Performance Evaluation of Unsupervised Learning Techniques for Enterprise Toll Fraud Detection

BY: MEHADI ALIYE

ADVISER: EPHREM TESHALE (PHD)

A Thesis submitted to the School of Electrical and Computer Engineering
Addis Ababa Institute of Technology

in Partial Fulfillment of the Requirements for the Degree of Master of Science in Telecommunication Engineering

Addis Ababa University
Addis Ababa, Ethiopia
November 16, 2018

Declaration

I, the undersigned, declare that the thesis comprises my own work in compliance with internationally accepted practices; I have fully acknowledged and referred all materials used in this thesis work.

Mehadi Aliye

Name	Signature

Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

This is to certify that the thesis prepared by **Mehadi Aliye**, entitled *Performance Evaluation of Unsupervised Learning Techniques for Enterprise Toll Fraud Detection* and submitted in partial fulfillment of the requirements for the degree of Master of Science Telecommunication Engineering complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

Internal Examiner	_____	Signature	_____	Date	_____
External Examiner	_____	Signature	_____	Date	_____
Adviser	Ephrem Teshale (PhD)	Signature	_____	Date	_____
Co-Adviser	_____	Signature	_____	Date	_____

_____

Dean, School of Electrical and Computer Engineering

DEDICATION

This research work is dedicated to my mother Wro. Meymuna Yusuf, my father Ato Aliye Mohammed, my beloved wife Wro. Rabiya Yusuf and my son Ramin.

ABSTRACT

In recent years, the impact of telecom frauds is increasing and their behavior is changing through time which makes them to remain the main challenge for telecom service providers in terms of revenue loss and degradation of quality of service. Toll fraud occurs whenever a criminal uses cheating or dishonest means with the intention to use telephony services free of charge, reduced rate or to make money.

This study focuses on detection of toll fraud committed through enterprise PBX hacking by analyzing Call Detail Record (CDR) data. Unfortunately, this data is mostly unlabeled, meaning no indications on which calls are fraudulent or non-fraudulent exist. A clustering model was developed and tested with CDR collected from ethio telecom. In order to test the model with big dataset additional synthetic CDR was also used in the research. Two user profile are constructed by summarizing the data on daily basis. WEKA machine learning tool has been used to come up with a model for predicting fraudulent activities.

The experimentation result showed that, the model from the K-means algorithm exhibited higher accuracy level, and can be applied in toll fraud detection. Since it was able to detect unusual changes in calling patterns which are highly likely as a consequence of fraud. The implementation of the model will enable telecom operators in general and ethio telecom in particular to detect such fraud at minimum cost of operation. Moreover, it can also be used to support enterprise customers by showing security vulnerability of their Private Branch eXchange (PBX).

KEYWORDS

Toll fraud, Unsupervised Learning, K-means,Expectation Maximization (EM)

ACKNOWLEDGMENTS

Above all Alhamdulilalahi rabil alemim (Praise be to Allah, Lord of the Worlds), my gratitude goes to ALLAH who is in control of the existing and the coming world. He has been with me in all bad and good times and will always be.

I am also thankful to Dr. Ephrem Teshale (PhD) who helped me to reach this level and provided all the support I needed in all situations. He also gave me the chance to use all my effort and welcomed my request for advice, given his tight schedule. Without his support this research would not have been successful.

My beloved wife Rabiya Yesuf and son Ramin, I love you so much. I can't wait to give you the time you deserve as a family. I am grateful for your patience when I was busy and not able to give the time and attention you deserve. My special thanks also goes to my Father Ato Aliye Mohammed and my brothers and sisters specially Nuriya Mohammed. My Father, his belief in education and the price he paid for all his children above all his sublication whenever his family become in trouble is the reason for my success today.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AAiT	Addis Ababa Institute of Technology
ARFF	Attribute-Relation File Format
CBP	Current Behavior Profile
CBS	Convergent Billing System
CDR	Call Detail Record
CFCA	Communications Fraud Control Association
CSV	Comma Separated Values
DISA	Direct In-line System Access
EM	Expectation Maximization
FMS	Fraud Management System
FP	False Positive
FPR	False Positive Rate
FN	False Negative
IP-PBX	Internet Protocol Private Branch eXchange
IPRN	International Premium Rate Number
IRSF	International Revenue Shared fraud
ITU	International telecommunication Union
LOF	Local Outliers Factor
MLE	Maximum Likelihood Estimate
PBP	Past Behavior Profile

PBX Private Branch eXchange

PRS Premium Rate Service

PSTN Public Switch telephone Network

SMS Short Message Service

TP True Positive

TPR True Positive Rate

TN True Negative

VoIP Voice over Internet Protocol

1

INTRODUCTION

1.1 BACKGROUND

Telecom industries have recently experienced an increase in the number of occurrences of unauthorized users accessing customer's equipment to place illegal international calls. So that telecommunication fraud detection is an increasingly important and difficult task in today's technological environment. Telecommunication Fraud can be defined as any activity by which service is obtained with the intention of not to pay. The intention of the perpetrators is progressed from not willing to pay to make money [1][2]. This problem is a major source of revenue loss for telecommunications industry. The Communications Fraud Control Association (CFCA) an industry organization on a mission to reduce fraud against carriers, conducts every two-year Global Fraud Loss Survey. According to their latest survey the telecommunications industry experienced $29.9 billion in fraudulent charges in 2017 [3]. The report has also shows the first top five countries from which fraud originates are US, Pakistan, Spain, Cuba and Italy. The report also gives Cuba, Somalia, Bosnia & Herzegovina, Estonia and Guinea as the top five countries were fraud terminate.

There exist different types of telecommunications fraud and these can occur in various ways. Different literatures categorized frauds in different ways. According to, [4], [5] divide as subscription fraud and superimposed fraud as the highest fraud types. Similarly, [6] lists the top three types of telecommunication fraud that cause a significant loss to be International Revenue Share Fraud, Premium Rate Service Fraud and Bypass fraud. According to [7] classifies as Contractual, Hacking, technical and Procedural frauds. This thesis is interested to focus on detecting toll fraud committed through hacking customers PBX.

Toll fraud sometimes called VoIP fraud is a theft or unauthorized use of making a long distance call with the intent of not to pay for the charges or makes others pay for. Voice over Internet Protocol (VoIP) frauds primarily affect enterprises or companies that use or sell the voice services, for instance, companies use Internet Protocol Private Branch eXchange (IP-PBX) could be primary targets. According to the CFCA 2017 Telecom Fraud Survey, annual global PBX and IP-PBX hacking losses amount to an estimated $1.94 billion independently [3]. This amount shows how the toll fraud still the big problems that costs millions of dollars for an enterprise uses PBX and telecom providers as well.

Like other telecom providers, ethio telecom and its customers affected by different telecom frauds, the company has lost around $52 Million to fraudsters in 2017 [8]. To overcome fraudulent activities, the company implement Fraud Management System (FMS) . Based on the FMS document of ethio telecom FMS is currently operating using rule-based approach. The rule-based approaches, methods use the pre-defined rules defined by experts [9].

The research is motivated by a toll fraud case against an enterprise PBX system and further investigation of an enterprise fraud by looking at customer's CDR and will provide an overview of state-of-art on toll fraud detection analysis and practice. In addition, we will propose a possible knowledge base detection approaches.

1.2 PROBLEM STATEMENT

Enterprise customers usually use a PBX/IP-PBX to manage their internal and external communication needs. In most of the case enterprises install PBX system in their network to reduce the cost of installing Public Switch telephone Network (PSTN) line for individual staffs. Also installing PSTN line for individual staffs are very difficult to manage. PBX is a switching system used to provides extensions, i.e., an internal phone number to reach each user within the enterprise and also has a connection (called a trunk) with an operator to reach the PSTN or mobile networks [10]. PBX system evolving from time to time moving from circuit switching to packet switching. Recently due to advancement towards VoIP technology

legacy PBX is moving on the way to IP-PBX. However, this advancement has also come with some security vulnerabilities. Then fraudsters use these vulnerabilities as advantage for the way to commit toll fraud. Toll fraud is the largest threat [11] to an enterprise voice system.

Accurate cost estimates for toll fraud are difficult to pin down because many companies are reluctant to publicly admit that they have been targeted, experts worldwide estimate the costs to run in the billions of dollars annually [11]. According to [3] the most recent 2017 CFCA survey shows that global fraud loss estimate $29.2 billion and annual global PBX and IP-PBX hacking losses amount to an estimated $1.94 billion independently. This figure shows how the toll fraud still the big problems that costs millions of dollars for an enterprise uses PBX and telecom providers as well. Like other telecom providers, ethio telecom and its customers affected by different telecom frauds, the company has lost around $52 Million to fraudsters in 2017 [8].

The primarily victims of toll fraud are customers, because toll charges incurred for all calls made over the enterprise PBX system are the responsibility of the PBX owner, regardless of how those charges were incurred. Hence, telecom company's beside giving a better service they are responsible to protect their customers from financial loss. So that the implementation of an effective fraud detection and prevention tools and techniques are the most important task for telecom companies in order to protect their business as well as their customers from huge financial damage. Particularly those customers with PBX in their network needs special attention.

Most operators still use rule-based systems as their primary FMS tool to detect fraud. A rule-based approach uses predefined rules developed by experts and notification is generated when a rule is satisfied [9]. Rules can do an excellent job of known fraud patterns. However, the fraud behavior are not statics and their behavior will change frequently. So this method could be ineffective for a new type of fraud or even when the behavior of the existing fraud changed [9]. In addition, the rules are creating by placing an upper threshold on the amount of calls made, the call duration, the frequency of calls and many other thresholds. So based on the

threshold if the rules or set of rules are violated the system will notify the fraud analyst by generating alarms and further analysis and prevention made accordingly. However due to what the expert limited as a maximum more number of false alarm rate will introduce. This mean legitimate customer numbers will wrongly have detected and suspended. This may lead to increase customer complain. On the other hand, the fraudsters try different options to make a call with less than the threshold, this is also another weakness of using threshold. This is where machine learning becomes necessary for fraud detection. There are different types of machine learning techniques introducing for fraud detection mainly classified as supervised and unsupervised. Supervised techniques used when there is a pre classified or labeled data used for training. While unsupervised are not required labeled data and this technique has the ability to find new behavior from the data.

In this thesis we try to propose unsupervised clustering algorithm with better classification accuracy and try to prepare data accordingly. Therefor the overall study attempt to answer the below three research questions.

Research questions

1. What type of CDR data features or attributes can be preferable in order to predict toll fraud?
2. What kinds of machine learning algorithm and models can be more effective in order to detect toll fraud?
3. How effective would the proposed model in terms of toll fraud detection rates over the existing methods?

1.3 OBJECTIVE

1.3.1 *General Objective*

The general objective of this study is to compare the performance of different unsupervised learning algorithms (k-means and expectation maximization) in detect-

ing toll fraud in the telecom environment. Additionally, to evaluate the efficiency of models' detection rates over rule-based methods.

1.3.2 *Specific Objectives*

The specific objectives of this thesis are:

- To identify appropriate unsupervised learning model and algorithm for toll frauds detection
- To prepare the data for analysis by cleaning and transforming the data into suitable format for the selected algorithms
- To identify toll fraud properties and select the relevant attributes that will help to predict toll fraud
- To build a model that will able to classify fraudulent calls form legitimate calls by analyzing CDR data of PBX and IP-PBX users
- To recommend future works based on observation of the research findings

1.4 SCOPE AND LIMITATIONS

1.4.1 *Scope of the Study*

There are many fraud types that exist in the telecom sector. International Data Corporation has identified more than 200 forms of telecommunication fraud [12]. Although there are several ways in which toll fraud can be committed in an enterprise network, PBX/IP-PBX hacking is highest on the list [3]. Therefore, the study limits itself to illegal outgoing long distance calls that are made via PBX/IP-PBX through the Internet and analyzes it's CDR data that help to predict these fraudulent calls. We will take ethio telecom as a case study. The fraud detection is limited

to PBX/IP-PBX customers only. Similar frauds could be found other systems as well which could be conducted by other researchers.

1.4.2 *Limitation of the Study*

The researcher primarily planned to get CDR data from Anti-fraud section. Unfortunately, it was not possible to get the data from the FMS for security reason however the CDR from Convergent Billing System (CBS) database is used for this study. In addition, the contribution of data base experts was very helpful and vital. It would have been impossible to finish this research without their help and support. But the participation of the domain expert was limited or with some reservation in order to maintain their business secret.

1.5 CONTRIBUTIONS OF THE RESEARCH

At the end of this research, the result will have both practical and theoretical contribution. The research enables ethio telecom to detect frauds in relation to illegal international outgoing call made from customer's PBX system. And provide insight about the company FMS, shows the company uncovers critical areas in the detection process that the system was not able to explore. It also enables ethio telecom to provides useful information and practical suggestions for PBX customers that may help them to take appropriate security measure or closed their system until they have solved potential security vulnerability. It also indicates and provides the suitable algorithm and model that can predict illegal international outgoing call (toll) frauds with best accuracy. The accuracy level is determined by the type of algorithm and techniques used during the research. This research also could be the starting point to conducted researches in toll fraud detection. It could be used as a reference for further researches in the area and explore major issues related to the toll fraud detection in a an enterprise network for designing better detection model as a base and make it available for academic reference.

1.6 LITERATURE REVIEW

In this section research that has been conducted on telecom and toll fraud detection using different unsupervised learning techniques and algorithms are discussed. Unsupervised learning methods are used when there are no prior sets of legitimate and fraudulent data. Some of the literature focused on telecom and toll fraud detection are reviewed and presented.

1.6.1 *Telecommunication fraud detection*

Telecommunication fraud occurs whenever a fraudsters uses deception to receive telephony services free of charge or at low cost. It is a global problem that causes substantial annual revenue losses for telecommunication companies [1][2].

In [13] the scholars present a rule-based approach to detect anomalous telephone calls. The method described used subscriber usage CDR data sampled over two observation periods: study period and test period. The study period contains call records of customers' non-anomalous behavior. Customers are first grouped according to their similar usage behavior such as average number of local calls per week and so on. They come with a probabilistic model to describe their usage for customers in each group. Maximum Likelihood Estimate (MLE) was used to estimate the parameters of the calling behavior. Then the thresholds were determined by calculating tolerable change within a group. MLE was used on the data in the test period to estimate the parameters of the calling behavior.

1.6.2 *Toll fraud detection*

In [9] the scholars used Local Outliers Factor (LOF) as a solution to overcome problem on rule-based and neural network approach. The LOF is an outliers detection algorithm proposed by Breuniq et al. In general, the experiment results show that the proposed approach can be effective for outliers detection on VoIP service. How-

ever, the results show graphically and no information about true and false positive rate.

