Analysis and Detection Mechanisms of SIM Box Fraud in The Case of Ethio Telecom

Thesis Submitted to The Department of Electrical and Computer Engineering at Addis Ababa University Institute of Technology In Partial Fulfillment of Master's Degree of Computer Engineering

By

Frehiwot Mola

Advisor

Yalemzewd Negash (PhD)

December 12, 2017

Addis Ababa, Ethiopia

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES INSTITUTE OF TECHNOLOGY

Analysis and Detection Mechanisms of SIM box Fraud in The Case of Ethio Telecom

By

Frehiwot Mola

Approved by Board of Examiners

Department head of electrical and computer Engineering	Signature
Advisor	Signature
Internal Examiner	Signature
External Examiner	Signature

Abstract

Telecommunication fraud can be defined as theft of services (fixed telephone, mobile, data and etc.) or measured abuse of voice or data networks. Fraud is one of the most severe threats to revenue and quality of service in telecommunication networks. The advent of new technologies has provided fraudsters with new techniques to commit fraud. Subscriber identity module box (SIMbox) fraud is one of such fraud that is used in international calls and it has emerged with the use of VOIP technologies.

In this thesis, the call detail records (CDR's) from ethio telecom were organized in order to develop models of normal and fraudulent number behavior via data mining techniques. And four classification algorithms namely decision trees, rule based induction, neural network and hybrid algorithms are used. First we have done data analysis on the data set and for classification we use nine selected features of data extracted from Customer Database Record.

The experimentation result will enable to understand the problem of SIM box fraud in the case of ethio telecom and clarifying the behavior of fraudulent and legitimate calls. Finally, we got a good result from PART rule based and hybrid (J48 and PART) algorithms and performed the best among the four algorithms. PART rule based induction classification algorithm had a better performance with an accuracy rate of 99.4906% with true positive and 0.5094 % false positive ratio and followed by hybrid of J48 and PART algorithm with accuracy rate 99.4795% with true positive and 0.5205% false positive ratios.

Key words: SIM Box Fraud, Telecom Fraud, Decision Tree, Multilayer perceptron

Acknowledgements

First I would like to thank GOD for all countless gifts and for giving the strength to finish this study.

It is a great pleasure to express my sincere gratitude to my advisor Dr. Yalemzewed Negash, for his valuable guidance and advice. He has been very-supportive and patient throughout the Progress of my thesis.

I would like to express my gratitude towards my family especially for my father, my brother and my husband for their encouragement and support which helped me in completion of this thesis.

Finally, I would like to thank my friends, classmates and ethio telecom domain experts especially Ato Mehari G/egizeabher, Wr. Selamawit Yitbarek and Ato Solomon Girma for their kind help and support during the study.

Dedication

This thesis is dedicated to the memory of my beloved mother Sirgut Gebre, for making me be who I am. She is a smart woman whom I still miss every day.

Table of Contents

List of Figures

List of Tables

List of Acronyms

SIM	Subscriber Identity Module
GSM	Global Stations for Mobile communications
CDR	Call Detail Record
HLR	Home Location Register
CBS	Convergent Billing System
SMS	Short Message Service
GPRS	General Packet Radio Service
IMEI	International Mobile Equipment Identity number
IMSI	International Mobile Subscriber Identity number
CRM	Customer Relation Management
ETC	Ethiopian Telecommunication Corporation
LTE	Long-Term Evolution
VOIP	Voice over Internet protocol
UNMS	Unified Network Management System
WEKA	Waikato Environment used for knowledge Analysis
DT	Decision Tree
NN	Neural Network
SVM	Support Vector Machine
MLP	Multi-Layer Perceptron
ANN	Artificial Neural Network
FMS	Fraud Management System

DM	Data Mining
ARFF	Attribute Relation File Format
CSV	Comma Separated Values
CART	Classification and Regression Tree
MSC	Mobile Switching Center
BTS	Base Transceiver Station
KDP	Knowledge Discovery Process
KDD	Knowledge Discover in Databases

Chapter One

1. Introduction

According to the Cambridge Advanced Learner's Dictionary, fraud is an intentional deception or cheating intended to gain an advantage while fraud in communication can be defined as the theft of service and misuse of voice as well as data networks of telecom providers. SIM box fraud is classified as one of the dominant type of fraud and this activity has been increasing dramatically each year due to the new modern technologies and the global super high ways of communication, resulting a decrement of revenue and quality of service in telecommunication providers, especially in Africa and Asia.

1.1 Background

Ethio telecom is the oldest public telecommunication operator (PTO) in Africa. The first long-distance telephone line was established in 1894 between Addis Abeba and Harar.

Ethio telecom is government owned and the sole telecom operator in Ethiopia. The name ethio telecom is coined in 2010 after France telecom took the management of Ethiopian Telecommunication Corporation due to government transformation plan. The introduction of telecommunications in Ethiopia dated back to 1894. The operator has passed through different names (brand names) and logos, by different governments that came in power, since the beginning. Fixed telephone (both wired and wireless), Internet (wireless and broadband), mobile (pre-paid and post-paid) including 3G (voice and data), 4G LTE (voice and data) and other value-added services are among the major telecom services provided by the company. [16].

The company placed mobile service division in its structure beginning from 1996. Mobile service in Ethiopia has existed since 1999 and at that time the network coverage was limited to Addis Ababa with a network capacity of not more than 60, 000 subscribers. After the introduction of mobile service in Addis Ababa, in April 1999, network expansion was a necessity, not only because of the demand from the subscribers but also due to government policy.[13].

According to ethio telecom corporate communication report on September 2017, for the past three years with 1.6 billion dollar investment ethio telecom did a vast telecom expansion. Based on the report after the expansion, ethio telecom currently has the capability to give services for more than 62 million mobile network, 21 thousand km optical fiber linkage, above 3 million fixed line network and have more than 42.6 Gbps international internet link.

Telecommunication fraud occurs whenever a person committing the fraud uses deception to receive telephony services free of charge or at a reduced rate. It is a worldwide problem with substantial annual revenue loss for many companies and countries as a whole.

Globally, telecommunications fraud is estimated to be around 55 billion US dollars and in the United States of America alone, 'telecommunication fraud has a share of 2% of network operators' revenue in 2015. However, it is difficult to provide precise estimates since some fraud may never be detected, and the operators are not willing to reveal figures on fraud losses. Sometimes they may not have the evidence and the technique to stop the fraud but they have only the information from different sources. [19].

The situation can significantly be worse especially for mobile operators in Africa, as a result of fraud they become liable for large hard currency payments to foreign network operators. Thus, telecommunication fraud is a significant problem which needs to be addressed, detected and prevented in the strongest possible manner. Popular examples of fraud in the telecommunication industry include subscription fraud, identity theft, SIM box or voice over internet protocol (VoIP) fraud, cellular cloning, billing and payment fraud on telecom accounts, prepay and postpaid frauds and PBX fraud.[11]

Among the revenue sources of ethio telecom, international traffic takes the lion share of it. As [12] indicated on reports 40% of Ethiopian Telecommunications Corporation revenue is from international traffic. SIM box fraud or gateway fraud is one of the fraud types that attack the revenue from international traffic. The researcher understood from domain experts working in international traffic area that the international incoming call termination tariff is currently 0.19 USD per minute. The fraudsters involved in SIM box fraud paying only 0.83 Ethiopia birr (for peak hour) and 0.35 Ethiopia birr (for off-peak hour) for international calls coming through the SIM box. In this regard ethio telecom is losing the difference per minute. In addition, it has negative effect on telecom security and quality of service.

SIM box fraud is affecting not only ethio telecom but also other telecom operators in Africa like Ghana, Congo, and Sudan. SIM box fraudsters re-route international calls by using SIM box device and local SIM cards. It is also one of the reasons that telecom operators lose millions of dollars every year. [15]

In this regard, fraud detection and prevention mechanisms help the company to increase revenue by minimizing loss through fraudulent practices in the country. Both security division and revenue assurance section of ethio telecom are working to protect and assure the revenue as well as existence of ethio telecom by minimizing loop-holes for revenue leakage. In this regard, this work will have its own contribution for companies as well as for the country as a whole.

1.2 Statement of the Problem

Mobile telecommunication fraud refers to illegal access to the mobile operator's network, using their services for unlawful interest to the detriment of the network operators and its subscribers. Ethiopia's telecom industry is a fast growing sector of the economy with increasing numbers of subscribers. But telecom fraud has been a major interference to the rapid growth of this industry as it has caused in telecommunication operators revenue loss and its subscriber's loss quality of service. Ethio telecom lose millions of dollars from its annual revenue due to fraud. Telecom fraud describes each attempt to use the telecom operator network without intention of paying for it.

