then we will try to build a model. Our model can simply classify entries into normal or attack. By doing so, we will generalise the model to attack varieties.
However, in our final approach we will use data analytics techniques for anomaly detection. We would like our model to be able to work well with known attack varieties and conjointly to present an approximation of the nearest attack kind. At the start we are going to do agglomeration victimization once more and see if we will beat our previous

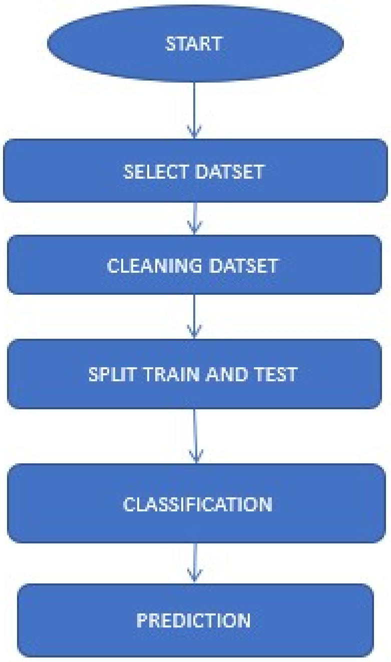
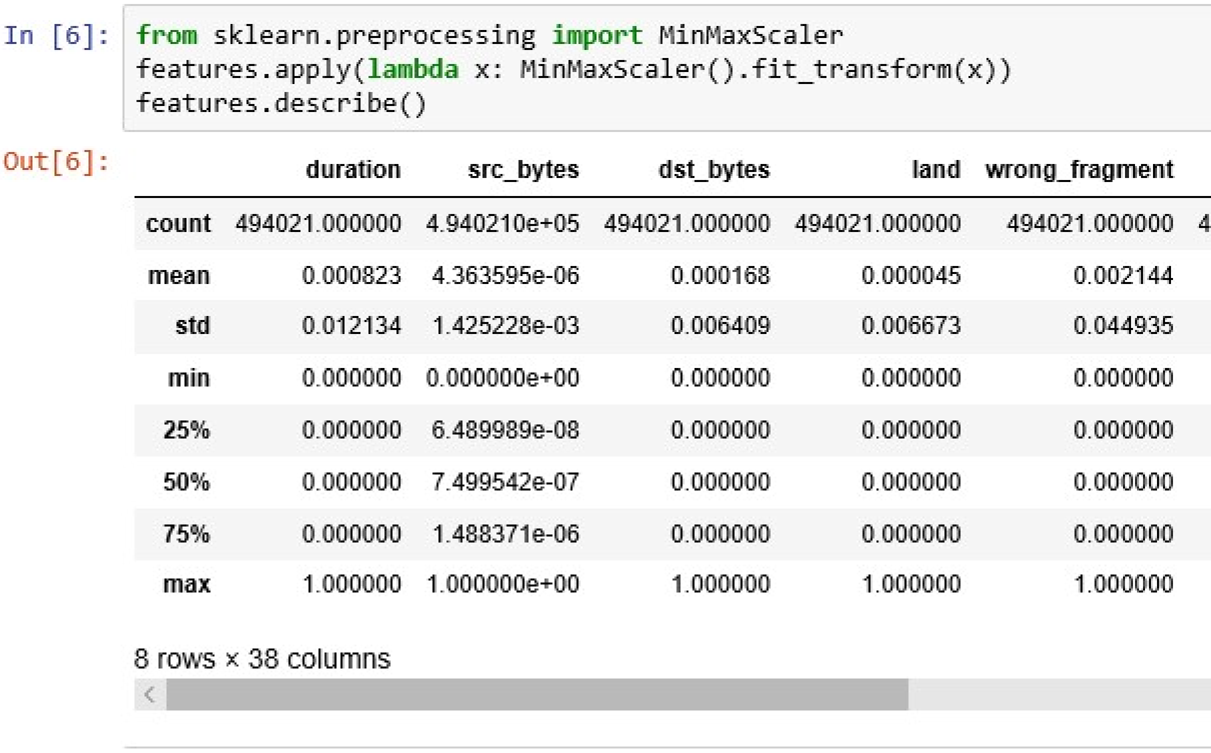