In [12], a detailed list regarding related work based on user profiling is provided and a method based on statistical user profiling is presented. Two user profiles containing statistical features are generated, representing the Past Behavior Profile (PBP) and the Current Behavior Profile (CBP), using a significant deviation of a user's behavior in contrast to his past behavior as an indication for possible fraud. Based on their findings True Positive Rate (TPR) of 90% and a False Positive Rate (FPR) of 1.22% is estimated.

In [14] the scholars used behavior pattern recognition based on the concept of clustering algorithms provided from user profiling. The grouping aspect of clustering algorithms regarding the similarity of objects leads to data depicting the behavior of a user to be matched against behavior patterns deviation from the assigned behavior patterns occurs, the call is considered fraudulent and the result of their findings shows that TPR measured is 98.4% and FPR is below 0.01 %.

In [15] the scholars used K-means to generate white-lists with call destinations, proposed white-lists are adaptive so they are changing according to user group's behavior over the time. White-list is a list of international destinations where every user is allowed to call. However, the results show graphically and no information about true and false positive rate.

1.6.3 *Summary*

As we seen from the above reviewed literatures all are focused on detecting toll fraud committed through VoIP network and even if they are using unsupervised learning model, they use some fraction of data that has fraudulent properties. No related work was conducted with unclassified data. Furthermore, the research conducted in this area is very limited. We have learn how customer profiling and data feature selection was benefited while using clustering algorithm. Therefore, in our study, first the data we are using is a real CDR data collected from telecom

company and some generated synthesize test data. In addition, additional feature that is not used in these related works are exercised.

1.7 METHODOLOGY

In this study, one of the machine learning technique called the unsupervised learning approaches are selected to reach to the desired knowledge. The researcher used clustering algorithms such as K-means and EM. These algorithms are selected based on previously made researches recommendations on fraud detection. Below is the steps following to perform the study.

1. Literature Review

   An extensive literature review covering research articles, books and industry white papers was performed. Not only formal researches but also other papers that have direct or indirect relation to this work were reviewed.

2. Data collection and preprocessing

   In order to get the data from ethio telecom, we have got a formal supporting letter from Addis Ababa Institute of Technology (AAiT) and deliver it to Chief Human Resource officer of ethio telecom. Accordingly, voice CDR of PBX users are collected and preprocessed. Then two user profile were constructed and four dataset were prepared. Finally the datasets were converted to appropriate format.

3. Tools and model experiment

   The models were experimented using the open source "Weka" machine learning tool. Four dataset using clustering models were tested. In addition to the clustering model the existing rule-based approach was tested with the same date use for the models.

4. Result Analysis and evaluation

   The output result were analyzed in line with toll fraud properties using database script. Then the results were evaluated applying variety of performance measure metrics. The other way of evaluating the models were

comparing the result found from the models with the existing rule based approach.

1.8 THESIS ORGANIZATION

This research paper contains six Chapters. Following this introductory Chapter, telecommunication fraud, its methods, types and specifically toll fraud including toll fraud are discussed in Chapter two. Then Chapter three covered the discussion on the machine learning methods that are used in this study. Chapter four is about data preparation, which deals with the data to make it ready for experimentation and analysis, experimentation, analysis and evaluation. The fifth Chapter focuses on experiment result and discussion. The last Chapter covered conclusions and recommended future work.

2

TOLL FRAUD

2.1 TELECOMMUNICATION FRAUD

Telecommunication fraud occurs whenever a fraudsters uses deception to receive telephony services free of charge or at low cost. It is a global problem that causes substantial annual revenue losses for telecommunication companies [1][2].

There are different types of telecommunications fraud exist and these can occur in many ways. There different category of telecommunication fraud that can be categorized by different scholars and authors. For instance, [16], [17] categorized subscription fraud and superimposed fraud as the most common fraud types, whereas [18] classifies as Contractual, Hacking, technical and Procedural frauds. Similarly, [3] can be divide in to major category, fraud method and fraud type. Fraud method is how they access the network or service to enable revenue gain from the attack and, fraud type is how they use the service or network to generate revenue from the attack . According to 2017 CFCA survey report, Subscription Fraud (Identity), PBX/IP-PBX Hacking, Subscription Fraud (Application) and Subscription Fraud (Credit Muling/Proxy) are among the first five top fraud method. Similarly, International Revenue Shared fraud (IRSF), Interconnect Bypass (e.g. SIM Box), Arbitrage, theft / Stolen Goods, Premium Rate Service (PRS) are among five top fraud types [3]. In this paper, based on CFCA fraud classification we have discussed PBX/IP-PBX from fraud methods and IRSF and PRS from fraud types

2.2 PBX/IP-PBX HACKING FRAUD

Enterprise customers usually use a PBX to manage their internal and external communication needs. A traditional PBX provides extensions, i.e., an internal phone number to reach each user within the enterprise [10]. This advantage PBX will be nothing if the PBX system compromises with hackers. Because one the hackers infiltrate the system they automatically change the advantage to make money out of it. Then the enterprise goes to loss big money. In short definitions PBX fraud is a method to make long illegal call.

2.3 INTERNATIONAL REVENUE SHARE FRAUD

IRSF fraud is committed by using someone else's phone service to inflates traffic by generating illegal calls to a high cost telephone number, typically an international premium rate with no intention to pay for the calls (or paying where there exists some form of arbitrage opportunity), or by stimulating calls by others to the number ranges. The victim receives a huge telephone bill for the unauthorized calls and the fraudsters collects a payout from the company that owns the premium rate number. In 2017 survey report the CFCA claims that traffic inflating fraud costs victims an estimated $6.10 billion each year worldwide. The IRSF fraud is explained in the following steps:

1. A fraudsters hacks into an enterprise's vulnerable PBX.
2. The fraudsters then uses the enterprise's PBX to make calls to premium rate numbers. The victim has the financial charge for these fraudulent calls and is obligated to pay for the service provider.
3. The service provider completes the international call to its wholesale international carrier. The service provider is obligated to pay the wholesale carrier for the calls. Once the call is handed off to an international carrier, it is likely that the call will traverse multiple carriers before being delivered to the pre-

mium rate number provider. Each carrier collects a share of the revenue and the premium rate number provider receives a fee for completing the call.

4. In the final step, the premium rate service provider shares part of its revenue with the fraudsters who generated the calls.

2.4 PREMIUM RATE SERVICE

A premium rate number is a telephone number that charges rates higher-than-normal rates and is designated for specific content services like adult chat lines, tech support, voting, or weather forecasts. In every country, the phone number regulatory authority allocates ranges for premium rate numbers, in addition to other number types like emergency, mobile, and shared cost ranges. Country number authorities are responsible for maintaining these allocations and reporting them to the International telecommunication Union (ITU), the international regulator for phone numbers. For example, 900 is the prefix for premium rate numbers based on the North American Numbering Plan in the United States [19].

Though there are specific number ranges in each country allocated for premium rate services, the "premium rate numbers" sold by International Premium Rate Number (IPRN) providers are very often not part of a country-defined premium rate number range. These numbers are often mobile numbers with prefixes deeply nested within legitimate number ranges. Because of this, it is not as easy as blocking ITU-defined premium rate number ranges within each country to mitigate the risk of toll fraud [19]. Fraudsters use all different types of numbers from all different countries in order to making preemptive blocking of numbers very difficult. Moreover, fraudsters regularly vary the ranges that they use.

2.5 TOLL FRAUD

The most common modes of telephone billing are the prepaid and postpaid. Prepaid mode is based on advance paying before the services are offered such as

airtime tokens. Most of enterprise business are on postpaid mode were calls are made and billing is at the end of the month. Consequently, fraudsters take advantage of this to make toll fraud. Toll fraud is the theft or unauthorized use of long distance phone service. Toll fraud takes many forms but is especially prevalent to phone systems that have not been secure, or where lax security measures are in place. Toll fraud is a problem worldwide, and fraudsters can easily rack up tens of thousands of dollars in long distance charges before the phone's administrator is even aware of a problem [20]. The increased use of IP-PBX in the enterprise environment makes it a new target of security attacks [21].

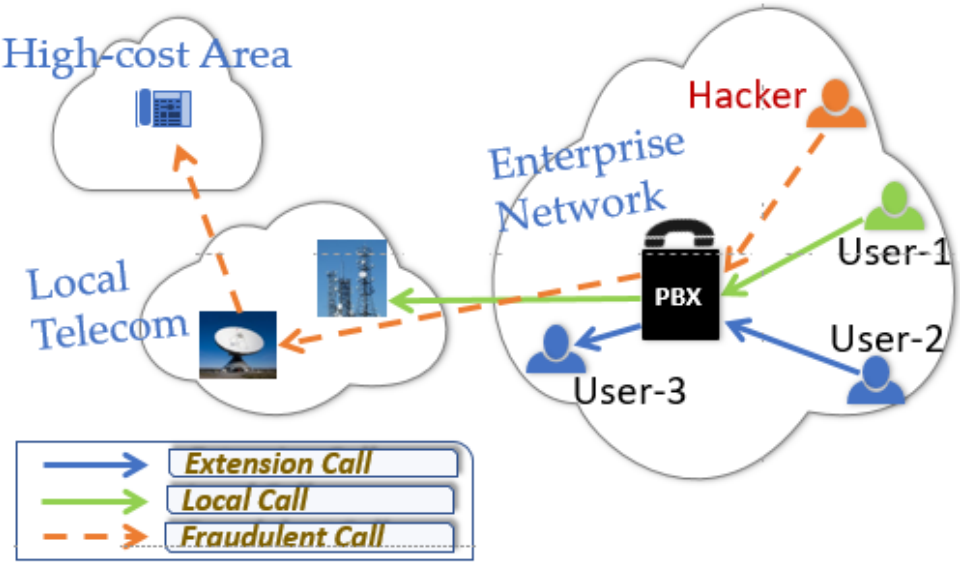


Figure 2.1: Toll fraud typical scenario adopted from [22].

2.5.1 *Toll fraud Scenario*

Fraudsters generally infiltrate the telephone system by using various techniques [23]. There are old traditional and modern PBX exchanges. So that the hacking methods are different, in case of old traditional PBX as shown in Figure 2.2 there five different ways of hacking each of them are discussed in detail. Figure Figure 2.1 shows a typical toll fraud scenario of an enterprise network modern PBX are less secure [24], if poorly administered, than the old traditional exchanges.

They are IP based and are connected to the public Internet, so that hackers will trawl the internet using specially designed scripts and looking for vulnerabilities like an open port. They are based in some case on open source software such as Asterisk. Once a weakness is detected, the hackers will try and authenticate their access and gain control of a PBX system. It is relatively simple for most fraudulent operators to access the telephone system if the passwords are easy to guess or if not modified the default passwords issued when the telephone system was activated [20]. Once the password has been cracked, they will have control over the PBX, then they can first test that the distant premium rate number is available and accessible with a few test calls. Then on a weekend, holidays and out of office hour, they can start to establish multiple calls to that expensive number, keeping them to a reasonable call length to avoid simple threshold of rule based FMS. They can forward calls to multiple numbers until eventually traffic increase are detected and the PBX owner are warned. Some more enterprising hackers prefer a slow and steady approach by sending a lower level of calls to the destination, but keeping it going for a very long period of time until the bill invoice arrives and the company realize, from the bill, what has been happening. Most of the time the PBX hacking approach are worked collaborate, hackers or fraudsters are hack the PBX and some of the profit shared with the fraudulent companies. Below are way of hacking the legacy PBX

Brute Force Attacks: Typically, the PBX security is breached via the toll-free Direct In-line System Access (DISA) number. Many PBXs allow dial-through, wherein a person calling into the PBX can access an external line by using appropriate passwords and control sequences. Large corporates use this feature extensively. Fraudsters hack into the PBX, obtain passwords and once a password is retrieved, use the PBX to generate outbound calls (typically for call selling or PRS revenue share). "War dialing" is one of the techniques used for obtaining passwords. It involves the fraudsters trying to break PBX code using an auto dialer, which keeps on dialing same number in a sequence making an exhaustive search of passwords until it breaks through.

Default Passwords: One of the largest contributing factors in PBX hacking and misuse is careless PBX installation and poor configuration, leaving default user and

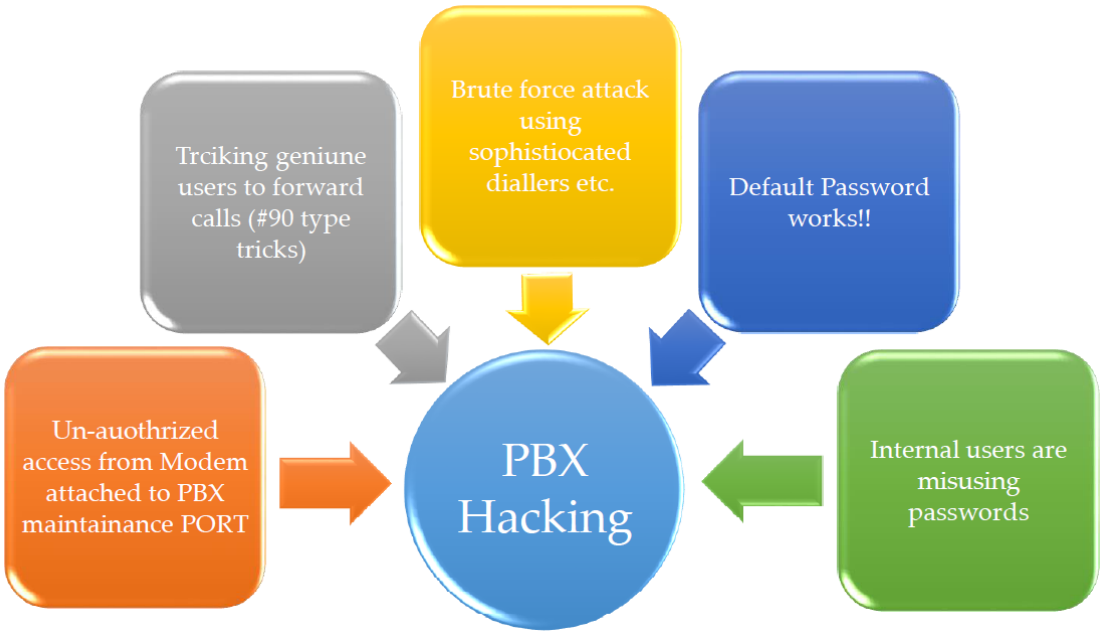


Figure 2.2: Ways of PBX hacking, adopted from [25] .

maintenance port passwords in place (fraudsters know the default passwords for the various switch vendors).PBX fraud commonly occurs when the customer fails to change the default password or do not change passwords frequently. Default passwords can be found on-line in the relevant PBX user manuals etc.

Internal Enemy: Many cases of PBX hacking result from insiders or vendors who disclose the phone numbers, IDs and passwords necessary for breaching PBX security. Sometimes the users may get hold of passwords by unauthorized means and use the corporate lines for making personal calls or colluding with external fraudsters to help in PBX hacking. Strict security policies should be in force for PBX password control. The physical security of PBXs and phone extensions is also an important factor to consider in avoiding misuse of PBXs.