Ethio telecom provides different services, like voice and data services including value added services using prepaid and postpaid payment mechanism. Prepaid is most popular service and all transactions in this service is by recharging the account before usage and postpaid services means credit facilities are given for services used for some period of time. Among the calls you received, one of them could be international call but the displayed number is a local mobile number. You might have asked yourself as to how this could be happen. If you see international call displayed as local number and if you find it as missed call, you might have tried to dial it but it didn't work. This kind of fraud is done using SIM boxes that are used by passing telecom operators normal international route. Such bypass results in reduced voice call quality, security

issues and also reduced revenues for the telecom operators. These kind of calls are coming through internet using broadband and 3G data connection, the SIM boxes are located anywhere in the country. The SIM boxes take many SIM cards and as they connect with internet, it will simply divert the international call to local call and as a result, ethio telecom will not have the chance to know these international calls are being terminated in its network. In this study, the algorithms will enable us to identify or predict such fraudulent calls, based on the behaviors of the fraudulent number. For detail analysis, call detail records (CDR) are used including, SMS, GPRS, HLR and CBS data. SIM Box fraud is very difficult to detect for telecom operators as it is coming through voice over IP (VOIP) and appears like local call. By analyzing ethio telecom CDR data and integrating the results, this study will enable to predict the fraudulent numbers.

Telecommunications in general produce a large amount of data every minute. Due to the nature and size of the data, it is almost impossible to analyze the data manually. A very large amount of data is generated from different network elements and telecom applications like CBS, HLR, CDR, UNMS, network alarms that are generated from each network devices, CRM customer service application system for sales and collection, Z-smart Trouble Ticket application system, log files of different services that the company provide for customers and many more. All these data are stored or pushed to different servers, by using excel and SQL integrated those data.

In this study, the power of data mining approaches will be appreciated which can help to extract different useful knowledge, pattern and prediction ability from these data. Special focus will be given on investigating clear data, the current real scenario of SIM box fraud in ethio telecom and selecting efficient algorithm to develop prediction model for identifying Fraudulent calls behavior and SIM box fraudulent numbers.

1.3 Objective

1.3.1 General Objective

The main objective of this research is to identify the fraudulent numbers and come up with best prediction algorithm by comparing the results of different data mining algorithms in detecting SIM box fraud in ethio telecom case.

1.3.2 Specific Objective

- To understand the fraudulent case and different telecom operators experience by reviewing literatures.
- Undertaking detail analysis on SIM box fraud by using appropriate data.
- To identify the SIM box fraud calls and identify which call detail record attributes more significant in order to identify the fraudulent numbers behavior and fraudulent calls.
- To measure and compare the prediction performance of PART, J48, Multilayer perceptron and Hybrid data mining algorithms.
- To select the best detection algorithms that will help to predict and detect the SIM box fraud.

1.4 Scope

The scope of this research is selecting the efficient data mining classification algorithms to predict and detect SIM box fraud in ethio telecom scenario. This study focus on SIM box fraud and used pre-paid subscribers call detail record, including SMS, GPRS, CBS and HLR data from ethio telecom that will consequently help to identify the fraudulent calls. Most CDR attributes used for this study are selected from customer relation management (CRM) database. This study also gives emphasis on analyzing valuable ethio telecom data, select efficient classification algorithms and build a hybrid algorithm to predict fraudulent numbers.

1.4.1 Limitation of The Study

Ethio telecom experts were very helpful in order to identify the behavior of fraudulent numbers but their involvement on data selection was limited because of company business security. During the preparation of this paper, request for IMEI (international mobile equipment identity) number and IMSI (International Mobile Subscriber Identity) was made**,** but unfortunately the data was unobtainable due to security reason. This research is done based on calling number, called number, SMS, GPRS, date and time, call fee, location number and duration of call detail records.

1.4.2 Current SIM Box Fraud Handling Practice of Ethio Telecom

Ethio telecom currently have fraud management section under security division. The fraud management section works on SIM box fraud detection and others fraud type prevention and detection mechanisms. But for security reasons, they didn't want to mention how they treat such kind of fraud type and the detection mechanisms. Additionally, there are different departments in the company like customer service, shops and back office departments that communicate to the section via email when they got such kind of information. Sometime the subscribers call to ethio telecom free call service 994 to ask what could happen on their mobile call when they receive international call via local service number. Still present day, ethio telecom suffer due to SIM box fraud and losses millions of dollars per year.

1.4.3 Justification of The Thesis

For this research, the data sets obtained from ethio telecom contains both labeled and row data set. Different data records are used in this paper, that have never used by previous researchers like CBS, HLR and CRM data. This study is also unique from previously made researches on fraud detection in many aspects, especially because it mainly focus on SIM box fraud in the case of ethio telecom. The previous local research focused on fraud detection in general and subscription fraud detection. [14] [7]. The data set is dissimilar with others and this thesis specifically uses decision

Tree, rule induction and neural network algorithms to build a hybrid algorithm from J48 and PART algorithm. The obtained result is also different from previously made researches.

1.4.4 Significance of The Thesis

Telecommunication industry has expanded dramatically in the last few years with the development of affordable mobile phone technology. With the increasing number of mobile phone subscribers, global mobile phone fraud is also set to rise. Now a days, it is becoming one of the severe threats to revenue and quality of service in telecom providers. It is a worldwide problem with substantial annual revenue losses of many companies. [17].

SIM box fraud is one of such frauds that has emerged with the use of VoIP technologies. Mostly, it occurs when the cost of terminating domestic or international calls exceeds than the cost of a local mobile to mobile call in a particular region or country. Because the fraudulent take the difference of the incoming international and local call tariff. Ethio telecom receive incoming international call with the cost of 0.19 dollar per minute and local mobile to mobile call tariff is 0.83 cent on peak hour and 0.35 on off peak hour.

In this research, a total of nine features that are useful in identifying SIM box fraud subscriber are derived from the attributes of the Customer Database record (CDR). This study also different with previously made researches on fraud detection in many aspects, mainly because it focuses on SIM box fraud.

SIM box fraud is becoming worse from time to time. Ethio telecom uses new systems like CBS, CRM. In this study uses Call Detail Records (CDR) including SMS, GPRS, CBS and other derived attributes those generated from different ethio telecom systems. The pervious researches used MATLAB, brain maker, J48 and neural network algorithms. But in this thesis, Decision Tree, rule induction and neural network algorithms are used, since data mining classification algorithms has shown promising solutions in SIM box fraud problem due to their generalization capabilities. The current scenario of SIM box fraud is more complicated and ethio telecom losses millions of dollars each year.

This study indicates and come up with the best detection algorithm and model that have the capability to predict SIM box fraud telephone numbers with best prediction accuracy. It also identifies the behavior of fraudulent numbers and how the fraudulent divert the incoming international call to local call.

The output of the research will contribute to the understanding of theoretical views and practical problems in the detection of SIM box or bypass telecom fraud. This study also expected to be a very useful reference for further research in telecom fraud area.

1.5 Methodology

For this study the selected data mining tool is WEKA, that stands for Waikato Environment used for Knowledge Analysis. It is developed at the University of Waikato in New Zealand, written in java (object oriented programming language) and tested under different operating systems. Fraud cases can be similar in content and appearance but usually are not identical. Fraud is an adaptive crime, so it needs special methods of data analysis to prevent and detect it. In this study different classification algorithms are used like J48, PART, hybrid and multilayer perceptron. Different researchers working on telecommunication and credit card fraud detection recommend those algorithms. [17][10]. By applying these algorithms on preprocessing data, the efficient algorithm is selected to detect SIM box fraud and predict fraudulent numbers. Due to the nature of the data bulkiness different tools like MS Excel and MS Access are used for data preprocessing. WEKA, Eclipse and MYSQL are used for data analysis and experimentation.

For the purpose of conducting this research the CRISP-DM process model is selected. The model includes six phases that address the main issues in data mining .The six phases includes business understanding, data understanding, Data preparation, modeling, evaluation and deployment.

In this study, data mining classification algorithms like decision tree, rule based induction, artificial neural network and hybrid (J48 and PART) classification algorithms are used in order to identify fraudulent numbers.

Chapter Two

2. Telecommunication fraud and Literature Review

In this section we will discuss about the overview of fraud in telecommunication, types of telecommunication frauds, fraud detection in telecommunication industry and some previous work related to SIM box fraud detection. Detection mechanisms of SIM box fraud and impacts of SIM box fraud are also delivered. By the end of the chapter, it is hoped that the reader will understand the types of fraud and different methods that can be used to commit fraud.

2.1 Fraud in Telecommunication Industry

Fraud is an unceasing risk to network operators' revenue and it remains difficult to predict exactly how, when, or where new fraud settings will attempt to attack services. Within the telecommunications industry, fraud is an ever-increasing and most prolific threat. Now a days, it is becoming more pervasive and sophisticated. Telecommunication fraud encompasses a variety of illegal activities on telecom operator network. There are different types of frauds, which adversely affect the carrier providers, not only financially but also in terms of extensive voice bandwidth , service quality and network resources. Some of them are: - call and SMS spamming fraud, private branch exchange fraud, subscription fraud, premium rate fraud, and domestic revenue share fraud and SIM box fraud.

2.1.1 Call and SMS Spamming Fraud

Call and SMS spamming is like email spam. Subscribers receive unwanted calls and SMSs about a deal. In the case of SMS spam, the message will have a text to call a specific number or visit a website, which will promote the subscriber to redeem the offer. After that, the subscriber presses or calls the provided link, which will result in premium charges. What distinguishes call and SMS spamming from email spam is that a subscriber might be charged for receiving a spam SMS from websites. Also, once the subscriber replies to the spam number, he or she will be charged regardless of the subscribed plan. Contrary to email, there is no filtration on call and SMS replies, unlike the

case of junk email. Some operators created a mechanism to fight SMS spam. The subscriber can report the spam SMS by forwarding it to specific numbers, but still, there is no built-in mechanism to separate spam SMS on an industry level. [7].