Figure 4.1: Flow Diagram



In [6]:

```
from sklearn.preprocessing import MinMaxScaler
features.apply(lambda x: MinMaxScaler().fit_transform(x))
features.describe()
```

Out[6]:

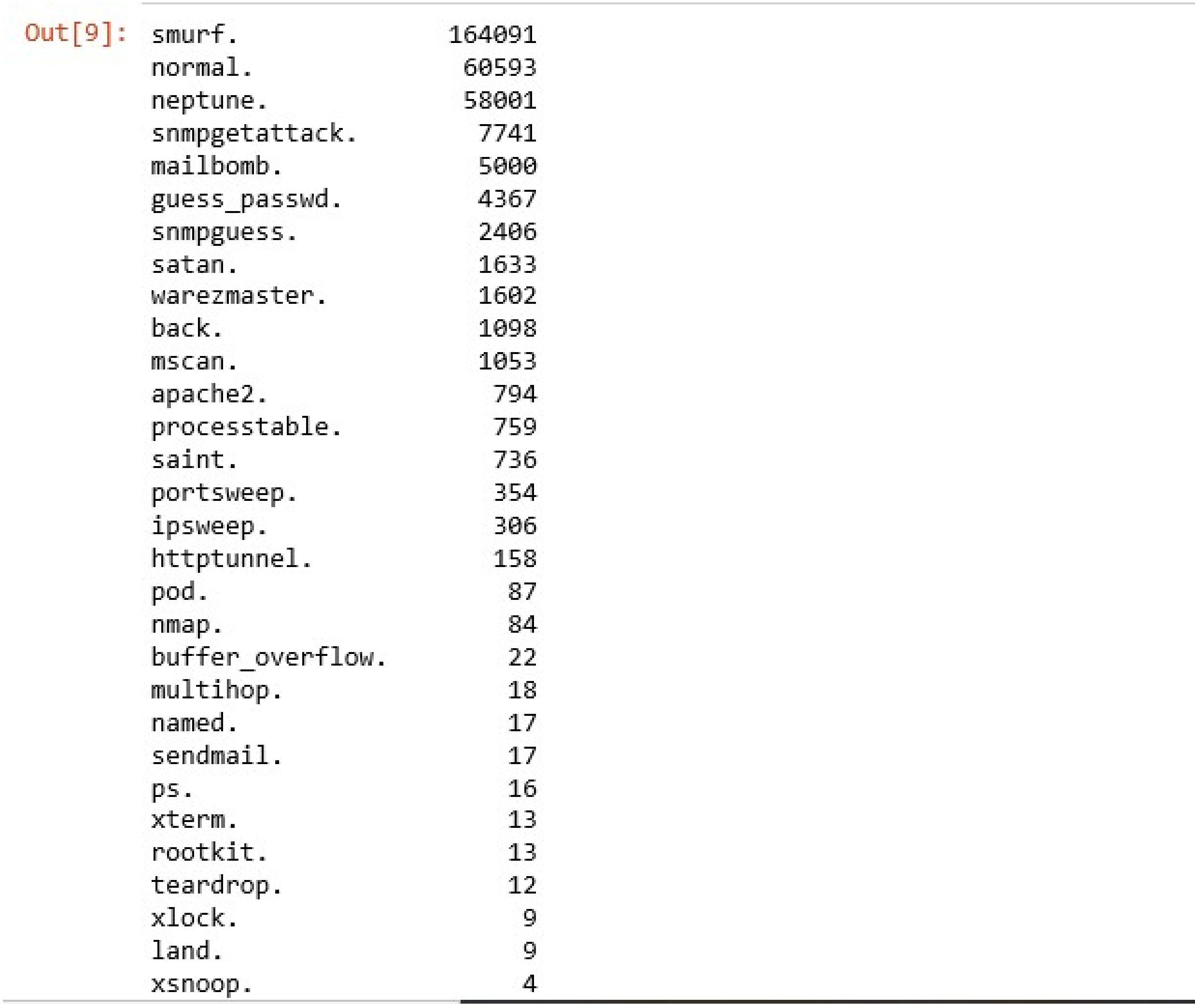
	duration	src_bytes	dst_bytes	land	wrong_fragment
count	494021.000000	4.940210e+05	494021.000000	494021.000000	494021.000000
mean	0.000823	4.363595e-06	0.000168	0.000045	0.002144
std	0.012134	1.425228e-03	0.006409	0.006673	0.044935
min	0.000000	0.000000e+00	0.000000	0.000000	0.000000
25%	0.000000	6.489989e-08	0.000000	0.000000	0.000000
50%	0.000000	7.499542e-07	0.000000	0.000000	0.000000
75%	0.000000	1.488371e-06	0.000000	0.000000	0.000000
max	1.000000	1.000000e+00	1.000000	1.000000	1.000000