Social Engineering: The old traditional (wired) phone scam involving the 90# buttons on corporate telephone lines is still around. Employees of a corporation using a PBX line from their desk, receives a call from someone claiming to be a telephone company employee investigating technical problems with line, or checking up on calls supposedly placed to premium rate services or other countries from your line. The caller asks the employee to aid the investigation by either dialing 90# (or similar combination) or transferring him/her to an outside line before then hang-

ing up the telephone receiver. By doing this the employee will be enabling the caller to place calls that are billed to the corporate telephone account. This attack only works on few PBXs today. Social engineering is another common technique used for obtaining passwords.

Service Port: PBX hackers may also target modems attached to the service PORT of a PBX. The facility is provided by PBX manufacturers to allow remote support of the PBX. Typically, the connection should be opened only when an authorized request goes from the PBX customer to the PBX vendor, but many PBX customers keep the connection always open and therefore prone to attack.

2.5.2 *Toll fraud call properties*

Toll fraud generally involves a third party making frequent long duration calls at the expense of an enterprise. Calls made from an enterprise PBX would be classified as either fraudulent or legitimate based on the following characteristics of the call:

1. **Caller Number-** This attribute shows the subscriber number that made the call. Legitimate callers should always bear the correct number within the company numbering plan.

2. **Destination Number-** This attribute specifies the number to which the call finally gets presented. This number should also be within the company's calling space for the outgoing call to be considered genuine. Fraudulent caller calls to fraudulent PRS number and also terminated on high cost area. The highness of call cost can be determined on which way (root) the call are terminated. let's see by scenario, if someone (2511134XXXXX) made call to some destination like Somalia (252XXXXXXX) he can use the normal path which cost him/her 20 birr per minute. However the caller select other way like satellite call which cost him/her 110 birr per/minute.

3. **Call Frequency-** This is the number of calls made from a particular source to a particular destination within a specific CDR dump. Fraudulent callers

make many number of calls than legitimate callers and those calls are to distinct numbers.

4. **Call Duration-** This is the total duration of calls made from a particular source to a particular destination within the specified CDR file. Fraudulent callers call durations are longer than legitimate, approximately more than 25 minute per call.

5. **Call Date and Time-** This is the days and time calls made from a particular source to a particular destination. The fraudulent caller's calls are mainly on weekend, out of office hour and on holidays.

3

MACHINE LEARNING

3.1 MACHINE LEARNING

The idea of learning is hard to define and is assimilated with terms as "to gain knowledge, or understanding, by study, instruction, or experience" [26]. By understanding how humans learn using artificial intelligence try to develop methods for accomplishing the acquisition and application of knowledge algorithmically (i.e. on computers), naming this machine learning. According to [27] [28], machine learning is type of Artificial Intelligence (AI) that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. Developing computer programs that detect a meaningful pattern from many data. Therefore, in machine learning, concepts and results from many fields can be found, including statistics, artificial intelligence, philosophy, information theory, engineering, biology, cognitive science, computational complexity, and other disciplines [26].

There different kinds of machine learning algorithms to discover patterns in big data that lead to actionable insights. At a high level, these different algorithms can be classified into three groups based on the way they "learn" about data to make predictions: supervised, unsupervised and reinforcement learning methods.

3.2 SUPERVISED LEARNING

Supervised machine learning is a most common learning approach where the model is attempt to learn from fraudulent and non-fraudulent examples; this is often referred to as pre-defined labeled data [29]. To train a supervised model, pre-

classified data as either fraudulent or non-fraudulent are needed, and the model then attempts to infer a function or instruction set that can predict whether fraud is present by applying it to new examples. Common supervised machine learning methods include logistic regression, neural networks, decision trees, gradient bosting machines, random forests of trees, support vector machines and many more.

There is limitation in supervised learning [30], it requires experts with high domain knowledge on the class of each training sample. If there is a fraudulent calls misclassified as legitimate then the constructed model will be tricky and this could leads the model to increase number of false positive rate. The same happens for a legitimate calls which is misclassified as a fraudulent and this could also lead to customer complain.

3.3 REINFORCEMENT LEARNING

Reinforcement learning is an important fast-growing concept and producing a wide variety of learning algorithms for different applications. It's a type of Learning where an agent learn how to behave in a environment by performing actions and seeing the results [31]. In order to produce intelligent programs (also called agents), reinforcement learning goes through the following steps:

1. Input state is observed by the agent.
2. Decision making function is used to make the agent perform an action.
3. After the action is performed, the agent receives reward or reinforcement from the environment.
4. The state-action pair information about the reward is stored.

3.4 UNSUPERVISED LEARNING

Unsupervised methods are used when there are no prior sets of legitimate and fraudulent observations [32], [33]. Unlike supervised machine learning, unsupervised machine learning methods cannot be directly applied to a regression or a classification problem. It focuses on finding hidden patterns in data. Data contains no output values which means that the purpose of these algorithms is to find patterns in the data that can help to give a structured representation.

Unsupervised learning tries to auto associate information from the inputs with an intrinsic reduction of data dimensionality or total amount of input data. Unsupervised learning is solely based on the correlations among the input data, and is used to find the significant patterns or features in the input data without the help of expert. There are some issues related to unsupervised learning, it is harder as compared to supervised Learning tasks. How do we know if results are meaningful since no answer labels are available? and this let us the experts to look at the result.

Since no labeled data were provided for this study, we are interested in the unsupervised learning category. Considering that a majority of the data should not be fraudulent [34], we aim at finding suspicious behavior in our dataset, i.e. data points which are significantly different from the others, also called outliers. If the dataset is big enough and the frauds in minority, we can expect those outliers to be either suspicious or fraud case.

3.5 CLUSTERING

Clustering is the process of grouping sets of objects in a way that data points with common characteristics are associated to the same "cluster". Data points in the same cluster are more similar to each other than data points in another clusters [35]. Many times clustering are chosen by researchers in the field of intrusion and fraud detection [35]. The main advantage of clustering algorithm is the ability

to learn from and detect suspicious (fraud) call in the CDR data without explicit description of fraudulent properties which usually provided by fraud experts.

Clustering-based fraud detection approach, which is called unsupervised clustering where the fraud detection model is trained using unlabeled data that consists of both legitimate as well as fraudulent calls. The idea behind this approach is that suspicious or fraudulent data forms a small percentage of the total data. Based on this assumption, suspicious and fraud can be detected based on cluster sizes, large clusters correspond to legitimate data and the rest of the data points, which are outliers, correspond to suspicious [36].

The clustering process shown in Figure 3.1 are describe how the data are clustered. First, the row data which is dirty by nature are collected, then these collected data are preprocessed. Second based on the attributes data are put on matrix format then, clustering algorithm will cluster the input data based on the number of cluster specified by the user.

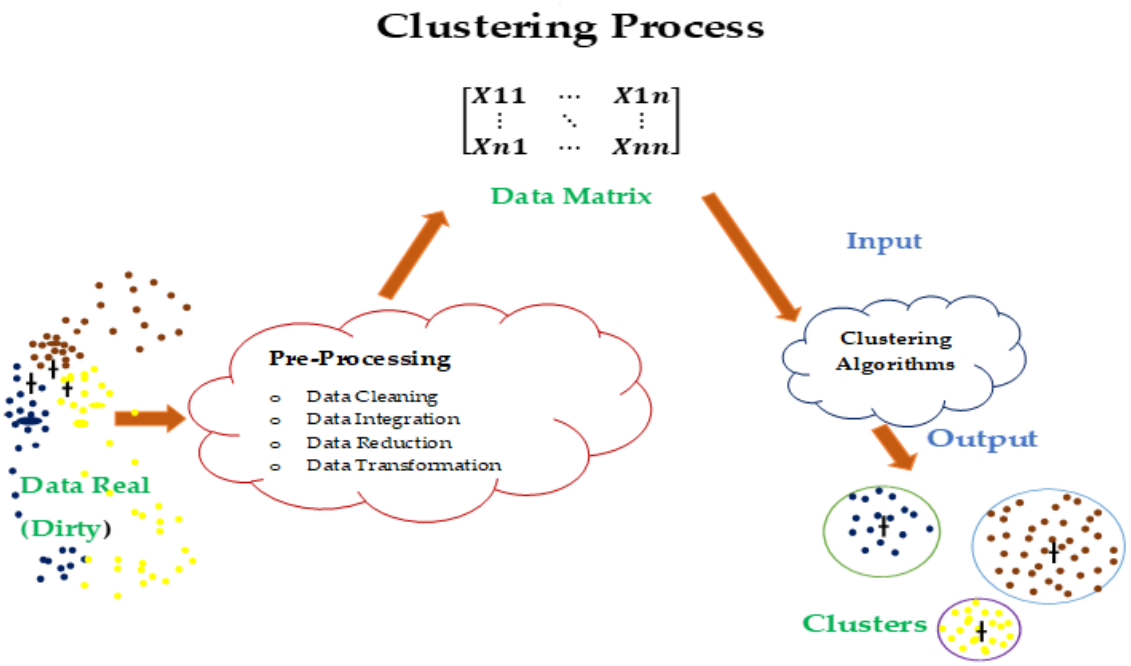


Figure 3.1: Clustering Process adopted from [37].

3.6 TYPE OF CLUSTERING

Cluster analysis is all about the kind of algorithms that can be used to find clusters automatically given the data. Clustering algorithms can be broadly classified [38][39] into three categories, the detail discussed in the following subsections:

- Partitioning Clustering algorithms
- Hierarchical Clustering algorithms
- Density-based Method

3.6.1 *Partitioning clustering*

A partitioning clustering a simply a division of the set of data objects into non over lapping group (clusters) such that each data object is in exactly one group. It discovers the grouping in the data by optimizing a specific objective function and iteratively improving the quality of partitions. Partitioning clustering methods are useful for the applications where a fixed number of clusters are required and further divided it into numerical methods and discrete methods. Partitioning algorithms assign a number of data into k number of clusters [40].

Partitioning tries to directly decompose the data set into a set of disjoint clusters. The criterion function that the clustering algorithm tries to minimize may emphasize the local structure of the data, as by assigning clusters to peaks in the probability density function, or the global structure. The global criteria involve minimizing some measure of dissimilarity in the samples within each cluster [30], while maximizing the dissimilarity of different clusters. A commonly used partitioning clustering method, represented by the centroid of the cluster, e.g. k-means, or by the closest instance to the medoid , e.g. k-medoids. Typically, k seeds are randomly selected and then a relocation scheme iteratively reassigns points between clusters to optimize the clustering criterion. The minimization of the square-error criterion - sum of squared Euclidean distances of points from their closest cluster centroid, is the most commonly used.

3.6.2 *Hierarchical clustering*

Hierarchical clustering, as the name suggests is an algorithm that builds hierarchy of clusters. The hierarchy can be formed in top-down (divisive) or bottom-up (agglomerative) fashion and need not necessarily be extended to the extremes [39]. This algorithm starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster. In the end, this algorithm terminates when merging or splitting stops once the desired number of clusters has been formed. Typically, each iteration involves merging or splitting a pair of clusters based on a certain criterion, often measuring the proximity between clusters. Some representative examples are: Clustering Using Representatives (CURE), CHAMELEON and so on.

3.6.2.1 *Density-based Method*

Density-based clustering methods group neighboring objects into clusters based on local density conditions rather than proximity between objects [39]. These methods regard clusters as dense regions being separated by low density noisy regions. Density-based methods have noise tolerance, and can discover non-convex clusters. Similar to hierarchical and partitioning methods, density-based techniques encounter difficulties in high dimensional spaces because of the inherent scarcity of the feature space, which in turn, reduces any clustering tendency [41]. Some representative examples of density based clustering algorithms are: Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Density-based Clustering (DENCLUE) and so on.

3.7 CLUSTERING ALGORITHM

Clustering is a task for which many algorithms have been proposed. Different techniques are in favor for different clustering purposes [41]. So an understanding of both the clustering problem and the clustering technique is required to apply a

suitable method to a given problem. Every methodology follows a different set of rules for defining the 'similarity' among data points. There are many number clustering algorithms known. however, in this specific study, two different clustering algorithms in suspicious (fraudulent) call detection model which are used popularly k-Mean and EM clustering algorithms has been discussed in detail:

3.7.1 *K-means Algorithm*

K-means clustering [42] is a well-known and widely used to automatically partition a data set into k groups [32]. It is unsupervised learning algorithm, which is used when you have unlabeled data (i.e., data without defined categories or groups). It is one of the simplest clustering algorithms in machine learning which can be used to automatically recognize groups of similar instances/items/objects/points in data training. The algorithm classifies instances to a pre-defined number of clusters specified by the user (e.g. assume k clusters). The first important step is to choose a set of k instances as centroids (centers of the clusters) randomly, usually choose one for each cluster as far as possible from each other. Next, the algorithm continues to read each instance from the data set and assigns it to the nearest cluster. There are some methods to measure the distance between instance and the centroid but the most popular one is Euclidean distance. The cluster centroids are always recalculated after every instance insertion. This process is iterated until no more changes are made. The k-Means algorithm is explained in this following pseudo code.

1. Select the total number of clusters (k)
2. Initially, random k points need to be chosen in the beginning and set as centroid

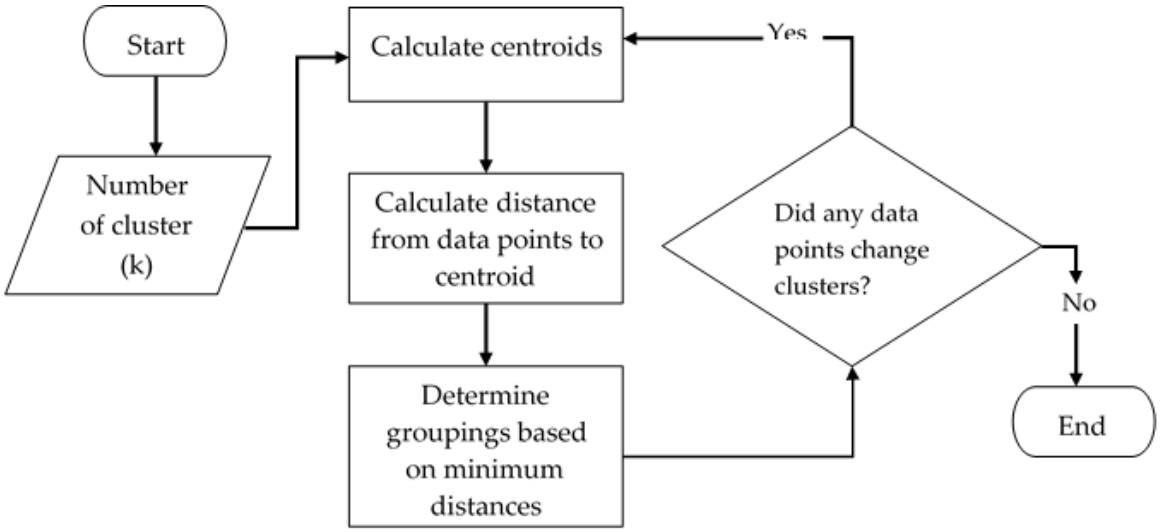


Figure 3.2: The K-Means Algorithm, adopted from [43].