As of March 2005, internationally SMS spamming is illegal to send to users who haven't specifically asked for them. However, there is a loophole in the law: solicitors are only prohibited from sending unwanted messages to cell phones from Internet domains. They can still send these messages from a cell phone.

2.1.2 Private Branch Exchange (PBX) Fraud

Private branch exchange (PBX) fraud happens when the fraudster takes over the private switching network and uses linked external phone lines to make calls to premium numbers owned by the fraudster. Private branch exchange fraud occurs when the internal network of an organization is not secure enough from outside attacks. A lot of ways are used to take control of a PBX. Companies might leave default passwords unchanged or they could be corrupted through social engineering, another option might be the attack comes from an internal employee or a vendor. [19].

2.1.3 Subscription Fraud

The subscription fraud is the most common since with a stolen or manufactured identity, there is no need for a fraudster to undertake a digital network's encryption or authentication systems. Subscription fraud occurs in the phase of signup. The fraudster uses stolen information (SSN, address, or credit card account) to login to services provided by an operator. After signup is complete, the fraudster will commit the fraud and will be billed for general usage. Once the fraudster does not pay the outstanding amount, the amount will be sent to collection agencies, which will rely on the account information that was fake or stolen. In this case, the account was opened using stolen information and now the original owner of the information will be required to pay the outstanding amount. Since the information was fake, no one can be required to pay the

outstanding amount, so that amount will be accumulated in the operator's account as a bad debt. [19].

According to GSM Association and the Communications Fraud Control Association, subscription fraud is the starting point for many other telecoms fraud and as such is recognized as the most damaging of non-technical fraud types.

2.1.4 Premium Rate Fraud

Premium rate service fraud is the second largest contributor to the $46.3 billion problem of mobile fraud in 2013. It rakes in $4.73 billion globally and $1.35 in North America of losses for subscribers annually. This type of fraud directly attacks subscribers by getting them to make calls to a premium rate telephone number. [26].

The most common occurrences of premium rate service fraud directly attack phone companies through the subscription fraud method. It is a fairly basic scheme that takes advantage of phone billing cycles. Fraudsters set up a premium - rate phone number through a carrier and subscribe for one or multiple phone lines through a different carrier using false information. They then run auto dialers on the subscriber lines that call the premium rate numbers, running up extremely large bills. They don't pay the subscription bills, but receive the profits from the premium -rate line. This goes on until the phone company begins to investigate a bill for non -payment, and then the fraudsters simply close out their services leaving the bills unpaid at the expense of the phone company. [7].

2.1.5 Domestic revenue share fraud

Domestic revenue share fraud pertains to the abuse of carrier interconnect agreements and is very similar to international revenue share fraud and premium rate service fraud. In all three scenarios, there is an artificial inflation of traffic to a premium rate phone number. The scheme is fairly simple: A fraudster gets hold of a premium rate service number a phone number where a portion

of the charges goes to the operator and not only the phone carrier like with regular phone numbers and inflates the traffic to the service to generate more revenue.

There are many different ways this is done and they range from very simple to calculated and organized. One of the simplest methods is by dialing a phone number just long enough to place a missed call on victims' phones but not long enough for them to pick up, so as to lure them into calling back. This fraud method has become popularly known as "One Ring" or, in its more advanced variant, Wangiri fraud. More sophisticated methods of artificial traffic inflation like PBX and voicemail hacking are very common today. And there are bluetooth- based attacks that can replace mobile phone numbers with premium rate phone numbers. This not only increases the fraud revenue, but also enables fraudsters to listen into phone conversations. VoIP hacking is also common, where hackers introduce their premium number into victims' communications as a call-through service. With the evolution of technology, it's only normal that fraud becomes more organized. [17].

2.1.6 SIM box Fraud

Fraudulent SIM boxes hijack international voice calls and transfer them over the Internet to a cellular device, which injects them back into the cellular network. By-Pass Fraud occurs when in-bound off-network traffic is disguised as on-network traffic (By-pass) to avoid high costs of terminating traffic. Most By-pass operations are performed on a large scale utilizing advanced SIM boxes that can be managed from anywhere.

Content Service Providers attacked can experience significant losses in their in-bound interconnect revenues. Service providers should constantly monitor in-bound and on-net traffic in order to detect any indications associated with SIM box fraud, such as suspected calling numbers or suspicious call pattern tendencies.

SIM box is a hardware which is used to bypass the legal or normal route for international incoming call.[1]. Figure 1 shows SIM box device which have SIM slots, antennas and Ethernet ports that can be used to get the SIM box equipment connected to the internet. SIM boxes are used as part of voice over IP gateway installation and the function of SIM box is used to make and

terminate international incoming call to local call. The fraudsters can forward international calls through local phone numbers in the respective country to make it appear as the call is a local call. (source:[1]).

Figure 1: Two models of SIM box device (New module 32 SIM card gsm SIM box and 128 SIMs cards call center SIM box device)

Current SIM box equipment have advanced features that help to fraudsters while forwarding the calls like SIM automated rotation, changeable international mobile equipment identity (IMEI), behavior pattern setup and etc. SIM box equipment can be found and purchased online easily throng companies like EBay and Amazon.

A typical SIM box has 32 modems and antennas, make calls continuously, it causes congestion problems. SIM box voice fraud mostly occurs where the cost of terminating international call exceeds the cost of mobile to mobile call in the country. Fraudulent SIM boxes hijack international voice calls and transfer them over the internet to a cellular device, which injects them back into the cellular network. And fraudsters make a profit by offering low cost international voice calls to the operators and to bypass call routing fees they buy or hijack large amounts of SIM cards and install them into hardware (SIM Boxes). Then the fraudsters transfer a call via the internet to a SIM box in the area of call recipient to deliver the call as local. As a result, the operators serving the called party do not receive the corresponding call termination fees.

SIM box fraud also creates a lot of quality issues, like delay, echoes and noise on the line .This quality issues, cause people to make shorter duration calls. The caller telephone number is not visible on the receiver phone, so someone is not sure who is calling. The fraudsters enjoy some portion of the difference between the international termination rate and local tariff. Countries, especially developing countries, in Africa like Ethiopia, Ghana, Congo and Asia suffer this loss due to high incoming traffic of international call for different reasons. [19]. The following Figure 2 shows how the SIM box works and the routes for both legal and illegal ones. (source: [20]).

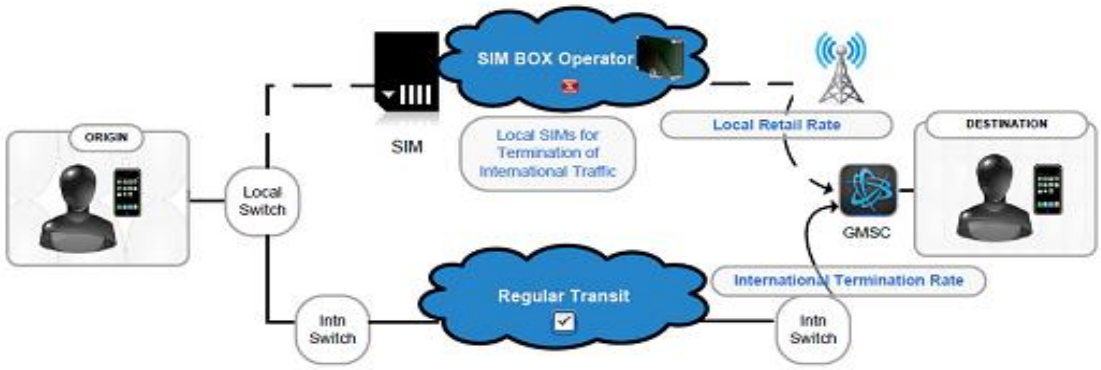

Figure 2: Example of Legitimate call and SIM box fraud hijacking of an international call SIM box fraud analysis

In order to identify the SIM box numbers or SIM cards that are used by fraudsters, the numbers have different behaviors those mentioned by different researchers and ethio telecom experts. The fraudulent numbers didn't have outgoing call detail records and most of SIM boxes are clustered around two areas without any legitimate account and legitimate accounts are clustered around another area. It can be observed that most of SIM boxes have originating calls more than terminating ones while legitimate accounts have comparable number of originating and terminating calls. That is because SIM boxes are used mainly to regenerate the calls received from the VOIP branch and make them GSM calls again. This feature is very useful to distinguish between SIM boxes and legitimate accounts. [17].

SIM boxes are installed to one location and could be moved from time to time, legitimate users are usually not tight to a specific location. This feature also very attractive to utilize in order to detect SIM box accounts. It can be noticed that most SIM boxes don't send SMS and not use data but most of the time legitimate accounts have SMS and GPRS call Details. Mostly fraudulent calls happen on peak of hour and weekend days. Based on the analysis above can conclude that from the total call of incoming and outgoing call, SMS, GPRS, call location, call fee are explored features. Call location number and total number of call per day, SMS and GPRS feature can give the highest distinguish rate between SIM boxes and legitimate accounts.