8 rows × 38 columns

Figure 4.2: Feature Extraction

classification.

Visualising information principal elements By studying victimization information using principal part analysis, we will cut back the spatial property of our information and plot it into a two-dimensional area. The PCA can capture those dimensions with the most variance, reducing the knowledge loss. **Building a classifier** Following the concept that new attack varieties are just like notable varieties, let's begin by attempting a k-nearest neighbours classifier. we have a tendency to should to avoid brute force comparisons within the Nxd area in the least prices. Being n the amount of samples in our information quite , and d the amount of options , we are going to find a possible modelling method. Our try our anomaly detection approach within the reduced data set. We are going to begin by doing k-means agglomeration. Once we've got the cluster centre's, we are going to use them to see the labels of the check information (unlabelled). Based on the idea that new attack varieties can fit previous kind, we are going to be able to sight those. Moreover, something that falls too aloof from any cluster, are thought-about abnormal and thus a potential attack



```
Out[9]: smurf.              164091
        normal.              60593
        neptune.             58001
        snmpgetattack.        7741
        mailbomb.             5000
        guess_passwd.         4367
        snmpguess.            2406
        satan.                1633
        warezmaster.          1602
        back.                 1098
        mscan.                1053
        apache2.               794
        processtable.          759
        saint.                 736
        portsweep.             354
        ipsweep.               306
        httptunnel.            158
        pod.                    87
        nmap.                   84
        buffer_overflow.        22
        multihop.               18
        named.                  17
        sendmail.               17
        ps.                     16
        xterm.                  13
        rootkit.                13
        teardrop.               12
        xlock.                   9
        land.                    9
        xsnoop.                  4
```

Figure 4.3: Number of Attacks

Chapter 5

Conclusion

Anomaly Detection is one of the important security problem in today's world.A significant number of techniques have been developed which are also based on data analytics techniques.So to identify anomaly we have to understand the flow of system and comparing the data sets we can find out pattern in intrusions or attacks. Attack classification and mapping of attack should be done corresponding to each attack.Machine learning techniques have played an important role in finding anomalies in existing systems.Current systems are evolving and getting mature day by day new attacks surface and along with them new techniques are also evolving. Among the existing approaches how can we increase the accuracy of anomaly detection and reduce false positives in cloud computing is an area which will keep attracting researchers in coming years as more and more techniques evolve.

Author's Biography

Tapas received his B.Tech+MBA dual degree in Information Technology from ABV-Indian Institute of Information Technology and Management,Gwalior in 2009. He has been pursuing M. Tech at the Department of Computer Science and Engineering, IIIT Guwahati, since July 2018.

Publications made out of this thesis

(listed in reverse chronological order)

1. Publication 1 in standard reference format
2. Publication 2 in standard reference format