3. Calculate the distance from each instance to all centroids using Euclidean method

$$D = \sqrt{\sum_{k=1}^{N} (X_i - Y_i)^2} \tag{3.1}$$

4. The next step is to take instances or points belonging to a data set and associate them to the closest centers.

5. Recalculate the positions of the centroids

6. Repeat step $3-5$ until no more changes are done

Finally, this algorithm aims at minimizing intra cluster distance (cost function also known as squared error function), automatically inter cluster distance will be maximized.

Main advantages:

1. K-means clustering is very ast, robust and easily understandable. If data set is well separated from each other data set, then it gives best results.

2. The clusters do not overlapping character & are also non-hierarchical within nature.

3.7.2 *Expectation Maximization Algorithm*

EM clustering is a variant of k-Means clustering and is widely used for density estimation of data points in an unsupervised clustering [32]. In the EM clustering, we use an EM algorithm to find the parameters which maximize the likelihood of the data, assuming that the data is generated from k normal distributions. The algorithm learns both the means and the covariance of the normal distributions. This method requires several inputs which are the data set, the total number of clusters, the maximum error tolerance and the maximum number of iteration. The EM can be divided into two important steps which are Expectation (E-step) and Maximization (M-step).

1. E-Step:

   The goal of E-step is to calculate the expectation of the likelihood (the cluster probabilities) for each instance in the dataset and then re-label the instances based on their probability estimations. The mathematical formulas of EM clustering are as follows:

$$P(C_j|x) = \frac{|\sum_j(t)|^{-1/2} \exp^{nj} P_j(t)}{\sum_{k=1}^{M} |\sum_j(t)|^{-1/2} \exp^{nj} P_k(t)} \tag{3.2}$$

2. M-step:

   The M-step is used to re-estimate the parameters values from the E-step results. The outputs of M-step (the parameters values) are then used as inputs for the following E-step. These two processes are performed iteratively until the results convergence. The mathematical formulas of EM clustering are as follows:

$$P_t(t+1) = \frac{1}{N} \sum_{k=1}^{N} P(C_j|X_k) \tag{3.3}$$

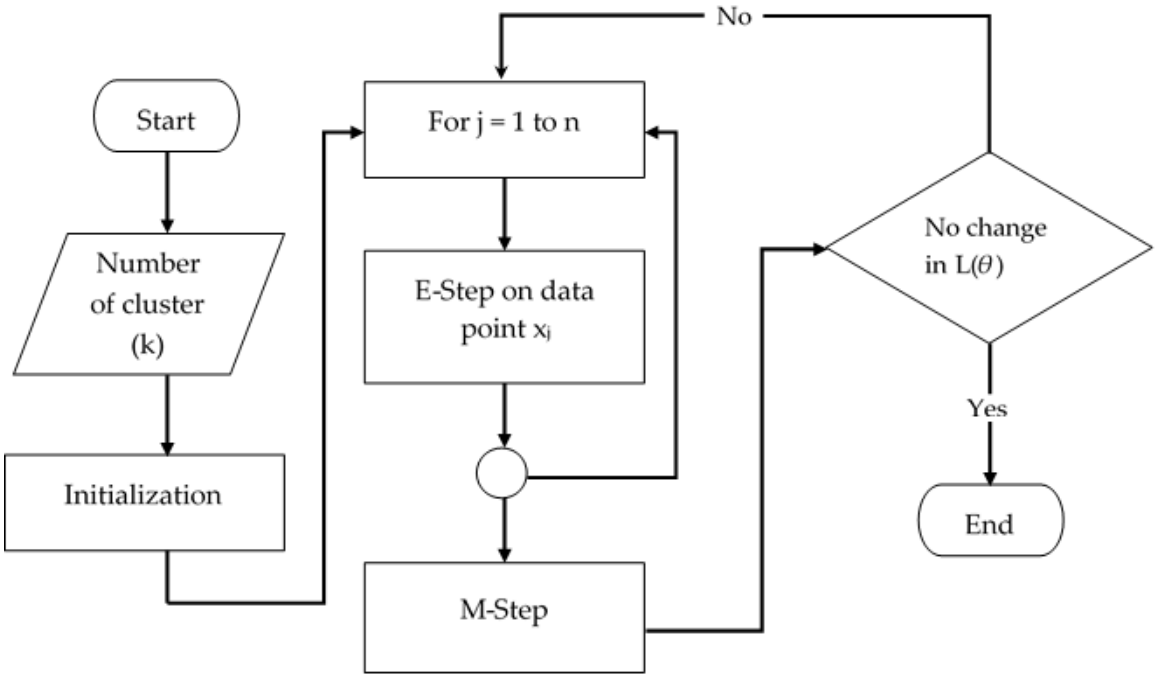


Figure 3.3: The Expectation Maximization Algorithm,adopted from [43].

4

CHAPTER FOUR: TOLL FRAUD DETECTION MODEL

4.1 INTRODUCTION

In this Chapter, the methodologies used in the construction of the toll fraud detection model are discussed in detail. Figure 4.1 refers to how the research was structured in order to meet the research objectives, from data collection until the model experimentation. The next chapter will discuss on the result found.

In Section 4.2, the data collection and preparation discusses in detail and the subsections start from, data selection, data preprocessing, synthetic test data generation, user profiling and data formatting. Section 4.3 present the experiment conducted, how the optimal number of cluster(k) determine with the reason behind the selection, and how the clustered data are further analyzed. Section 4.4 describes the evaluation techniques used to determine the performance of the model. Finally, Section 4.5 will conclude the chapter.

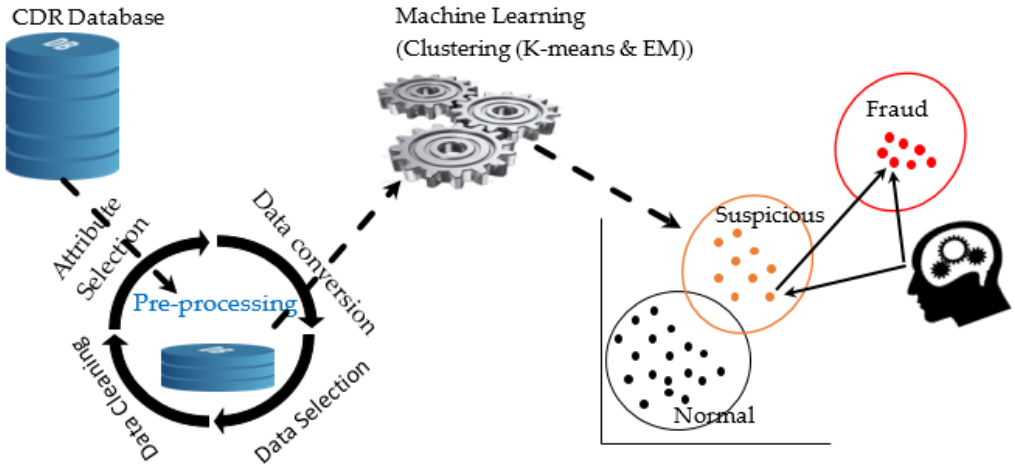


Figure 4.1: System Model.

4.2 DATA COLLECTION AND PREPARATION

4.2.1 *Data collection*

As mentioned in subsection 1.7, the data required for the study has been requested from ethio telecom using a formal cooperation letter from AAiT. Following the approval of data access request from the responsible division, department and section, the required data was collected from ethio telecom information system division. The data is found on CBS database. It is a huge billing database system used to process all billing data of the services provided by the company. Due to the bulkiness of the data, the first preprocess has been done while collecting the data by clearly specify what kind of data are we looking for. The data type need for this research is a voice CDR data. A CDR is the information captured by the telecom companies during Call, Short Message Service (SMS), and Internet activity of a customer [44]. Voice CDR as the name indicates, holds each and every record of calls made by the customer with details like calling number, called number, date and time of call, duration, amount charged and other details.

The study is focused on detecting fraudulent activities performed through PBX hacking on an enterprise network and the company has 4998 customer categorized under PBX users. So that only CDR records of these specific customers are needed. Six month (February to July 2018) period total of more than five million CDR records are collected from CBS database server. The collected data was stored on the temporally server provided by the IT operation department for further studied, selected, preprocessed and transformed to the appropriate format. The detail processes are presented on the subsequent subsections.

4.2.2 *Data Selection*

The first step for data selection is to establish the criteria on which the data will be chosen [45]. Our ultimate target is to study on how to detect fraudulent activities

performed through PBX hacking on an enterprise network, specifically on illegal outgoing international call using a CDR data. Therefore, informal interviews were held with domain experts from billing and fraud management sections. Using the knowledge gathered through discussion and by consulting literatures [12], [46] we have been set some criteria in order to identify what we are looking for, and this criterion was:

1. Since calling or originating number are already identified, so that called or destination are used as a criterion. The criteria are selecting all the destination number where the area code is different from +251.
2. To remove the incoming calls, the criteria are selecting all the destination number where the numbers are those 4998 PBX customer numbers.

The above two criteria are written using database script and after processed it, we have found 65633 CDR records. So that the data size is tremendously decreased. Then, these selected CDR records are further preprocessed.

4.2.3 *Data Preprocessing*

Row data are generally incomplete, noisy, and inconsistent [45]. So one of the important stages of data mining or machine learning is preprocessing. Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the experiment. There are a number of different tools and methods used for preprocessing,for this study there are number of oracle database scripts used. The scripts are found on Appendix A.

- **Data cleaning:** fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
- **Data integration:** using multiple databases, data cubes, or files.
- **Data transformation:** normalization and aggregation.

- **Data reduction:** reducing the volume but producing the same or similar analytical results.
- **Data discretization:** part of data reduction, replacing numerical attributes with nominal ones.

4.2.3.1 *Data Cleaning*

The presence of missing data has to be investigated carefully while performing clustering analysis, because the existence of missing value in the data will be replaced with some estimates [26]. Accordingly, the received CDR are checked and found 500 records with missing values and discarded them.

4.2.3.2 *Feature selection and data reduction*

Feature selection is the process of using domain knowledge to choose which data metrics to input as features into a machine learning algorithm. It is the most important but undervalued step of machine learning and, it is an important part of building statistical models and algorithms [47]. Feature selection plays a key role in clustering; using meaningful features that capture the variability of the data is essential for the algorithm to find all of the naturally-occurring groups. In addition, it reducing complexity that might hamper the analysis and increase fraud detection effectiveness. A number of attributes were reduced, and the selection process has been done with the help of domain experts and by consulting different literatures [48][12][46].

Feature selection started by removing attributes having zeros or ones and other constant values, redundant value containing attributes, like billing number and calling number, only one of them is used. For instance, the billing number is 113693551 and the calling number is 251113693551. Due to this fact, only calling number has been selected. In addition, non-relevant attributes are also removed by consulting related literatures and domain experts. Due to this from the total of 33 attributes only 6 attributes are selected. Table 4.1, describes the attributes

selected with their corresponding data type and description and all attributes of CDR are listed in Appendix A

Table 4.1: CDR Attribute Fields, Data Types and Description

No.	Attributes	Value	Descriptions
1	Calling Number	Number	Fixed line number initiating or originating the call.
2	Called Number	Number	International call received the call
3	Call date	Date	Data of call initiation
4	Call start time	Time	Time of call initiation (calling time).
5	Call stop time	Time	Time of call terminated
6	Duration	Number	Call duration in second
7	Charge(cost)	Number	Amount of paid in cent
8	Call Type	Number	Forward call or not

4.2.3.3 *Data transformation*

The field for start day and time is transformed. As shown in Table 4.2 and Table 4.3 the date is converted to weekend, holiday and office day also the time was converted to office hour (8am to 12pm) and out of office hour (12pm to 8am) periods. Office day is represented by 1, weekend and holiday are represented by 2, office hour is represented by 1, and out of office hour represented by 2.
The field duration values originally was on seconds and converted to minutes, and the filed call fee values was also on cents and converted to birr. This is to minimize the overhead for the algorithms.

Table 4.2: Transformed call days

Day	Represented value
Office day	1
Weekend, Holiday	2

Table 4.3: Transformed call times

Time	Period	Represented value
Office hour	8am–12pm	1
Out of office hour	12pm–8am	2

4.2.4 *Synthetic test data generation*

Since the data we have found from the system database is small in number for this specific study. There is a need to test the algorithm with a large number of dataset. Therefore, generating synthetic data which has the same attributes with the original CDR data but large number of instance. The use of generating synthetic CDR data helps us in the following ways:

1. Create large volumes of data, this is especially useful in performance testing, where large volumes of test data are needed.
2. Improve the efficiency of our model testing.

How to generate the data?

Data generation has been performed using oracle database with the function called dbms_random.value. We have used 4998 customer numbers as a calling number and 17,000 international (normal) and 1300 international (identified as a fraudulent number) as a called number with the time span of six months (Feb – July 2018). The call fee is calculated based on the current ethio telecom tariff format. The database script used for this data generation is presented in Appendix A.

Totally six million sample CDR records are generated. After generating the dataset, there is a need to sample the data when the generated data is a large in number. This will help to reduce the model training time. The next tasks are sampling from a larger dataset using random sampling. Out of 6,000,000 records 300000 are selected for the test.

4.2.5 *User profiling*

User profile is a collection of characteristics of a user. A process that refers to construction of a user profile via the extraction from a set of data [12]. For this specific study, two different profile types of each users are constructed and dataset are prepared accordingly. The first profile (Profile A) is a detailed daily behavior of a user which is constructed by separating the number of calls per day and their corresponding duration per day, call fee per, and the time of the day, i.e., working hours, out of working hours. Last, the second profile (Profile B) is an accumulated per day behavior. It consists of the number of calls and their corresponding duration separated. So, the two prepared dataset are convert into format suitable for the tool and the detail are discussed on the next section. Table 4.4 shows the two profiles with their corresponding attributes.

4.2.6 *Data formating*

Before dealing with the experimentation both dataset has to be formatted in a way that suitable for the tool to be used for modeling. In this study WEKA 3.8.2 is used which requires file formats like Comma Separated Values (CSV), Attribute-Relation File Format (ARFF) and so on. For this study ARFF format preferred to use. Since the preprocessing done using oracle database and the data base provided the data in such format.