2.2 Methods Used to Detect SIM box Fraud in Telecommunication Industry

Operators and regulators have devised several fraud detection schemes for generated revenue assurance. There are different techniques and tools to detect the SIM box fraudulent internationally. Test call generation is one of them and it has proven with the identification of SIM box fraud and it is an effective method for detecting fraudulent numbers. Test calls are initiated to those numbers from various countries by using different interconnect voice routes worldwide. By using test call generation they identify the paths followed to reach the SIM boxes in the home country .Test call generation is a probabilistic method in which the number of fraudulent SIM boxes identified increase as more calls using more routes are generated. These methods are relatively new to network operators, an understanding of these methods is the key to managing revenue assurance. [5].

The other detection technique is CDR analysis and analytics. CDRs are used to identify fraudulent activities through extensive analysis while performing analytics on fraud indicators by comparing different fields of the CDR like calling number, called number, call type, IMEI, IMSI, time and call duration. In SIM box detection the fraud management system (FMS) uses CDRs user based profiling that distinguish between fraudulent SIMs installed in SIM boxes and legitimate users. Call generation providers and FMS tools providers collaborate to pool their alerts in order to more efficiently detect SIM box fraud. In this study ethio telecom CDR data used to identify the fraudulent number behavior and SIM box fraudulent calls. [4].

2.3 Impacts of SIM box Fraud in Telecommunication Industry

SIM box fraud have different effects on telecom operators, regulators and subscribers. Few major impacts are: -

Revenue loss due to call termination

Many developing countries like Ethiopia, Ghana and Nigeria telecom operators, generates high revenue from international call. Due to SIM box fraud international call are redirected, intercepted and terminated, then the fraudster proceeds the incoming international revenue. Then Telecom operators lost their international incoming call revenue.

Revenue loss due to service inaccessibility and missing call backs

SIM box fraud have the negative effect on multiple telecom service like voice mail, SMS and calling back. Due to SIM box fraudulent international incoming call displayed as local number on receiver side. Then it have immediate impact on the ability to call back to the caller or sending SMS and setting voice mail. This all resulting a major opportunity loss of retail revenue.

Image loss due to bad quality of service

SIM box fraud generally is based upon redirecting calls over inadequate, highly compressed IP connections, resulting in poor voice quality , echo's and increased call failure rates because of congestion caused through use of a Voice over IP. Call setup time or routing delays are extended which also leads to the impression of an overall bad service quality by the local network operator.

Additional Investment

Sometimes traffic hot-spots and congestion caused by bypassed traffic can lead to substantial unnecessary site acquisition and roll-out costs for new radio access equipment like Basic Transceiver station and mobile switching center.

2.4 Literature Review

Shawer and Burge [1] investigated detection of fraud in mobile communication, European Project ASPecT (Advanced security for personal communications technologies). The ASPecT fraud detection tool is based on investigating sequences of call detail records (CDRs), which contain the details of mobile call attempt for billing purpose. The information produced for billing contains usage behavior for fraud detection. A differential analysis is performed to identify a fraudster through profiling the behavior of a user. ASPecT fraud detection tool utilizes a rule based system for identifying certain frauds and neural networks to deal with abnormal scenarios.

Bolton and Hand [2] study based on the statistical and machine learning technology for fraud analysis and detection including their application to detect activities in credit card fraud, telecommunication fraud and computer intrusion. Their paper has identified the different types of fraud, such as bankruptcy fraud, counterfeit fraud, theft fraud, application fraud and behavioral fraud, and they used different methods to detect them. Such methods have included pair-wise matching, decision trees, clustering techniques, neural networks, and genetic algorithms.

According to Ilona Murynets and Adam Panagia [5] study the fraudulent traffic from SIM boxes operating with a large number of SIM cards. They processes hundreds of millions of anonymized voice call detail records (CDRs). In addition to overloading voice traffic, fraudulent SIM boxes are observed to have static physical locations and to generate disproportionately large volume of outgoing calls. Based on these observations, novel classifiers for fraudulent SIM box detection in mobility networks are done.

Farvaresh and Seperi [10] applied decision tree (DT), neural network (NN) and support vector machine (SVM) in order to identify customer with residential subscription of wire line telephone service but used it for commercial purposes to get lower tariffs which is classified as subscription fraud. The employed data mining approach consist of preprocessing, clustering and classification phases. Combination of SVM and K-Means were used in the clustering phase and decision tree (C4.5), Neural Network, SVM as single classifiers were examined in the classification phase. The results are presented in terms of confusion matrix. DT, NN and SVM as single classifiers were able to correctly classify 88.1%, 84.9% and 88.2% respectively. Therefore, SVM has shown best performance among all the classifiers.

Krenker et al. [9] proves that using bi-directional Neural Network (bi-ANN) in predicting generic mobile phone fraud in real time gave high percentage of accuracy. Bi-ANN is used in prediction the time series of call duration attribute of subscribers in order to identify any unusual behavior. The results show that bi-ANN is capable of predicting these time series, resulting 90% success rate in optimal network configuration. However call duration is the only parameter used, therefore other relevant parameters are missing to accurately predict customer behavior.

Abdikarim and Roselina Sallehuddin [17] outlines the Artificial Neural Network (ANN) and Support Vector Machine (SVM) to detect Global System for Mobile communication (GSM) gateway bypass in SIM Box fraud. The suitable features of data obtained from the extraction process of Customer Database Record (CDR) are used for classification in the development of ANN and SVM models. The performance of ANN is compared with SVM to find which model gives the best performance. From the experiments, it is found that SVM model gives higher accuracy compared to ANN by giving the classification accuracy of 99.06% compared with ANN model, 98.71% accuracy.

Bülent [18] examine the call detail records (CDR's), demographic data and payment data of mobile subscribers in order to develop models of normal and fraudulent behavior via data mining techniques. First they have done some Exploratory Data Analysis (EDA) on the data set and discovered that some variables like Account length, Package type, Gender, Type, Total Charged Amount showed important tendency for fraudulent use and then they applied k-means cluster method to cluster the customer, based on their call behaviors. Standard variables with ranked attributes and variables obtained from factor analysis due to some correlated variables were used as two different set of variables performed the data mining techniques. Decision trees and Neural Networks for both training and test sets and then discussed the collected results based on performance measures such as accuracy, sensitivity, specificity, precision and RMSE.

Chapter Three

3. Data Mining and Knowledge Discovery

3.1 Data Mining

According to the Gartner Group, "Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques.

Data mining, which is also called knowledge discovery in databases (KDD), can also be defined as “the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both”.[19]. Data mining combines different techniques from statistics, databases, machine learning and pattern recognition to extract (mine) concepts, concepts interrelations and interesting patterns automatically from large business databases.

Innovative organizations worldwide are already using data mining to locate and appeal to higher value customers, to reconfigure their product offering to increase sales, and to minimize losses due to errors or fraud. Data mining is a process that users a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid prediction. Data mining is also a tool for increasing the productivity of people trying to build predictive models.

Data mining can be broadly divided into two, verification driven and discovery driven data mining. Verification driven data mining extracts information in the process of validating a hypothesis postulated by user, it uses techniques such as statistical and multidimensional analysis. Discovery driven data mining applies tools such as symbolic neural clustering association discovery and supervised induction to automatically extract information.

Telecommunications and credit card companies are two of the leaders in applying data mining to detect fraudulent use of their services. Insurance companies and stock exchanges are also interested in applying this technology to reduce fraud. [18]. The following Figure 3 shows an illustration of the process of extracting knowledge from data using data mining. (source: [18]).

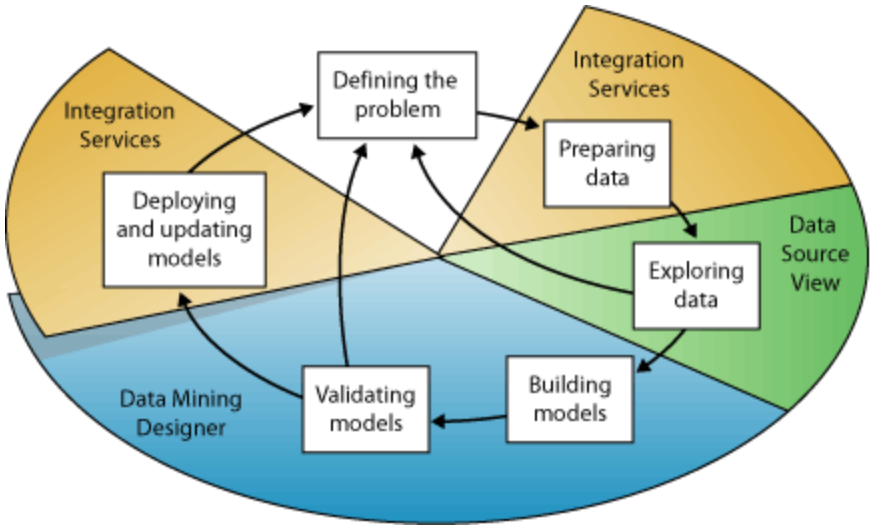

Figure 3: A typical data mining process

3.2 Knowledge Discovery Process

Knowledge discovery process model helps organizations to better understand the knowledge discovery process and provides a roadmap to follow while planning and performing the research. The knowledge discovery process (KDP) also called knowledge discovery in large databases, seeks new knowledge in different application domains like in machine learning ,pattern recognition, databases , statistics, artificial intelligence, knowledge acquisition for expert systems and data visualization . Knowledge discovery concerns the entire knowledge extraction process including how data are stored and accessed, how to use efficient and scalable algorithms to analyze massive datasets, how to interpret and visualize the results and how to model and support the interaction between human and machine. Data mining is the core part of the knowledge discovery process. The data mining and KDD are often used interchangeably because Data mining is the key part of KDD process. [11].