Table 4.4: Attributes of prepared dataset

Profile A dataset fields or attributes	Calling Number No_of_Calls Normalize_Calling_data Normalize_Calling_time Ava_Duration Ava_Fee Ratio_Dur/fee
Profile B dataset fields or attributes	Calling Number No_of_Calls Ava_Duration Ava_Fee

4.3 EXPERIMENTATION

We had performed three sets of experiments to examine the classification performance of the proposed model. Two experiments were designed to evaluate the quality of the k-means and expectation maximization in detecting suspicious behaviors, an evaluation of several test with real CDR and synthetic data was performed. The first experiment examines the classification performance of the k-means algorithm on four datasets prepared from real CDR and synthetic test data. The next four experiments examine the classification performance of expectation maximization with the same datasets used for k-means. The final experiment examines the classification performance of the evaluation metrics in comparing machine segmentations to the current rule-based approach implemented on ethio telecom fraud management system.

The working principle of both k-means and expectation maximization comprehensively discussed in Chapter 3. The dataset used in this experiment are the classified as dataset one, dataset two, dataset three and dataset four. Further explanation are given below :

1. Dataset one (DS-One) has 48440 number of records and seven attributes.

2. Dataset two (DS-Two) has 48440 number of records and four attributes.
3. Dataset one (DS-Three) has 275061 number of records and seven attributes.
4. Dataset one (DS-Four) has 275061 number of records and four attributes.

Determining the number of clusters

There is no general theoretical solution to find the optimal number of clusters for any given dataset. The number of clusters (K) in this study is two in accordance with the grouping of objects in the dataset. The ultimate goal is to find classification of fraudulent calls over legitimate. The reason to set number of cluster as 2 is with the context of the test, the ultimate goal of the test is to classify the dataset into fraudulent and legitimate group. The second parameter is seed which is 10 for k-means and 100 for expectation maximization.

Following data clustering the record has been examined using the toll fraud properties discussed on Section 2.5.2 together with the knowledge gathered during discussion with domain expert on real scenario of ethio telecom enterprise customers call behaviors and the real fraudulent property identified through their experience on the detection of related fraud types. Therefor we set criteria like from the loneness of call duration greater than or equal to 25 minutes per call, from highness of call fee call costs 50 birr and above per minute, from frequency of call 5 number of call all calls should be distinct and from call day and time point of view calls made during unusual time such as weekend, out of office hour and holidays. The database script used for analysis is presented in Appendix A. The next subsequent subsection are discussed each experiment conducted.

4.3.1 *Experiment* 1*:using K-means Algorithm*

The first experiment has been performed using k-means algorithm. Four dataset were used during testing and, these datasets are presented in **??**. As mentioned in the Section 4.3 cluster size were set as two. Then the test was run on the laptop having 8GB RAM and 1.4 GHz processor. After running four test the clustered data are as follows:

1. Using the first dataset, from the total of 48440 instances 29658 (61%) number of instance grouped under cluster 0 and the rest 18782 (39%) instance under cluster 1. The algorithm took 4 number of iteration and 0.13 second until all data object are clustered.

2. From the second dataset of total 48440 instances 5807 (12%) number of instance grouped under cluster 0 and the rest 42633 (39%) instance under cluster 1. The algorithm took 6 number of iteration and 0.09 second until all data object are clustered.

3. The third dataset resulted from the total of 275061 instances 115114 (42%) number of instance grouped under cluster 0 and the rest 159947 (58%) instance under cluster 1. The algorithm took 2 number of iteration and 0.23 second until all data object are clustered.

4. The fourth dataset resulted from the total of 275061 instances 140667 (51%)number of instance grouped under cluster 0 and the rest 134394 (49%) instance under cluster 1. The algorithm took 6 number of iteration and 0.71 second until all data object are clustered.

The graphical test result of dataset one, two, three and four are shown in Figure 4.2. Additionally, the detail output information is summarized in Table A.2.

Following the clustering experiment further analysis were performed based on the defined threshold on Section 4.3. The total instances found in cluster 0 and cluster 1 of each dataset was again grouped as suspicious and legitimate call using the database script. Based on the defined threshold the instances were classified as shown in the following Table 4.5 and Table 4.6

4.3.2 *Experiment 2:using EM Algorithm*

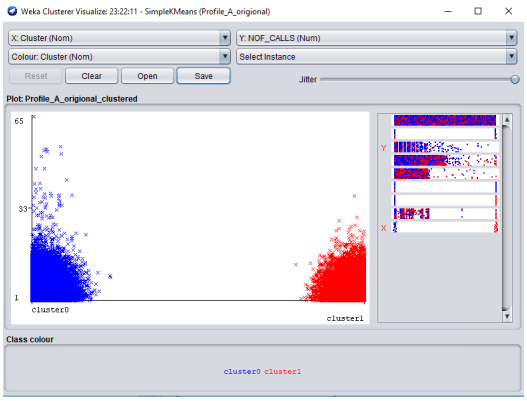
The second experiment has been conducted with a model using expectation maximization algorithm with the same dataset. It is possible to run it without specifying number of clusters, instead select number of cluster as -1. However, to check the EM algorithm performance the same procedure was done by specifying num-

Table 4.5: Ext.1 Analysis result of dataset one and two

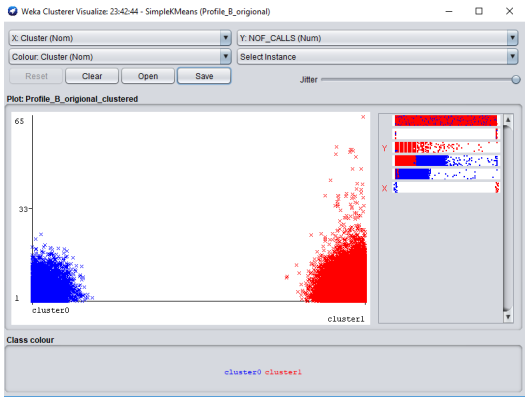
Properties	Datasets one		Datasets Two	
	Cluster 0	Cluster 1	Cluster 0	Cluster 1
Total dataset exhibits toll fraud call properties	2038 (6.87%)	2915 (15.52%)	66 (0.15%)	4887 (84.16%)
Total dataset exhibits legitimate call properties	27620 (93.13%)	15867 (84.48%)	42567 (99.85%)	920 (15.84%)

Table 4.6: Ext.1 Analysis result of dataset three and four

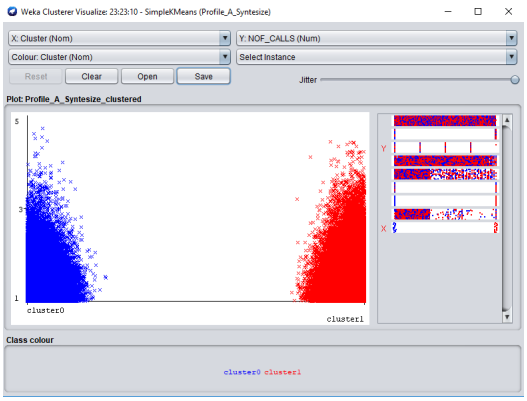
Properties	Datasets Three		Datasets Four	
	Cluster 0	Cluster 1	Cluster 0	Cluster 1
Total dataset exhibits toll fraud call properties	43219 (27.02%)	29921 (25.99%)	69054 (49.09%)	1527 (1.14%)
Total dataset exhibits legitimate call properties	116728 (71.98%)	85193 (74.01%)	71613 (50.91%)	132867 (98.86%)



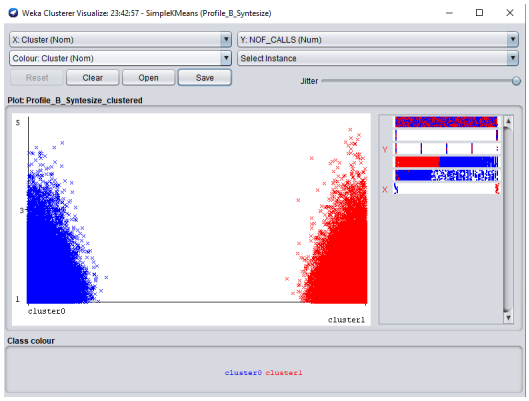
(a) Graphical output of experiment-1 k-means with profile A.



(b) Graphical output of experiment-1 k-means with profile B.



(c) Graphical output of experiment-2 k-means with profile A.



(d) Graphical output of experiment-2 k-means with profile B.

Figure 4.2: Graphical result of experiment one. **DRY!**

ber of cluster (2) and run to compare against K-means. And the output obtained after running the tests are as follows:

1. Using the first dataset, from the total of 48440 instances 35754 (74%) number of instance grouped under cluster 0 and the rest 12686 (26%) instance under cluster 1. The algorithm took 19 number of iteration and 1.98 second until all data object are clustered.

2. From the second dataset of total 48440 instances 39231 (81%) number of instance grouped under cluster 0 and the rest 9209 (19%) instance under cluster 1. The algorithm took 11 number of iteration and 1.24 second until all data object are clustered.

3. The third dataset resulted from the total of 275061 instances 40012 (15%) number of instance grouped under cluster 0 and the rest 235049 (85%) instance under cluster 1. The algorithm took 13 number of iteration and 7.91 second until all data object are clustered.

4. The fourth dataset resulted from the total of 275061 instances 84148 (31%) number of instance grouped under cluster 0 and the rest 190913 (69%) instance under cluster 1. The algorithm took 6 number of iteration and 7.51 second until all data object are clustered.

The graphical test result of dataset one, two, three and four are shown in Figure 4.3. Additionally, the detail output information is summarized in Table A.3. The analysis are performed with the same procedure follow during the previous experiment and, the total instances found on clusters (cluster 0 and cluster 1) of each dataset was again grouped as suspicious and legitimate call. The analysis results are shown in table Table 4.7 and Table 4.8

Table 4.7: Ext.2 Analysis result of dataset one and two

Properties	Datasets one		Datasets Two	
	Cluster 0	Cluster 1	Cluster 0	Cluster 1
Total dataset exhibits toll fraud call properties	67 (0.19%)	4886 (38.51%)	66 (0.17%)	4887 (53.07%)
Total dataset exhibits legitimate call properties	35687 (99.81%)	7800 (61.49%)	39165 (99.83%)	4322 (46.93%)

Table 4.8: Ext.2 Analysis result of dataset three and four

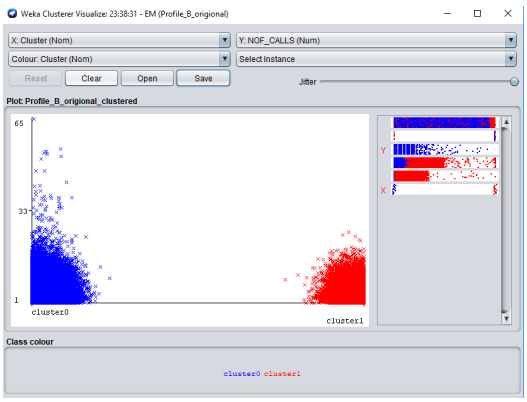
Properties	Datasets Three		Datasets Four	
	Cluster 0	Cluster 1	Cluster 0	Cluster 1
Total dataset exhibits toll fraud call properties	62076 (24.54%)	28948 (72.35%)	2908 (8.95%)	70232 (83.46%)
Total dataset exhibits legitimate call properties	190857 (75.46%)	11064 (27.65%)	29571 (91.05%)	13916 (16.54%)

4.3.3 *Experiment 3:using rule-based approach*

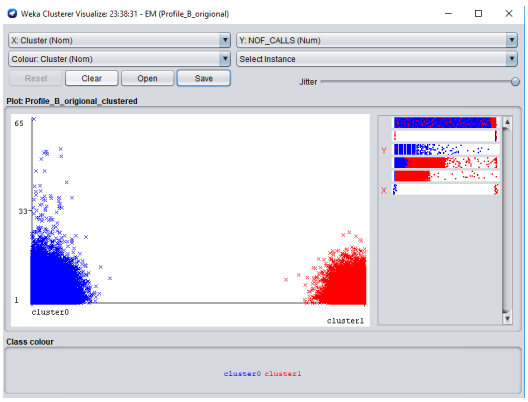
The intention of the study was to compare the result from the proposed model with the existing rule found on ethio telecom FMS. However due to some difficulties the test was not conducted. So that we forced to test the rule without their system. We try to write the rule with Java language but it takes time and decide to write the rule using database script. Accordingly, seven rule and four rule sets were written and tested. The result obtained from four rule sets are show in Table 4.9. The database scripts used to create a rule and rule set are presented on **??**, however the thresholds are excluded because of security reason.

Table 4.9: Rule sets result

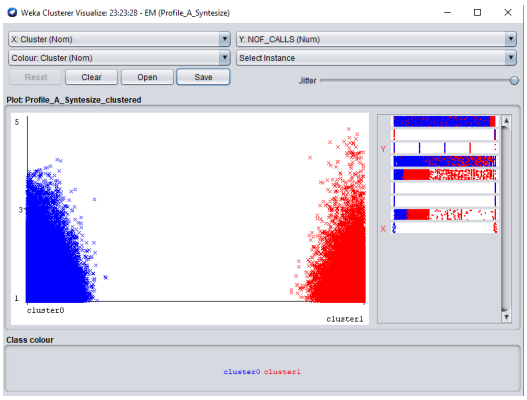
	suspicious call records	
Rule sets	Real CDR dataset	Synthesized test dataset
Ruleset-1	1387	13200
Ruleset-2	1810	29400
Ruleset-3	1095	11400
Ruleset-4	1323	16900



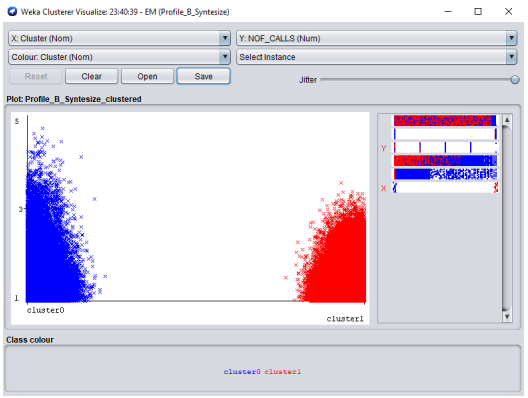
(a) Graphical output of experiment-1 EM with profile A.



(b) Graphical output of experiment-1 EM with profile B.



(c) Graphical output of experiment-2 EM with profile A.



(d) Graphical output of experiment-2 EM with profile B.

Figure 4.3: Graphical result of experiment two. **DRY!**

4.4 PERFORMANCE EVALUATION

The task of clustering seems to be intrinsically difficult to evaluate. Classification and association has an objective criterion of successes-prediction made on the test cases are either right or wrong whereas it is not so with clustering. A clustering evaluation demands an independent and reliable measure for the assessment and comparison of clustering experiments and results. In theory, the clustering researcher has acquired an intuition for the clustering evaluation, but in practice the mass of data on the one hand and the subtle details of data representation and clustering algorithms on the other hand make an intuitive judgment impossible. An intuitive, introspective evaluation can therefore only be plausible for small sets of objects, but large-scale experiments require an objective method.

There exist certain standard evaluation metrics which are commonly used to describe different aspects of the systems. The most common forms are accuracy, precision and recall. However, it is hard to select according to which of all the existing measures one should rely on to determine the performance or make a comparison of classification methods.

Definitions:

For this specific study some of the definitions of word that has been used while evaluation.

Suspicious call: calls that have toll fraud properties

Legitimate call: calls that have normal call properties

True Positive (TP) = the number of instance correctly classified as suspicious call

False Positive (FP) = the number of instance incorrectly classified as suspicious call

True Negative (TN) = the number of instance correctly classified as legitimate call

False Negative (FN) = the number of instance incorrectly classified as legitimate call

1. **Accuracy** describes the ratio of correct classification. It measures how many of samples are assigned to the correct classes and how many are assigned incorrectly. It is the most common form of performance metrics especially in machine learning. Although it tells about each class, it may fail in class-imbalanced situations as in anomaly detection. Formally, it is defined as follows.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{4.1}$$

2. **False Positive Rate** measures how many of the irrelevant class samples are labeled as relevant. In anomaly detection, false alarm rate refers to the ratio between the number of incorrectly detected normals and total number of normals.

$$FalsePositiveRate = \frac{FP}{FP + TN} \tag{4.2}$$

3. **False Negative Rate** measures how many of the relevant class samples are labeled as irrelevant. In anomaly detection, false negative rate refers to the ratio between the number of incorrectly detected normals and total number of normals.

$$FalseNegativeRate = \frac{FN}{FN + TP} \tag{4.3}$$

4.5 CONCLUSIONS

In this chapter we showed the methodology that attempts to offer a solution for a better classification for fraudulent over legitimate calls starting from data collection and preparation process, how raw data can be aggregated and make it suitable for the experiment. We also have conducted three experiment with k-means algorithm, expectation EM and rule based approach. In the next Chapter result obtained from the experiment will be further discussed.

5

RESULT AND DISCUSSION

5.1 INTRODUCTION

This Chapter will present and discuss the results found during the experiment made on Chapter 4. Two models with k-means and EM algorithm were tested using real and synthetic CDR data. These data further classified in four datasets. Dataset one and three are prepared based on profile A and dataset 2 and three were prepared based on profile B. Accordingly the results obtained from k-means algorithm with four datasets are explained in subsection 5.2.1. In subsection 5.2.2 discuss the results found using EM algorithm. The last subsection 5.2.3 present the results found from rule based approach. Finally, in Section 5.3 covered the discussions on results and the Chapter will conclude by answering the research questions.

5.2 RESULT

The results obtained from clustering model with k-means and EM algorithms are presented in Figure 5.1 and Figure 5.2 respectively. Each dataset was clustered into two group naming cluster 0 and cluster 1. The clustered data records further analyzed and their result will be discussed on each subsection.

5.2.1 *Result of experiment one (k-means algorithm)*

As mentioned in the above section the clustering model with k-means algorithm were tested using four datasets and the result are shown in below Figure 5.1.

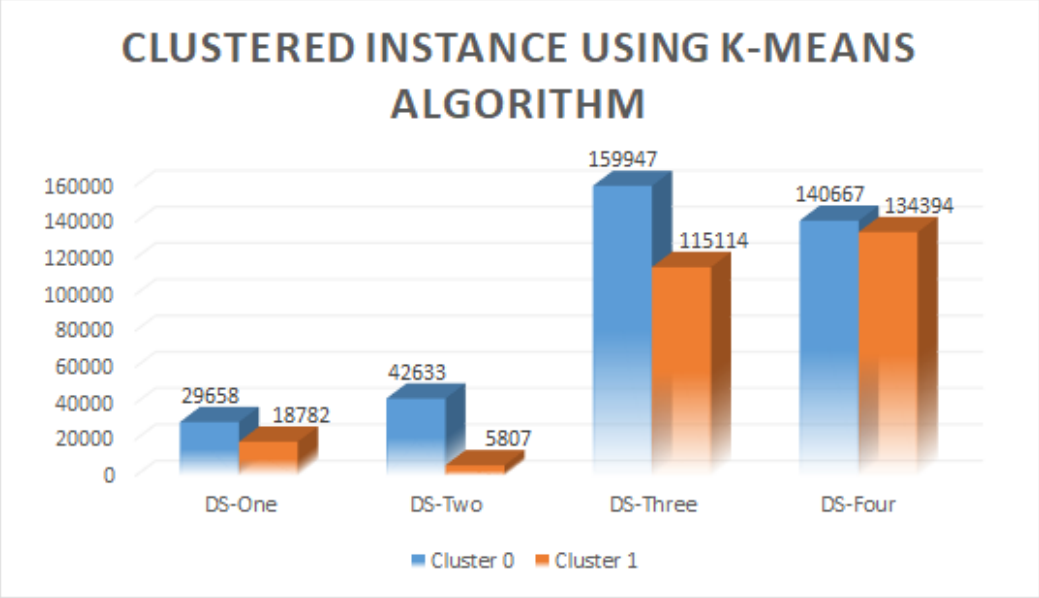


Figure 5.1: Clustered instance using k-means algorithm.

5.2.1.1 *Result obtained using real CDR*

Table 5.1: No.of legitimate and fraudulent records under each cluster(real CDR)

Cluster	**Profile A (DS-1)**		**Profile B (DS-2)**	
	Legitimate	**Fraudulent**	**Legitimate**	**Fraudulent**
0	27620	2038	42567	66
1	15867	2915	920	4887

The result using profile A dataset (DS-1) shown in Table 5.1 shows that in both cluster, large number of legitimate call records and few number of fraudulent call records were found. When we compare the two clusters more number of fraudulent call record where found on cluster 1 while more number of legitimate call were found on cluster 0. Following data records found on each cluster further

Table 5.2: Performance of k-means using dataset one and two

Algorithm	Dataset	FPR	FNR	Accuracy
	DS-1	36.49%	41.15%	63.04%
K-Means	***DS-2***	2.12%	1.33%	97.96%

analysis were done in order to examine the classification performance of the algorithm. Based on the classification performance metrics mentioned in Section 4.4, k-means algorithm showed an overall classification performance of 63.04% (see Table 5.2).

In similar fashion, the result using profile B dataset (DS-2) shows that large number of legitimate call record and small number of fraudulent call record are found in cluster 0 while small number of legitimate call record and large number of fraudulent call record are found in cluster 1. Here the algorithm showed an overall classification performance of 97.96% (see Table 5.2).

The above two results showed that k-means algorithm during the second test achieved significantly high number classification (legitimate from fraudulent call) performance compared to the first test. Number of data record in both datasets are equal. However, they are different by the number of attributes.

5.2.1.2 *Result obtained using synthetic CDR*

Following the same procedure made during the above test, in similar way results k-means algorithm with synthetic data are presented.

Table 5.3: No. legitimate and fraudulent records under each cluster(Synthetic CDR)

Cluster	Profile A (DS-3)		Profile B (DS-4)	
	Legitimate	Fraudulent	Legitimate	Fraudulent
Cluster 0	116728	43219	69054	71613
Cluster 1	85193	29921	132867	1527

Table 5.4: Performance of K-means using dataset three and four

Algorithm	Dataset	FPR	FNR	Accuracy
Expectation	DS-3	57.81%	40.91%	46.68%
Maximization	DS-4	65.80%	97.91%	74.34%

The result using profile A dataset (DS-3) shows that in both cluster, large number of legitimate call records and few number of fraudulent call records were found. Large number of fraudulent call record were found from cluster 0 whereas large number of legitimate call records were found in cluster 1. Based on the analysis made on the records found in both cluster classification performance of the algorithm was made and the result show that the overall performance is 46.68% (see Table 5.4).

The result with similar test using profile B dataset (DS-4) shows that, large number of fraudulent call records were found in cluster 0 whereas large number of legitimate call records were found in cluster 1. So that the analysis result showed during this test that, the EM algorithm has achieved the overall classification accuracy is 74.34% (see Table 5.4).

The above two results showed that k-means algorithm with profile B has achieved significantly high number classification performance compared to profile A.

5.2.2 *Result of experiment two (EM algorithm)*

Using the same datasets and same procedure test were performed on model with EM algorithms and, the result showed during clustering are presented in the below Figure 5.2. further result will be presented in the following two sub subsection.

5.2.2.1 *Result obtained using real CDR*

The result using profile A dataset (DS-1) shows that in both cluster, large number of legitimate call records and few number of fraudulent call records were found.

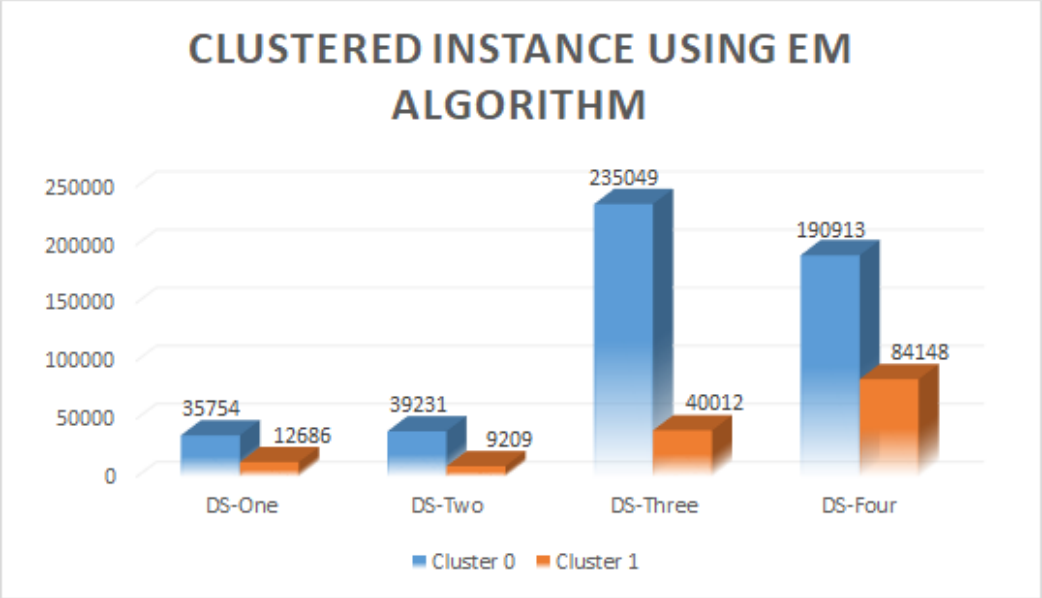


Figure 5.2: Clustered instance using Expectation maximization algorithm.

Table 5.5: Performance of EM algorithm with dataset one and two

Algorithm	Dataset	FPR	FNR	Accuracy
Expectation	DS-1	17.94%	1.35%	83.76%
Maximization	DS-2	9.94%	1.33%	90.94%

Large number of fraudulent call record were found in cluster 1 while large number of legitimate call were found in cluster 0. Following further analysis, the overall classification accuracy achieved is 83.76%. (see Table 5.5). Likewise, the result using profile B dataset (DS-2) shows that large number of legitimate call record and small number of fraudulent call record have found in cluster 0 while small number of legitimate call record and large number of fraudulent call record were found in cluster 1. Here the algorithm showed an overall classification performance of 90.94% (see Table 5.5). The above two results showed that EM algorithm on the second test has achieved better classification accuracy compared to the first test.

5.2.2.2 *Result obtained using synthetic CDR*

The result found during the experiment using dataset three and four are presented in below paragraph. The result using profile A dataset (DS-3) shows that

Table 5.6: performance of EM algorithm using dataset three and four

Algorithm	Dataset	FPR	FNR	Accuracy
Expectation	DS-3	5.48%	60.42%	79.91%
Maximization	***DS***-4	6.89%	3.98%	93.88%

in cluster 0, large number of legitimate call records and few number of fraudulent call records were found, whereas in cluster 1large number of fraudulent call records and small number of legitimate call records were found. Based on the analysis made on the records found in both cluster classification performance of the algorithm was made and the result show that the overall accuracy scored by the algorithm is 79.91% (see Table 5.6). The result shows with similar test using profile B dataset (DS-4), large number of fraudulent call records were found in cluster 1 whereas large number of legitimate call records were found in cluster 0. So the analysis result showed during this test implies the EM algorithm has achieved the overall classification accuracy is 93.88% (see Table 5.6). The above two results showed that EM algorithm during the first test has achieved much better classification accuracy as compared the second test.

5.2.3 *Result of experiment three (rule-based)*

1. ***Result***

   The result obtained from third experiment are shown in Table 5.7

Table 5.7: Result of experiment 3(Using rule set)

Dataset	TP	TN	FP	FN
Real CDR	3588	41460	2027	1365
Synthetic Data	58767	189788	12133	14373

The result from the third test using real CDR data has over all accuracy of 93.00%. The result of the Table 5.7 indicates that the detection of suspicious

Table 5.8: Performance of rule-based approach

Dataset	FPR	FNR	Accuracy
Real CDR	4.66%	27.56%	93.00%
Synthetic Data	6.01%	19.65%	90.36%

records is 72.44%. This means from the total of 4953 suspicious records 3588 of them are correctly detected. The remaining 1365 (27.564%) records were detected wrongly.Similarly,The result from the test using synthetic data has over all accuracy of 90.36%. The result of the Table 5.7 indicates that the detection of suspicious records is 80.35%. This means from the total of 73140 suspicious records 58767 of them are correctly detected. The remaining 14373 (19.65%) records were detected wrongly

5.3 DISCUSSION

The overall objective of the study is to propose model that detect fraudulent calls over legitimate with better accuracy as compare with the current implemented system (rule based approach). So that in order to meet this objective we need to assess clustering algorithm that classify instance with better accuracy. In addition to algorithm one of the major part of this study is to find CDR attributes that have play major role in the process of detection. Therefor with this in mind we will discuss on the result obtained during the experiment in line with our objective, are we meet or not.

To start from the algorithm clustering algorithms used throughout the study were k-means and EM algorithms. These two algorithms were selected based on the classification accuracy performed achieved on similar research [14] [15]. The overall accuracy achieved in three experiment are shown in Table 5.10, K-means algorithm has achieved the highest classification accuracy as compared with the classification accuracy achieved by expectation maximization and the rule-based approach. In similar ways while we compared

Table 5.9: Accuracy achieved per datasets

Attributes/fields of datasets	Acuuracy achieved	
Dataset 1 & 3 (Profile A)	k-means	EM
Calling Number No_of_Calls, Normalize_Calling_data, Normalize_Calling_time, Ava_Duration, Ava_Fee and Ratio_Dur/fee	63.04% 46.68%	83.76% 79.91%
Dataset 2 & 4 (Profile A)	k-means	EM
Calling Number No_of_Calls,Ava_Duration and Ava_Fee	97.96% 74.34%	90.94% 93.88%

the accuracy achieved with respect to the datasets, As shown in Table 5.9 the dataset with only 4 attributes were significantly higher than that of datasets with 7 attributes. This implies that from 7 attributes 3 of them (normalized call day, normalized call time and ratio of duration per fee) are not determinant regarding classification. In addition, the filed called calling number has no effect on increasing or decreasing the accuracy level. We just use to identifies from which calling number the fraudulent activities made and to take action accordingly.