The term Knowledge Discovery in Databases (KDD) refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. The goal of KDD process is to extract knowledge from data in the context of large databases by using data mining methods(algorithms) to identify what is deemed knowledge, according to the specifications of measures and thresholds, using a database along with

any required preprocessing, sub sampling, and transformations of that database . The KDD process is shown diagrammatically in figure 4 below, (source: [16]).

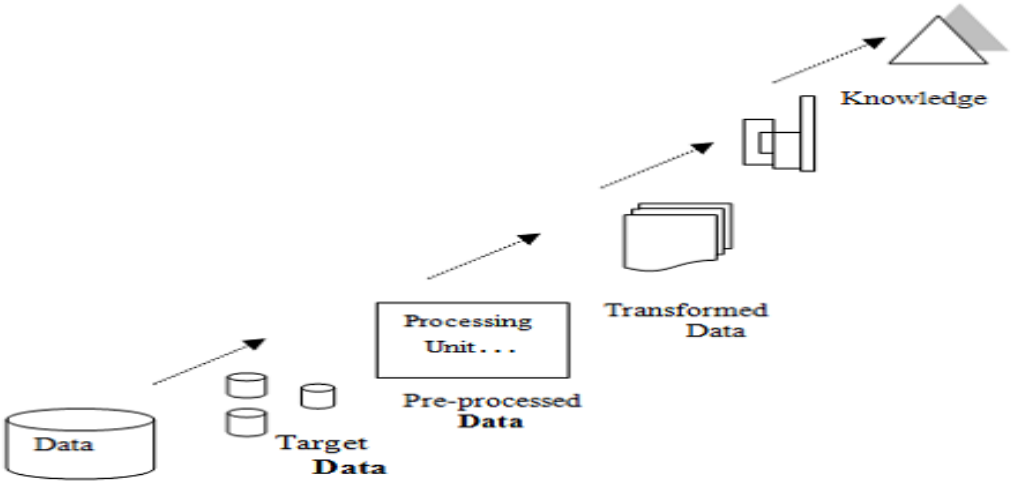

Figure 4: A typical knowledge discovery process

3.3 Data Mining and Knowledge Discovery Process Models

The Knowledge discovery process model consists of a set of processing steps to be followed by practitioners when performing a knowledge discovery research. There are different process models including Academic Research Models, Industrial Models and Hybrid Models.

The efforts to establish a KDP model were initiated in academic world in the mid-1990s, when the DM field was being shaped, researchers started defining multistep procedures to guide users of DM tools in the complex knowledge discovery world. The main emphasis was to provide a sequence of activities that would help to execute a KDP in an arbitrary domain. The nine-step model by Fayyad et al. and the eight-step model by Anand and Buchner academic research models developed in 1996 and 1998. [10].

While industrial models quickly followed academic efforts and developed two industrial models by a large consortium of European companies. Those are the five-step model by Cabena et al., with support from IBM and the industrial six-step CRISP-DM model. In this study CRISP-DM model will be discussed.

3.3.1 CRISP-DM Model

The general CRISP-DM process model includes six phases that address the main issues in data mining. [21]. The six phases fit together in a cyclical process, illustrated in the following figure 5 below. [21].

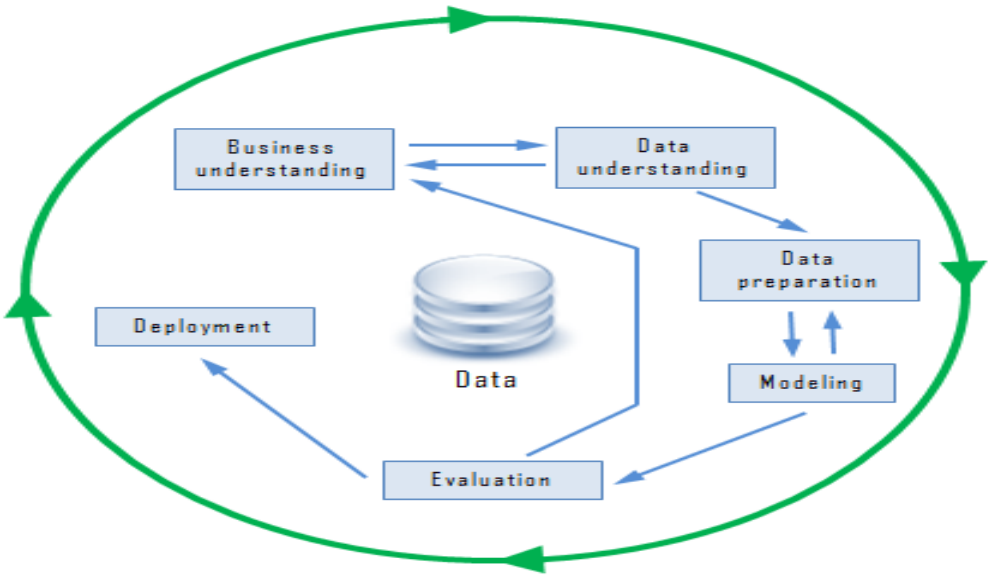

Figure 5: Phases of the CRISP-DM reference model

These six phases cover the full data mining process, including how to incorporate data mining into large databases. The six phases are defined as follows:

- Business Understanding: This is perhaps the most important phase of data mining. Business understanding includes determining business objectives, assessing the situation, determining data mining goals, and producing a project plan.
- Data Understanding: Data provides the "raw materials" of data mining. This phase addresses the need to understand what your data resources are and the characteristics of these resources. It includes collecting initial data, describing data, exploring data, and verifying data quality.

- Data Preparation: After cataloging the data resources, the data will be prepared for mining. Preparations include selecting, cleaning, constructing, integrating, and formatting data.
- Modeling: This is, of course, the flashy part of data mining, where sophisticated analysis methods are used to extract information from the data. This phase involves selecting modeling techniques, generating test designs and building and assessing models.
- Evaluation: Once models are chosen, users are ready to evaluate how the data mining results can help to achieve their business objectives. Elements of this phase include evaluating results, reviewing the data mining process, and determining the next steps.
- Deployment: This phase focuses on integrating new knowledge into everyday business processes to solve original business problems. This phase includes plan deployment, monitoring, and maintenance, producing a final report, and reviewing the project.

3.4 Data Mining Tasks

Data mining is the field in which useful outcome that is being predicted from large database. It uses different tools to get out the useful hidden patterns, trends and prediction of future can be obtained using the techniques. [8]. Figure 6 below shows the classification of data mining tasks and models. [16].

The goal of any data mining effort can be divided in one of the following two types.

- To generate descriptive models to solve problems.
- To generate predictive models to solve problems.

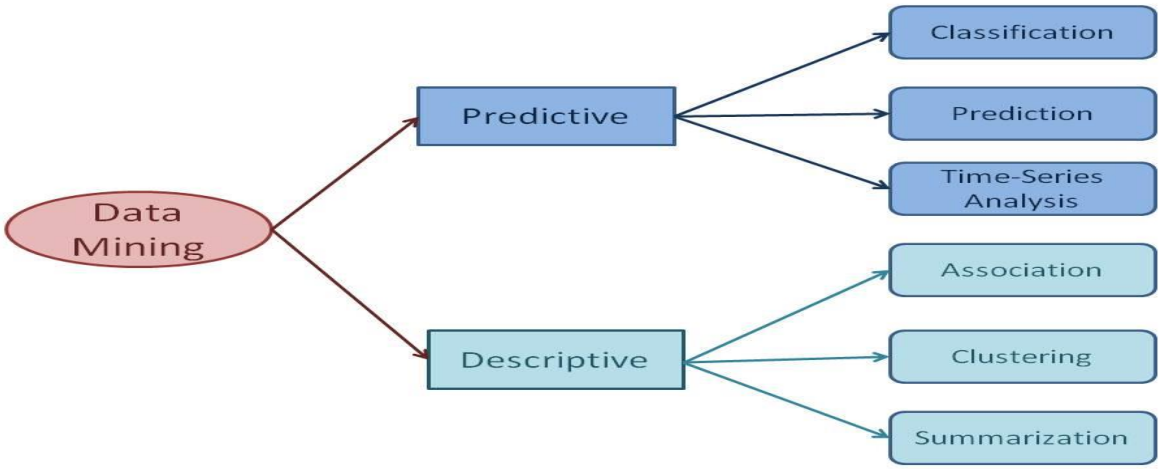

Figure 6: Data mining tasks and models

3.4.1 Descriptive Model

The descriptive data mining tasks characterize the general properties of the data in the database, while predictive data mining tasks perform inference of the current data in order to make prediction.

Descriptive data mining focus on finding patterns describing the data that can be interpreted by humans, and produces new, nontrivial information based on the available data set. The goal of descriptive data mining is to gain an understanding of the analyzed system by uncovering patterns and relationships in large data sets.