Therefore, attributes like number of calls made, average call duration and average call fee are the three determinant CDR features or attributes. As discussed previously on chapter four this three attribute were derived from the original CDR attributes. Further conclusion are made during answering research questions.

Evaluations were made in line with the objectives of the research. The overall objectives of the study is proposing the model that has a better classification accuracy of fraudulent records that of legitimate.The highest score from each model are summarize and shown in the below Table 5.10. Both expectation

maximization model and rule based approach resulted in best accuracy level however, k-means algorithm is selected since its accuracy level is higher than the two. Following the above discussion, the major findings of the study are

Table 5.10: Summary Top Scored Models from k-means, EM Algorithms and rule-based

Model	***Acurracy***
K-means	*97.96%*
EM	*93.88%*
Rule-based	*93.00%*

summarize by ***Answering the research question***

Answering to the first research question, Research question one "What type of CDR data features or attributes can be preferable in order to predict toll fraud?", we found useful to represent data information as aggregated. Aggregated data resulted from transforming raw data into new attributes/variables that measure a certain goal concept, that is not yet explicit in the raw data. In process of aggregating data how (time span) and what (data feature) was the two main concern. From the real CDR data we found that call duration and call fee were the two main attributes to define the fraud properties.

Answering the second research question, research question two "What kinds of machine learning algorithm and models can be more effective in order to detect toll fraud?" As shown in Table 5.10 the experiment result showed that 97.96% accuracy is achieved. Therefore, the model from K-means algorithm is then selected as effective model.Bear in mind that data aggregation has also a vital role on increasing the detection capability. In general, from the result point of view unsupervised clustering approach are can give a better result in detecting fraud like toll call fraud.

Answering the third and last research question, research question three "How effective would the proposed model in terms of toll fraud detection rates over the existing methods?" this question somehow has similarity with the second question. The proposed model, model with k-means algorithm were

detecting more number of fraudulent call and wrongly detect a few number of legitimate calls as fraud call. In general, the proposed model has scored 4.96 % higher accuracy than that of rule based approach.

In general for a technique like unsupervised learning or clustering data preparation has the most important on classification of the input data as needed. Because the techniques hasn't take the classified data, it tries to learn pattern from the given data records. So that in this study having this in mind we do have an effort on preparing data try to fine a correct attributes which could play a key roll in finding the records that shows the toll fraud properties. Therefore as we first mentioned the toll fraud call has to main characteristics that identified the call such as call cost are very high and call duration is long than normal call behavior. Accordingly the attributes from CDR records which has priority on defining these characteristics is call duration and call fee. Apart from these two features or attributes the call date and time has also its own benefits on enhancing accuracy of classification.

6

CONCLUSION AND FUTURE WORK

This chapter concludes the thesis work. It starts with a conclusion and goes to future work.

6.1 CONCLUSION

Currently for telecom operator's fraud detection is an increasingly important and difficult task in today's technological environment. Since fraudsters do exist and always will, they always try to find a new way to make fraudulent activities. Similarly, telecom operator will work hard to detect these malicious activities before huge damage happens to the industry and their customers as well.

In this study an effort has been made to, propose a toll fraud detection model using machine learning (unsupervised clustering) algorithms. Performance of two clustering (k-means and EM) algorithm were evaluated in the detection process of fraudulent over legitimate call record. These algorithms were tested using real CDR collected from ethio telecom. In order to evaluate performance of the proposed algorithms with large number of datasets additional synthetic CDR was also used in the study. Two user profile (Profile A and Profile B) were constructed with different number of attributes. The first profile has seven attributes and the second has four attributes.

In the data preparation phase both real and synthetic CDR data were preprocessed and profile was constructed by aggregating the call made within a day. Following the data preparation three experiment were conducted. The first and second experiment were made using four datasets and a model

with k-means and EM algorithm respectively. The third experiment were made using rule-based approach. For the first two experiment to proceed the number of cluster (k) should be set primarily. Accordingly, we choose k=2, there is no general solution to find the optimal number of cluster for any given dataset. The reason behind choosing k=2 was, we need the dataset to be grouped into two cluster assuming legitimate cluster and fraudulent cluster.

After finalizing the data clustering process, the clustered data records were further analyzed in line with the fraudulent call properties. Records that shows fraudulent behavior were classified as fraudulent call. To do so database scripts were used to write the rules that differentiate the call records. So that in this process of fraud detection a model with k-means algorithm has achieved high classification accuracy (97.96%) with much less run time (0.09 seconds). However, this algorithm scores the worst accuracy result in the rest two tests, this means the algorithm is ineffective with the dataset that has higher number of outliers as compared to EM algorithm and rule-based approach. The second highest (93.88%) accuracy was achieved in the model with EM algorithm comparatively this algorithm scored better classification in the rest three tests. When we compare the results achieved with these two algorithm with that of rule based approach (93.00%), the result using k-means and EM algorithms showed higher in 4.96% and 0.88% respectively than that of the accuracy achieved using rule-based approach.

Another findings of the study were, both algorithms have scored higher classification performance during the test made with dataset constructed based on profile B. In other word the test made on dataset with 4 attributes has score better classification accuracy than that of the datasets with 7 attributes. So it is proved that the three additional attributes (call started day, call start time and fee/duration ratio) were not determinate attribute while classification.

In general, a model with k-means algorithm is less complex, simple to use and fast in terms of data clustering as compared to a model with EM algo-

rithms but needs to specify number of cluster prior while EM not need. Both models are less complex and fast as compared to the existing rule-based method. Finally, in this study few features were selected and with the addition of more features, the results can be improved as more fraudulent calls could be detected. Even though k-means has been failed to achieve better accuracy with the rest of three test the result found from the study shows that k-means algorithm seems to be a desirable choice, with high clustering quality and relatively low time complexity and it is possible to detect toll fraud using the presented approach.

6.2 FUTURE WORK

Due to limited number of real CDR dataset used in this thesis, there is a need to use synthetic CDR data. Future work on increasing the detection accuracy and achieve low false positive rate would consist of selecting a better clustering algorithm and use real CDR data that contain large number of records. In addition, use unseen CDR attributes during this research, construct different user-profile to compare a user's present behavior to the same users past behavior, thus making outliers more noticeable.

REFERENCE

[1] J. Shawe-Taylor, K. Howker, and P. Burge, "Detection of fraud in mobile telecommunications," *Information Security Technical Report*, vol. 4, no. 1, pp. 16–28, 1999.

[2] R. Becker, C. Volinsky, and A. Wilks, "Fraud detection in telecommunications: History and lessons learned," *Technometrics*, vol. 52, no. 1, pp. 20–33, 2010.

[3] C. F. C. Association *et al.*, "Global fraud loss survey," *Press Release, New Jersey, NJ (CFCA)*, vol. 10, p. 2013, 2017.

[4] Y. Kou, C.-T. Lu, S. Sirwongwattana, and Y.-P. Huang, "Survey of fraud detection techniques," in *Networking, sensing and control, 2004 IEEE international conference*, IEEE, vol. 2, 2004, pp. 749–754.

[5] M. I. Akhter and M. G. Ahamad, "Detecting telecommunication fraud using neural networks through data mining," *International Journal of Scientific and Engineering Research*, vol. 3, no. 3, pp. 601–6, 2012.

[6] I. Ighneiwa and H. Mohamed, "Bypass fraud detection: Artificial intelligence approach," *arXiv preprint arXiv:1711.04627*, 2017.

[7] P. Gosset and M. Hyland, "Classification, detection and prosecution of fraud in mobile networks," *Proceedings of ACTS mobile summit, Sorrento, Italy*, 1999.

[8] Apanews. (Mar. 6, 2017). Ethiopia loses over $52m to telecom fraud-official. [Online; 2017-03-06], [Online]. Available: `https://mobile.apanews.net/en/news/ethiopia-loses-over-52m-to-telecom-fraud-official`.

[9] K.-I. Kim, T. Kim, N.-W. Cho, and M. Kim, "Toll fraud detection of voip service networks in ubiquitous computing environments," *Inter-*

national Journal of Distributed Sensor Networks, vol. 11, no. 9, p. 276 408, 2015.

[10] M. Sahin, A. Francillon, P. Gupta, and M. Ahamad, "Sok: Fraud in telephony networks," in *Proceedings of the 2nd IEEE European Symposium on Security and Privacy (EuroS&P17), EuroS&P*, vol. 17, 2017.

[11] N. Koiser, "Toll fraud detection in voip networks using artificial neural networks," Master's thesis, University of Nairobi, April 2016.

[12] A. Wiens, T. Wiens, and M. Massoth, "A new unsupervised user profiling approach for detecting toll fraud in voip networks," in *The Tenth Advanced International Conference on Telecommunications (AICT 2014) IARIA*, 2014, pp. 63–69.

[13] R. K. Gopal and S. K. Meher, "A rule-based approach for anomaly detection in subscriber usage pattern," in *Proceedings of World Academy of Science, Engineering and Technology*, 2007, pp. 396–399.

[14] S. Kübler, M. Massoth, A. Wiens, and T. Wiens, "Toll fraud detection in voice over ip networks using communication behavior patterns on unlabeled data," *ICN 2015*, p. 203, 2015.

[15] S. I. M. Reyes, G. R. Salgado, and J. P. Ortega, "Defining adaptive whitelists by using clustering techniques, a security application to prevent toll fraud in voip networks," in *Proceedings of the International Conference on Information and Knowledge Engineering (IKE)*, The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2016, p. 100.

[16] J. W. Phelps, *Toll fraud detection system*, US Patent 5,602,906, 1997.

[17] P. A. Estévez, C. M. Held, and C. A. Perez, "Subscription fraud prevention in telecommunications using fuzzy rules and neural networks," *Expert Systems with Applications*, vol. 31, no. 2, pp. 337–344, 2006.

[18] A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system: A survey," *Journal of Network and Computer Applications*, vol. 68, pp. 90–113, 2016.

[19] A. Mason, *Premium rate services: International markets and regulation: Final report for phonepayplus*, 2011.

[20] Frontiernetworks. (Feb. 27, 2012). How to prevent toll fraud. [Online; 2018-04-25], [Online]. Available: http://www.frontiernetworks.ca/how-to-prevent-toll-fraud..

[21] J. Yu, "Prevention of toll frauds against ip-pbx," in *Proceedings of the International Conference on Security and Management (SAM)*, The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2015, p. 259.

[22] Tollshield. (2016). Pbx hacking: How it works. [Online; 2016-090-15], [Online]. Available: https://tollshield.com/news/pbx-hacking-how-it-works.

[23] S. Hofbauer, K. Beckers, G. Quirchmayr, and C. Sorge, "A lightweight privacy preserving approach for analyzing communication records to prevent voip attacks using toll fraud as an example," in *Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th International Conference on*, IEEE, 2012, pp. 992–997.

[24] Securelogic. (Oct. 26, 2012). Toll-fraud-use-case. [Online; 2018-01-23], [Online]. Available: http://download.securelogix.com/library/Toll_fraud_use_case.pdf.

[25] TransNexus. (Oct. 18, 2012). Introduction-to-voip-fraud. [Online; 2018-01-23], [Online]. Available: https://transnexus.com/resources/telecom-industry-topics/fraud/introduction-to-voip-fraud.

[26] L. Maruster, *A machine learning approach to understand business processes*. Technische Universiteit Eindhoven, 2003.

[27] M. Mohammed, M. B. Khan, and E. B. M. Bashier, *Machine learning: algorithms and applications*. CRC Press, 2016.

[28] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[29] R Frank, N Davey, and S Hunt, "Applications of neural networks to telecommunications systems," in *In: Procs of the European Congress on Intelligent Techniques and Soft Computing (EUFIT'99)*, ELITE Foundation, 1999.

[30] C. Phua, V. Lee, K. Smith, and R. Gayler, "A comprehensive survey of data mining-based fraud detection research," *arXiv preprint arXiv:1009.6119*, 2010.

[31] P. Sagar, "Analysis of prediction techniques based on classification and regression," *International Journal of Computer Applications*, vol. 163, no. 7, 2017.

[32] I. Syarif, A. Prugel-Bennett, and G. Wills, "Unsupervised clustering approach for network anomaly detection," in *International Conference on Networked Digital Technologies*, Springer, 2012, pp. 135–145.

[33] S. Dua and X. Du, *Data mining and machine learning in cybersecurity*. CRC press, 2016.

[34] R. Domingues, *Machine learning for unsupervised fraud detection*, 2015.

[35] D. S. Matusevich *et al.*, "A fast clustering algorithm merging the expectation maximization algorithm and markov chain monte carlo," PhD thesis, 2015.

[36] R. Alves, P. Ferreira, O. Belo, J. Lopes, J. Ribeiro, L. Cortesão, and F. Martins, "Discovering telecom fraud situations through mining anomalous behavior patterns," in *Proceedings of the DMBA Workshop, on the 12th ACM SIGKDD*, 2006.

[37] M. K. Rafsanjani, Z. A. Varzaneh, and N. E. Chukanlo, "A survey of hierarchical clustering algorithms," *The Journal of Mathematics and Computer Science*, vol. 5, no. 3, pp. 229–240, 2012.

[38] L. Bijuraj, "Clustering and its applications," in *Proceedings of National Conference on New Horizons in IT-NCNHIT*, vol. 1, 2013, pp. 169–172.

[39] D. Sisodia, L. Singh, S. Sisodia, and K. Saxena, "Clustering techniques: A brief survey of different clustering algorithms," *International Journal of Latest Trends in Engineering and Technology (IJLTET)*, vol. 1, no. 3, pp. 82–87, 2012.

[40] W. Kim, "Parallel clustering algorithms: Survey," *Parallel Algorithms, Spring*, 2009.

[41] S. Kaushik, "An introduction to clustering and different methods of clustering," *Analytics Vihya Learn everything about analytics*, 2016.

[42] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Oakland, CA, USA, vol. 1, 1967, pp. 281–297.

[43] T. Prajwala and V. Sangeeta, "Comparative analysis of em clustering algorithm and density based clustering algorithm using weka tool.," *International Journal of Engineering Research and Development*, vol. 9, no. 8, pp. 19–24, 2014.

[44] C. S. Hilas, P. A. Mastorocostas, and I. T. Rekanos, "Clustering of telecommunications user profiles for fraud detection and security enhancement in large corporate networks: A case study," *Applied Mathematics & Information Sciences*, vol. 9, no. 4, p. 1709, 2015.