The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined. Ex. Clustering, Summarization, Association rule, Sequence discovery etc. Clustering is similar to classification except that the groups are not predefined, but are defined by the data alone. It is also referred to as unsupervised learning or segmentation. It is the partitioning or segmentation of the data in to groups or clusters. The clusters are defined by studying the behavior of the data by the domain experts. The term segmentation is used in very specific context; it is a process of partitioning of database into disjoint grouping of similar tuples. Summarization is the technique of presenting the summarize information from the data. The association rule finds the association between the different attributes. Association rule mining is a two-step process:

Finding all frequent item sets, and generating strong association rules from the frequent item sets. Sequence discovery is a process of finding the sequence patterns in data and this sequence can be used to understand the trend.

3.4.2 Predictive Model

Predictive data mining involves using some variables or fields in the data set to predict unknown or future values of other variables of interest, and produces the model of the system described by the given data set. The goal of predictive data mining is to produce a model that can be used to perform tasks such as classification, prediction or estimation. It is used to predict the future outcomes based on previous records with known solutions.

The predictive model makes prediction about unknown data values by using the known values. Ex. Classification, Regression, Time series analysis, Prediction etc. Many of the data mining applications are aimed to predict the future state of the data. Prediction is the process of analyzing the current and past states of the attribute and prediction of its future state. Classification is a technique of mapping the target data to the predefined groups or classes, this is a supervised learning because the classes are predefined before the examination of the target data. In the time series analysis, the value of an attribute is examined as it varies over time and the distance measures are used to determine the similarity between different time series, the structure of the line is examined to determine its behavior and the historical time series plot is used to predict future values of the variable.

The data mining task of generating models are divided into the following two approaches:

- Supervised or directed data mining modeling – use the available data to build a model that describes one particular variable of interest in terms of the rest of the available data. The task is to explain the values of some particular field and the user selects the target field and directs the computer to determine how to estimate, classify or predict its value.

- Unsupervised or undirected data mining - The goals of predictive and descriptive data mining are achieved by using specific data mining techniques that fall within certain

primary data mining tasks. The goal is rather to establish some relationship among all the variables in the data. The user requests the computer to identify patterns in the data that may be significant. Undirected modeling is used to explain those patterns and relationships once they have been found.

In this study a predictive model is developed, in order to predict the SIM box fraudulent calls and to identify the fraudulent numbers behavior.

3.5 Classification Algorithm

Classification involves the discovery of a predictive learning function that classifies a data item into one of several predefines classes. It involves examining the features of a newly presented object and assigning to it a predefined class. First , a model is built describing a predetermined set of data classes or concepts and secondly the model is used for classification.

Prediction can be viewed as the construction and use of a model to assess the class of a unlabeled sample, or to assess the value or value range of an attribute that a given sample is likely to have. In this research typical business related questions that can be answered using classification or prediction tasks like which call is fraudulent, which one is legitimate call or what kind of behavior fraudulent calls have.

Classification is one of data mining technique, which is useful for predicting group membership for data instances. It is a supervised kind of machine learning, which provides the training data as trained .Classification predicts categorical continuous valued functions. Prediction is the form of predicting the class value to which data belong. [18]

Classification is the derivation of model, which determines the class of an object based on its attributes .In this study different classification algorithms are used to build prediction model and to categorize either the calling number is fraudulent number or legitimate number. A set of object is given as training set in which every object is represented by vector of attributes along with its class, by analyzing the relationship between attributes and classes of the objects in the training set, classification model can be constructed. Such classification model can be used to classify future

testing datasets or objects. The training data set includes (calling number, called number, SMS, GPRS, call location number calling time, call fee and duration). Based on this attributes a classification model can be built which have the capacity to predict SIM box fraud calls.

Classification techniques and its algorithms are respectively used in this research, to predict the SIM box fraudulent calls by identifying the behavior of the legitimate and fraudulent numbers. The algorithms are decision tree, rule based induction and multilayer perceptron, which will be discussed in detail in the following sections.

3.5.1 Decision Tree

Decision trees find use in a wide range of application domains. They are used in many different disciplines including diagnosis, cognitive science, artificial intelligence, game theory, engineering and data mining. Decision trees model has two goals: producing an accurate classifier and understanding the predictive structure of the problem. The classification accuracy of decision trees has been a subject of numerous studies.

The machine learning algorithm is meant to identify patterns based on different characteristics or features and then make predictions on new, unclassified data based on the patterns learned earlier. The input data is usually numerous instances of relations between the different variables or features relevant to the data. [17] There are various different approaches to machine learning namely decision trees, random forests, neural networks, clustering, Bayesian networks, reinforcement learning, support vector machines, genetic algorithms, and many more. Decision tree learning is a method commonly used in data mining. Decision trees are powerful and popular tools for classification and prediction. They also represent rules, which can be understood by humans and used in knowledge system such as database. A decision tree represents a multi-stage decision process, where a binary decision is made at each stage.

A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called “root” that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves (also known as terminal

or decision nodes). In a decision tree, each internal node splits the instance space into two or more subspaces according to a certain discrete function of the input attributes values. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute's value. In the case of numeric attributes, the condition refers to a range.

Each leaf is assigned to one class representing the most appropriate target value. Alternatively, the leaf may hold a probability vector indicating the probability of the target attribute having a certain value. Instances are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path. The tree complexity is measured by one of the following metrics: the total number of nodes, total number of leaves, tree depth and number of attributes used. Decision tree induction is closely related to rule induction. Each path from the root of a decision tree to one of its leaves can be transformed into a rule simply by conjoining the tests along the path to form the antecedent part, and taking the leaf's class prediction as the class value. The resulting rule set can be simplified to improve its comprehensibility to a human user, and possibly its accuracy. [18].

Figure 7 shows a simple decision tree that solves this problem while illustrating all the basic components of a decision tree: the decision node, branches and leaves.

Figure 7: Example of a decision Tree

The first component is the top decision node, or root node, which specifies a test to be carried out. The root in this example is "Income > $ 40,000." The results of this test cause the tree to split into branches, each representing one of the possible answers. In this case, the test "Income> $40,000" can be answered either "yes" or "no", so we get two branches.

Decision trees models are commonly used in data mining to examine the data and induce the tree and its rules that will be used to make predictions. A number of different algorithms may be used for building decision trees including CHAI D (Chi-squared Automatic Interaction Detection), CART (Classification and Regression Trees), QUEST, and C5.0. Decision trees which are used to predict categorical variables are called classification trees because they place instances in categories or classes. Decision trees used to predict continuous variables are called regression trees. [18]

Decision trees are known as highly efficient tools of machine learning and data mining, capable to produce accurate and easy-to-understand models. They are robust and perform well with large data in short time. As we can see in different researches, decision tree is a very efficient predictive model.

Advantage of Decision Tree

Decision Trees are simple to understand and interpret. Most people were able to understand decision tree models after getting simple explanation. When decision tree is used, a value will be acquired even with little hard data, important insights can be generated based on experts describing a situation (its alternatives, probabilities and costs) and their preferences for outcomes. The algorithm are robust to noisy data and capable of learning disjunctive expressions. It helps to determine worst, best and expected values for different scenarios. [13].

Decision trees offer advantages over other methods of analyzing alternatives. Some of them are:

- Efficient. It makes possible to quickly express complex alternatives clearly. A decision tree can easily be modified as new information becomes available. Standard decision tree notation is easy to adopt.
- Revealing. It can compare competing alternatives-even without complete information-in terms of risk and probable value. The Expected Value (EV) term combines relative investment costs, anticipated payoffs, and uncertainties Decision Trees into a single numerical value. The EV reveals the overall merits of competing alternatives.
- Complementary. Decision trees can be used in conjunction with other project management tools. For example, the decision tree method can help to evaluate project schedules.

- Decision trees are self-explanatory and when compacted they are also easy to follow. In other words, if the decision trees have a reasonable number of leaves, it can be grasped by non-professional users. Furthermore, decision trees can be converted to a set of rules. Thus, this representation is considered as comprehensible.
- Decision trees can handle both nominal and numerical attributes.
- Decision trees representation is rich enough to represent any discrete-value classifier.
- Decision trees are capable of handling datasets that may have errors or missing values.
- Decision trees are considered to be a nonparametric method. This means that decision trees have no assumptions about the space distribution and the classifier structure.

Disadvantage of Decision Tree

On the other hand, decision trees have disadvantages such as:

- Most of the algorithms (like ID3 and C4.5) require that the target attribute should have only discrete values.
- As decision trees use the "divide and conquer" method, they tend to perform well if a few highly relevant attributes exist, but less so if many complex interactions are present. One of the reasons is that other classifiers can compactly describe a classifier that would be very challenging to represent using a decision tree.
- The greedy characteristic of decision trees leads to another disadvantage that should be pointed out, i.e . its over-sensitivity to the training set, to irrelevant attributes and to noise.

3.5.2 Neural Networks

Artificial Neural Networks (ANN) represent a very basic imitation of the non-linear learning mechanisms of biological neural networks. ANNs have the capability to learn from the environment and enhance their performance through learning which is achieved by an iterative process of adjusting the weights and bias level. [17].

A neuron has a number of inputs and one output. It combines all the input values (Combination), does certain calculations, and then triggers an output value. There are different ways to combine inputs. One of the most popular methods is the weighted sum, meaning that the sum of each input value is multiplied by its associated weight. Therefore, for a given node g we have:

$$Net_g = \Sigma w_{ij} x_{ij} = w_{0j} x_{0j} + w_{1j} x_{1j} + ..w_{ij} x_{ij}$$

Where x_{ij} represents the i 'th input to node j, w_{ij} represents the weight associated with the i 'th input to node j and there are I + 1 inputs to node j. The value obtained from the combination function is passed to non -linear activation function as input. One of the most common activation functions used by Neural Network is the sigmoid function. This is a nonlinear functions and result in nonlinear behavior.

Sigmoid function is used throughout this study, and is defined as follows.

$$Sigmoid = \frac{1}{1+e^{-x}}$$

Where x is the input value and e is base of natural logarithms, equal to around 2.718281828. The output value from this activation function is then passed along the connection to the connected nodes in the next layer. Back-propagation algorithm is a commonly used supervised algorithm to train feed -forward networks. The whole purpose of neural network training is to minimize the training errors. The next equation shows one of the common methods for calculating the error for neurons at the output layer using the derivative of the logistic function:

$$Err = O_i(1 - O_i)(T_i - O_i)$$

In this case, O_i is the output of the output neuron unit i ,and T_i is the actual value for this output neuron based on the training sample. The error calculation of the hidden neurons is based on the errors of the neurons in the subsequent layers and the associated weights as shown in here below.

$$Err_i = O_i(1 - O_i)\Sigma_j Err_i W_{ij}$$

O_i is the output of the hidden neuron unit i, which has j outputs to the subsequent layer. Err_j is the error of neuron unit j, and W_{ij} is the weight between these two neurons. After the error of each neuron is calculated, the next step is to adjust the weights in the network accordingly using the next equation.

$$W_{ij,new} = W_{ij} + l * Err_j * O_i$$

Here l, is value ranging from 0 to *1*.The variable l is called learning rate. If the value of l is smaller, the changes on the weights get smaller after each iteration, signifying slower learning rates.[17].

Neural networks are of particular interest because they offer a means of efficiently modeling large and complex problems in which there may be hundreds of predictor variables that have many interactions. Neural nets may be used in classification problems (where the output is a categorical variable) or for regression (where the output variable is continuous).

A neural network starts with an input layer, where each node corresponds to a predictor variable. These nodes are connected to a number of nodes in a hidden layer. Each input node is connected to every node in the hidden layer. The nodes in the hidden layer may be connected to nodes in another hidden layer, or to an output layer. The output layer consists of one or more response variables.

3.5.3 Rule Induction

Rule induction is one of the most important techniques of machine learning. Since regularities hidden in data are frequently expressed in terms of rules, rule induction is one of the fundamental tools of data mining at the same time. Usually rules are expressions of the form.

If (attribute-1, value-1) and (attribute-2, value-2) and

(attribute-n, value-n) then (decision, value).

Some rule induction systems induce more complex rules, in which values of attributes may be expressed by negation of some values or by a value subset of the attribute domain. [26].

Rule induction is a method for deriving a set of rules to classify cases. Although decision trees can produce a set of rules, rule induction methods generate a set of independent rules which do not necessarily (and are unlikely to) form a tree. Because the rule inducer is not forcing splits at each level, and can look ahead, it may be able to find different and sometimes better patterns for classification. [18].

Unlike trees, the rules generated may not cover all possible situations. Rules may also conflict in their predictions, in which case it is necessary to choose which rule to follow. One common method to resolve conflicts is to assign a confidence to rules and use the one in which you are most confident. Alternatively, if more than two rules conflict, you may let them vote, perhaps weighting theirs votes by confidence you have in rule.

Chapter Four

4. Data Preparation

This chapter focuses on data preparation process starting from data understanding, initial data collection, data description, data preparation, data integration and transformation up to data formatting.

4.1 Type of Telecommunication Data

In telecommunication field there are some critical data on which any decision can be made. This data can be distributed on three main groups call detail data (CDR), network data, customer data.

Most of the time CDR includes sufficient information describing the call such as: the originating and terminating phone numbers, the date and time of the call and the duration of the call. CDR is generated in real-time every time a call is placed on the telecommunication network. The number of CDRs that are generated and stored is huge. The CDR should be kept online several months and billions of CDRs stored. In ethio telecom case CDR details wait for six months then it sends to data warehouse.

Telecommunication networks are comprised of thousands of interconnected components. Each network component generates tremendous amount of data containing errors and status messages. This data must be analyzed to support network management functions. This data will mainly include a timestamp and a string that identified the component generating the message and a code that explains the cause of message generation. Due to the enormous number of network messages generated, technicians cannot handle every message. For this reason expert systems have been developed to automatically analyze these messages and take appropriate action and technicians are only involved when a problem cannot be automatically resolved.

Telecommunication companies may have millions of customers. Their databases maintain information on these customers. This information includes Customers‟ name, address, service plan, contract information, credit score and payment history. [6].

4.2 Data Source

Every time a call is placed on telecommunications network, descriptive information about the call is saved as call detail records. The dataset used for this thesis is obtained from ethio telecom. The data collected from different databases of the mobile operator, like CBS, CRM, HLR and CDRs.

4.3 Data set and Descriptors

Call detail records data was large because of that two months selected CDR data used for this research. Data filtered by ethio telecom domain experts based on time especially weekend, off peak hour calls and call duration .The data was large and filtering was a big challenge, by using different data preprocessing mechanisms we try to select sample data. After applying different selection criteria like call duration, total call per day, calling time ,the data minimize into 23,904 records, finally for experimentation purpose 12,686 call detail records used.

This study is based on Global Systems for mobile communications (GSM) network and specifically the Customer Data Record (CDR) database of ethio telecom prepaid subscribers. The below table contains subscriber id, the calling number, called number, duration of the call, date and time, location of calling number, SMS and GPRS and other details. The location of calling number data is found in the HLR database but ethio telecom experts generate the data with the CDR detail with attribute field ‘’CELL-A’’ . It indicates the place where the customer initiated the call and ethio telecom have detail information about each base transceiver station (BTS). The data also contains SMS and GPRS record in order to identify the SIM box fraudulent numbers .Most of the time the SIM box fraudulent numbers didn’t have SMS and GPRS records.

A Total of nine features have been identified to be useful in detecting SIM box fraud. Table 1 shows the list of these features and their description.

Table 1: Selected Data Descriptors

No	Field Name	Description	Data Type
1	A_NBR	Calling Number - Mobile number initiating or originating the call.	Number
2	B_NBR	Called number - Mobile number receiving the call.	Number
3	START _TIME	Time of call initiation (Calling time)	Date
4	DURATION	Call duration (in seconds)	Number
5	CALL_FEE	Amount paid (in cents)	Number
6	CELL_A	The place where the customer initiated the call.(BTS)	Number
7	SMS	Number of SMS made by the service(mobile) Number. Derived from SMS CDR table.	Nominal
8	GPRS	Number of GPRS connections made by the service (mobile) number	Nominal
9	TOTAL CALLS	This feature is derived from counting the total Calls made by each subscriber on a single day	Number

4.4 Data Preparation

Data preparation or data preprocessing is a crucial activity for data analysis and to build a model. Call detail records are not used directly for data mining, since the goal of data mining is to extract knowledge at the customer level, not at the level of individual calls. The call detail records associated with customers must be summarized into a single record that describes the customer's calling behavior. All relevant data for each subscriber was assembled in the form of columns and row dataset and it represents a unique subscriber with all the data related to this subscriber includes the choice of features is critical in order to obtain a useful description of the subscriber. Under preparation of the data like data cleaning, data integration and data reduction are done.

4.4.1 Data Cleaning

Data cleaning used for making sure that the data is free from different types of errors and in order to make the data complete. In this thesis first the data prepared by removing the records that had incomplete record and missing value under each column. Removing such kind of data set doesn't affect the entire dataset. Ethio telecom domain experts generate twelve attributes from different data base but some of them have zero and missing value .Only nine selected attributes (SMS, GPRS, call location number, call fee, calling number, called number, date and time, duration and total Number of call per day) are used for training, testing dataset and also for model building. Microsoft access 2013 and Microsoft excel 2013 used for data cleaning purpose.

4.4.2 Data Integration

Data integration is the core element to get full information about each caller number .Data integration method used for retrieving important attributes from different files and tables to make ready the data for experimentation. In this study different attributes generated from different databases and the data integration process took a lot of time of the thesis. Because of the reason that the data was very huge and after preprocessed by oracle database the data saved in different excel files. Finally by using excel data integration mechanisms the data is integrated and put together into a single excel file.

4.4.3 Data formatting

To apply data mining algorithms, before building the prediction model the first task is preparing the data in a file format that is acceptable by WEKA 3.8.1 data mining tool. The tool accepts file formats like comma delimited CSV and ARFF.

The original data extracted from oracle database saved by excel format, after applying data cleaning, reduction and integration the excel data changed into a comma delimited CSV file format. Then by using WEKA data mining tool the CSV file format saved into ARFF (attribute

relation file format) file extension. Finally the dataset in ARFF format is ready to process in WEKA tool by applying different data mining algorithms.