[45] M. Kamel, "Data preparation for data mining," in *Encyclopedia of Data Warehousing and Mining, Second Edition*, IGI Global, 2009, pp. 538–543.

[46] A. K. Muli, "A model to detect and protect toll fraud in voip pbx infrastructure," Master's thesis, KCA University, 2017.

[47] D. D. Gutierrez, *Machine learning and data science: an introduction to statistical learning methods with R*. Technics Publications, 2015.

[48] L. Cortesão, F. Martins, A. Rosa, and P. Carvalho, "Fraud management systems in telecommunications: A practical approach," in *Proceeding of ICT*, 2005.

A

APPENDIX

Detailed CDR table

Table A.1: detail records.

NO	Field Meaning	Field Name	Remark
1	CDR ID	EVENT_INST_ID	CDR ID
2	Event	RE_ID	1: voice 2: SMS 5: GPRS (Data)
3	Billing Number	BILLING_NBR	
4	CDR type	CDR_TYPE	1: MO 2: MT 3: MF
5	Calling Number	CALLING_NBR	
6	Called Number	CALLED_NBR	
7	Calling IMEI	CALLING_IMEI	
8	Calling IMSI	CALLING_IMSI	
9	The Third Party Number	THIRD_NBR	Used for call forward
10	Call start time	START_TIME	
11	Call end time	END_TIME	
12	Call duration	DURATION	
13	Call fee	CALL_FEE	

Continued on next page

Table A.1 – continued from previous page

NO	Field Meaning	Field Name	Remark
14	Called country	CALLED_COUNTRY	
15	Calling carrier	CALLING_CARRIER	
16	Called carrier	CALLED_CARRIER	
17	Calling district	CELL_A	
18	Called district	CELL_B	
19	Status date	STATE_DATE	Billing date
20	Calling SUB id	CALLING_SUB_ID	
21	Billing cycle ID	BILLING_CYCLE_ID	
22	Charge 1	CHARGE1	
23	Charge 2	CHARGE2	
24	Rate ID1	PRICE_ID1	
25	Account item ID1	ACCT_ITEM_ID1	
26	Upload traffic	TRAFFIC_UP	
27	Download traffic	TRAFFIC_DOWN	
28	Billing offering id	BILLING_OFFERING_ID	
29	Error CDR type	ERROR_CDR_TYPE	Normal (N)/Error(E)
30	Call Forward Indicator	CALL_FORWARD_INDICATOR	
31	Hot Line Indicator	HOT_LINE_INDICATOR	9: voice mail
32	Calling Trunk ID	CALLING_TRUNK_ID	
33	Called Trunk ID	CALLED_TRUNK_ID	

Experiment Result

Database Scripts for Data preparation and preprocess

Table A.2: Summary output information of Experiment 1

Dataset		DS-One	DS-Two	DS-Three	DS-Four
No instance		48440	48440	275061	275061
Number of Attributes		7	4	7	4
Test mode		Evaluate on training data			
Number of iterations		4	6	2	6
Sum of squared errors		6679.514593	201.0986644	79538.13069	8114.620977
Time taken to build model (second)		0.13	0.09	0.23	0.71
Clustered Instances	Cluster 0	29658 (61%)	5807 (12%)	115114 (42%)	140667 (51%)
	Cluster 1	18782 (39%)	42633 (39%)	159947 (58%)	134394 (49%)

Table A.3: Summary output information of Experiment 2

Dataset		DS-One	DS-Two	DS-Three	DS-Four
No instance		48440	48440	275061	275061
Number of Attributes		7	4	7	4
Test mode		Evaluate on training data			
Number of iterations		19	11	13	6
Log likelihood		-34.42517	-32.18496	-36.66553	-34.14406
Time taken to build model (second)		1.98	1.24	7.91	7.51
Clustered Instances	Cluster 0	35754 (74%)	39231 (81%)	40012 (15%)	84148 (31%)
	Cluster 1	12686 (26%)	9209 (19%)	235049 (85%)	190913 (69%)

a) **Database script to create table for storing PBX users numbers.**

CREATE TABLE KASS.PBX_cdr_sample3 (CALLING_NBR VARCHAR2(64 BYTE), CALLED_NBR VARCHAR2(64 BYTE), START_date VARCHAR2(32 BYTE), START_TIME VARCHAR2(32 BYTE), End_date VARCHAR2(32 BYTE), End_time VARCHAR2(32 BYTE), DURATION NUMBER(20), CALL_FEE NUMBER(16,2))drop table KASS.PBX_cdr_sample2

b) **Database script to generate synthetic sample data.**

create table ET_1837.pbx_NON_fraud as ((select CALLING_NBR,CALLED_NBR, sysdate - round (dbms_random.value(-6.25,174),2) start_date , round (dbms_random.value
(60,1020),0) duration from (select b.CALLING_NBR CALLING_NBR , a.CALLED_NBR CALLED_NBR from ET_1837. PBX_NON_risky_no a , ET_1837.PBX_ACC b order by dbms_random.value)where rownum <200001));

c) **Database script to summarize CDR data per no of call per day**

select CALLING_NBR,NORM_DAY, NORM_TIM, sum(DURATION) duration ,sum(call_fee) call_fee, count(CALLED_NBR) NOF_calls, round(sum(DURATI
dur_call_ratio, round(sum(call_fee)/count(CALLED_NBR),2) fee_call_ratio , round(sum(call_fee)/sum(DURATION),2) fee_dur_ratio from PBX_SAMPLE_SEP group by CALLING_NBR,NORM_DAY,NORM_TIM,START_DAY order by 1,2,3

d) Database script to summarize CDR data per no of call per day

select CALLING_NBR,NORM_DAY, NORM_TIM, sum(DURATION) duration ,sum(call_fee) call_fee, count(CALLED_NBR) NOF_calls, round(sum(DURATI
dur_call_ratio, round(sum(call_fee)/count(CALLED_NBR),2) fee_call_ratio , round(sum(call_fee)/sum(DURATION),2) fee_dur_ratio from PBX_SAMPLE_SEP group by CALLING_NBR,NORM_DAY,NORM_TIM,START_DAY order

by 1,2,3
scr:analysis

e) select distinct upper(to_char (to_date (START_DAY ,'YYYY-MM-DD') ,'DAY')) from PBX_SAMPLE1

f) select * from PBX_SAMPLE1 where DURATION >=1800

g) select * from PBX_SAMPLE1 where trim(to_char (to_date (START_DAY ,'YYYY-MM-DD') ,'DAY')) in ('SUNDAY','SATURDAY')

h) select * from PBX_SAMPLE1 where trim(to_char (to_date (START_DAY ,'YYYY-MM-DD') ,'DAY'))='SATURDAY'

i) select count(1) from kmeans_dataset_4 where CALLED_NBR in (select * from FRAUD_B_NUM) and CLASTER ='clustero'

j) select count(1) from kmeans_dataset_4 where trim(to_char (to_date (START_DAte ,'YYYY-MM-DD') ,'DAY')) in ('SUNDAY','SATURDAY') and CLASTER ='clustero'

k) select count(1) from kmeans_dataset_4 where CALLED_NBR in (select * from FRAUD_B_NUM) and trim(to_char (to_date (START_DAte ,'YYYY-MM-DD') ,'DAY')) in ('SUNDAY','SATURDAY') and CLASTER ='clustero'

l) select * from kmeans_dataset_2 –where norm_day= 2

m) select * from kass.pbx_sample1_sep –where norm_day= 2 – trim(to_char (to_date (START_DAY ,'YYYY-MM-DD') ,'DAY')) in ('SUNDAY','SATURDAY')

n) select count(1) from PBX_SAMPLE1 where trim(to_char (to_date (START_DAY ,'YYYY-MM-DD') ,'DAY')) in ('SUNDAY','SATURDAY')

o) select * from PBX_SAMPLE2_sep –where trim(to_char (to_date (START_DAY ,'YYYY-MM-DD') ,'DAY')) in ('SUNDAY','SATURDAY')

p) create table PBX_SAMPLE2_sep as select CALLING_NBR, CALLED_NBR, START_DAY, STAR_HOUR, END_DAY, END_HOUR, DURATION, FEE,0 norm_day, 0 norm_tim from PBX_SAMPLE2

q) update kass.pbx_sample1_sep set norm_day= 2 where trim(to_char (to_date (START_DAY ,'YYYY-MM-DD') ,'DAY')) in ('SUNDAY','SATURDAY')

r) update PBX_SAMPLE2_sep set norm_tim= 2 where norm_tim =0

s) update PBX_SAMPLE2_sep set norm_tim= 2 where STAR_HOUR like '19

Database scripts for creating a rules

Rule 1

SELECT * FROM PBX_ORIGINAL_SUM_DAY WHERE DURATION > DT AND DURATION_RATIO >RT AND DURATION >HIS_DURATION * RT

Rule 2

SELECT * FROM PBX_ORIGINAL_SUM_DAY WHERE NOF_CALLS > NT AND CALL_RATIO >RT AND NOF_CALLS >HIS_NOF_CALLS * RT

Rule 3

SELECT * FROM PBX_ORIGINAL_SUM_DAY WHERE HDAY_DURATION > 300 AND HDAY_DURATION_RATIO >RT AND HDAY_DURATION > HDAY_HIS_DURATION * RT

Rule 4

SELECT * FROM PBX_ORIGINAL_SUM_DAY WHERE HDAY_NOF_CALLS > NT AND HDAY_CALL_RATIO > RT AND HDAY_NOF_CALLS > HDAY_HIS_NOF_CALLS * RT

Rule 5

SELECT * FROM PBX_ORIGINAL_SUM_DAY WHERE PRS_NOF_CALLS > NT

Rule 6

SELECT * FROM PBX_ORIGINAL_SUM_WEEK WHERE DURATION > DT AND DURATION_RATIO > RT AND DURATION >HIS_DURATION * RT

Rule 7

SELECT * FROM PBX_ORIGINAL_SUM_WEEK WHERE NOF_CALLS > NT AND CALL_RATIO > RT AND NOF_CALLS >HIS_NOF_CALLS * RT

Database script to create rule sets

Rule set 1

(SELECT CALLING_NBR FROM PBX_ORIGINAL_SUM_DAY WHERE DURATION > DT AND DURATION_RATIO >RT AND DURATION > HIS_DURATION * RT)
INTERSECT
(SELECT CALLING_NBR FROM PBX_ORIGINAL_SUM_DAY WHERE NOF_CALLS > NT AND CALL_RATIO >RT AND NOF_CALLS > HIS_NOF_CALLS * RT)

Rule set 2

SELECT * FROM PBX_ORIGINAL_SUM_DAY WHERE HDAY_DURATION > NT AND HDAY_DURATION_RATIO >RT AND HDAY_DURATION > HDAY_HIS_DURATION * RT
INTERSECT
SELECT * FROM PBX_ORIGINAL_SUM_DAY WHERE HDAY_NOF_CALLS > NT AND HDAY_CALL_RATIO >RT AND HDAY_NOF_CALLS > HDAY_HIS_NOF_CALLS * RT

Rule set 3

SELECT * FROM PBX_ORIGINAL_SUM_DAY WHERE PRS_NOF_CALLS > NT

Rule set 4

SELECT * FROM PBX_ORIGINAL_SUM_WEEK WHERE DURATION > DT AND DURATION_RATIO >RT AND DURATION > HIS_DURATION * RT
INTERSECT
SELECT * FROM PBX_ORIGINAL_SUM_WEEK WHERE NOF_CALLS > NT AND CALL_RATIO >RT AND NOF_CALLS > HIS_NOF_CALLS * RT

Sample dataset used for weka

```
@relation 'Profile_A_origional'

@attribute CALLING_NBR numeric
@attribute NOF_CALLS numeric
@attribute Avr_Duration numeric
@attribute Avr_Fee numeric
@attribute NORM_DAY numeric
@attribute NORM_TIM numeric
@attribute FEE_DUR_RATIO numeric

@data
251[illegible]047,1,1,7.5,1,1,0.13
251[illegible]047,1,3,22.5,1,1,0.13
251[illegible]047,1,2,15,1,1,0.13
251[illegible]047,1,1,7.5,1,1,0.13
251[illegible]047,1,1,7.5,1,1,0.13
251[illegible]047,2,1,7.5,1,1,0.13
251[illegible]047,1,5,44.75,1,2,0.15
251[illegible]047,1,1,8.95,1,2,0.15
251[illegible]047,1,6,53.7,1,2,0.15
251[illegible]047,1,9,67.5,2,1,0.13
251[illegible]047,1,1,7.5,2,2,0.13
251[illegible]098,1,2,17.9,1,1,0.15
251[illegible]098,1,9,67.5,1,1,0.13
251[illegible]098,1,2,15,1,1,0.13
251[illegible]098,1,1,7.5,1,1,0.13
251[illegible]098,1,13,97.5,1,2,0.13
251[illegible]118,1,1,8.95,1,1,0.15
251[illegible]118,1,1,8.95,1,1,0.15
251[illegible]118,2,2.5,18.75,1,1,0.13
251[illegible]118,1,1,8.95,1,1,0.15
251[illegible]118,1,38,8668.56,1,1,3.8
251[illegible]118,1,3,22.5,1,2,0.13
251[illegible]118,1,3,25.88,1,2,0.14
251[illegible]131,1,3,26.85,1,1,0.15
251[illegible]131,3,1.666666667,12.5,1,1,0.13
251[illegible]131,1,13,116.35,1,1,0.15
251[illegible]131,1,4,30,1,1,0.13
```

Figure A.1: Sample Profile A dataset.

```
1  @relation 'Profile_B_Syntesize'
2
3  @attribute CALLING_NBR numeric
4  @attribute NOF_CALLS numeric
5  @attribute Avr_Duration numeric
6  @attribute Avr_Fee numeric
7
8
9  @data
10 25        047,1,60,17184
11 25        047,1,4,35.8
12 25        047,1,20,179
13 25        047,1,55,3911.05
14 25        047,1,33,8981.94
15 25        047,1,22,165
16 25        047,1,4,30
17 25        047,1,23,172.5
18 25        047,2,43,3030.79
19 25        047,1,5,37.5
20 25        047,1,37,277.5
21 25        047,2,16,141.75
22 25        047,1,9,67.5
23 25        047,1,40,358
24 25        047,1,22,165
25 25        047,1,38,285
26 25        047,1,11,82.5
27 25        047,1,39,292.5
28 25        047,1,37,10596.8
29 25        047,2,26,195
30 25        047,2,44,2807.64
31 25        047,1,18,161.1
32 25        047,1,33,7527.96
33 25        047,1,19,142.5
34 25        047,1,29,259.55
35 25        047,2,34.5,283.4
36 25        047,1,46,3271.06
37 25        047,2,39,5327.2
38 25        047,1,38,340.1
```

Figure A.2: Sample Profile B dataset.