4.4.3.1 Attribute selection

In this thesis important attribute selection is done based on WEKA attribute selection technique .Data mining methods such as attribute selection and attribute relevance ranking help to identify the most important attributes and eliminate non relevant ones. Most machine learning algorithms are designed to learn in which are the most appropriate features to make a decisions. In classification algorithms we can choose the most important features to classify the problem correctly.

The importance of reducing the number of attributes not only speed up the learning process but also prevents most of the learning algorithms from getting fooled into generating an inferior model by the presence of many irrelevant or redundant attributes. In order to select the best attributes from WEKA tool, first we try to evaluate the attribute content based on previously made research recommendations.

4.5 Performance measures for Classification Algorithms

Confusion matrix

The confusion matrix is used to measure the performance of two class problems for a given dataset. True positive (TP) and True negative (TN) means that correctly classify instances as well as false positive (FP) and False negative (FN) means incorrectly classify instances. The following figure 8 shows the confusion matrix (correctly classified instances (TP, TN) and incorrectly classified instances (FP, FN)). [27]

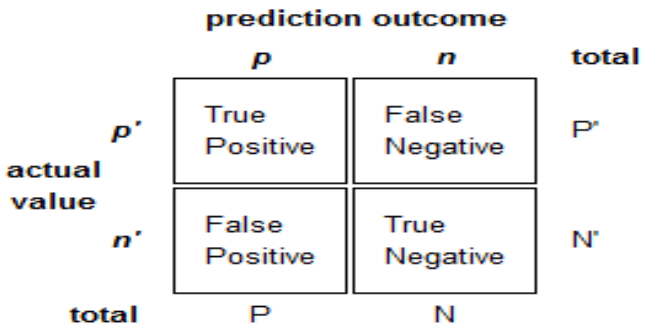

Figure 8: Confusion matrix performance measure

Total number of Instance = correctly classified instances +incorrectly classified instances.

Cost matrix

A cost matrix is similar to confusion matrix but minor difference is with finding the value of cost accuracy through misclassification error rate.

Misclassification error rate =1- accuracy

Recall

Recall is the ratio of modules correctly classified as fault prone to the number of entire fault modules.

Recall =TP/TP+FN

Precision

Precision is the ratio of modules correctly classified to the number of entire modules classified fault-prone. It is proportion of units correctly predicted as faulty.

Precision =TP/TP+FP

F-Measure

It is a combination of recall and precision .It is defined as harmonic mean of precision and recall

F-Measure =2*Recall*precision / Recall + precision

Accuracy

It is defined as the ratio of correctly classified instances to total number of instances.

Accuracy = TP+TN / TP+FP+TN+FN

TPR, TNR, FPR and FNR

One can calculate the value of true positive rate, true negative rate, false positive rate and false negative rate by methods shown here below.

TPR =TP/ TP+FN

TNR=TN/FP+TN

FPR=FN/TP+FN

FNR=FP/FP+TN

Chapter Five

5. Result and Discussion

In this thesis, the experimentation part is completed by using WEKA (Wakito environment knowledge analysis) data mining tool version 3.8.1. To identify the SIM box fraudulent numbers and their behavior different experiments are made by using 9 attributes (SMS, GPRS, call location number, call fee, calling number, called number, date and time, duration and total Number of call per day). During the experimentation important attributes are selected based on Weka selection mechanism. Such experiments are conducted using decision tree (J48), rule based (PART) and from neural network (multilayer perceptron) algorithms. Finally, by combining PART and J48 algorithms, a better prediction model is built. The rules generated from classification algorithms especially J48, PART and Hybrid algorithms have the highest performance to predict the fraudulent numbers. Multilayer perceptron have also a good accuracy but when we compare with the others three algorithms, it is a little bit inferior. Different parameters and techniques used during the experiment to get an optimal result. In this section, the result of this thesis will be discussed.

5.1 Model Building

To achieve the objective of this thesis, the hybrid algorithm has been used by taking the strength of J48 and PART algorithm. Four classification algorithms are used in this paper. The first is PART algorithms from rule based, the second one is J48 algorithms from tree based algorithm, thirdly multilayer perceptron form neural network is used and the last one is hybrid algorithm from the combination of PART and J48 algorithms, which are tree based and rule based algorithms. Based on sampled CDR Dataset, those algorithms are tested and compared with their prediction performance. The algorithms were evaluated based on processing time, confusion metrics, performance measures such that Precision, Recall, F-measure and Accuracy. Finally, the best prediction model were built by using PART algorithm for SIM box fraudulent number prediction because it have a better classification accuracy than J48, multilayer perceptron and hybrid algorithms.

5.1.1 WEKA Interface

For testing and model building WEKA data mining tool is used. The WEKA workbench provides four main ways: - Explorer, Experimenter, knowledge flow and Simple CLI.

WEKA explorer give the chance to play with the data and think about what transforms to apply to the dataset and what algorithms you want to run in experiments. [22]. The Explorer interface is divided into 5 different tabs:

- **Preprocess**: Load the dataset and manipulate the data into a form that is desired to work with.
- **Classify**: Select and run classification and regression algorithms to operate on the data.
- **Cluster**: Select and run clustering algorithms on the dataset.
- **Associate**: Run association algorithms to extract insights from the data.
- **Select Attributes**: Run attribute selection algorithms on the data to select those attributes that are relevant to the feature wanted to predict.
- **Visualize**: Visualize the relationship between attributes.

WEKA experimenter

This interface is used for designing experiments with selected algorithms and datasets, running experiments and analyzing the results. The tools for analyzing results are very powerful, allowing to consider and compare results that are statistically significant over multiple runs.

Knowledge Flow

Applied machine learning is a process and the knowledge flow interface allows to graphically design that process and run the designs that we create. This includes the loading and transforming of input data, running of algorithms and the presentation of results. It's a powerful interface and metaphor for solving complex problems graphically.

Simple CLI

It provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface. Figure 9 below shows the graphical user interface of WEKA data mining tool. (source: [22]).

Figure 9: WEKA graphical user Interface

WEKA Test Options

The result of applying the chosen classifier will be tested according to the options that are set by clicking in the Test options box. There are four test modes:

1. Use training set- The classifier is evaluated on how well it predicts the class of the instances it was trained on.

2. Supplied test set- The classifier is evaluated on how well it predicts the class of a set of instances loaded from a file. Clicking the Set... button brings up a dialog allowing you to choose the file to test on.

3. Cross-validation- The classifier is evaluated by cross-validation, using the number of folds that are entered in the Folds text field. Cross-validation (CV) method used in order to validate the predicted model. CV test basically divide the training data into a number of partitions or folds. The classifier is evaluated by accuracy on one phase after learned from other one. This process is repeated until all partitions have been used for evaluation [13]. The most common types are 10-fold, n-fold and bootstrap result obtained into a single estimation.

4. Percentage split-The classifier is evaluated on how well it predicts a certain percentage of the data which is held out for testing. The amount of data held out depends on the value entered in the % field.

5.2 J48 algorithm

Decision tree is most popular algorithm used for fraud detection and prediction. Decision trees represent a supervised approach to classification. A decision tree is a simple structure where non-terminal nodes represent tests on one or more attributes and terminal nodes reflect decision outcome s. J.R. Quinlan has popularized the decision tree approach with his research. J48 is one of the most common decision tree algorithms that are used today to implement classification technique using WEKA. The latest public domain implementation of Quinlan's model is C4.5. The Weka classifier package has its own version of C4.5 known as J48. Nine call detail attributes (SMS, GPRS, call location number, call fee, calling number, called number, date and time, duration and total Number of call per day) are used in this thesis, WEKA data mining tool have mechanism to select most important attributes based on that the analysis is done.

As listed Table 2 for J48 algorithm with 10-fold cross validation scored an accuracy of 99.47%. This result shows that out of 9030 training datasets 8983(99.47%) instances are correctly classified, while 47 (0.53%)of the instances are incorrectly classified. Table 3 indicates the accuracy, True positive and false negative rate, precision, recall, f-measure values that uses to measure the performance of J48 classification model and values of time taken to build J48 classification model.

Table 2: J48 Confusion matrix and details of classification model

Confusion Matrix			No of Instances	No of Leaves	Size of the tree	Correctly classified instances	Incorrectly Classified instances	Test Mode
a	b	Classified as						
6579	18	a= Non fraudulent	9030	16	31	8983	47	Cross-validation
29	2404	b=fraudulent						

Table 3: J48 classification accuracy

J48 Algorithm Performance							
Class	TP Rate	FP Rate	Precision	Recall	F-measure	Accuracy	Time
NonFR	0.997	0.012	0.996	0.997	0.996	99.47%	1.42 sec
FRAUD	0.988	0.003	0.993	0.988	0.990		

Figure 10 and 11 shows that the ROC curve for both training set and test set for fraudulent subscribers and non-fraudulent subscribers. On the x axis plots the false positive rate and on the y axis plots true positive rate. ROC curves with J48 for class value of non-fraudulent subscriber shows (0.012 False positive rate value with 0.997 True positive rate value). And for fraudulent subscriber the ROC curve shows (0.003 False positive rate value with 0.988 True positive rate value). The closer the curve follows the left hand border and then the top border of the ROC spaces shows the more accurate. The closer curve comes to the 45-degree diagonal of the ROC space shows the less accurate test.

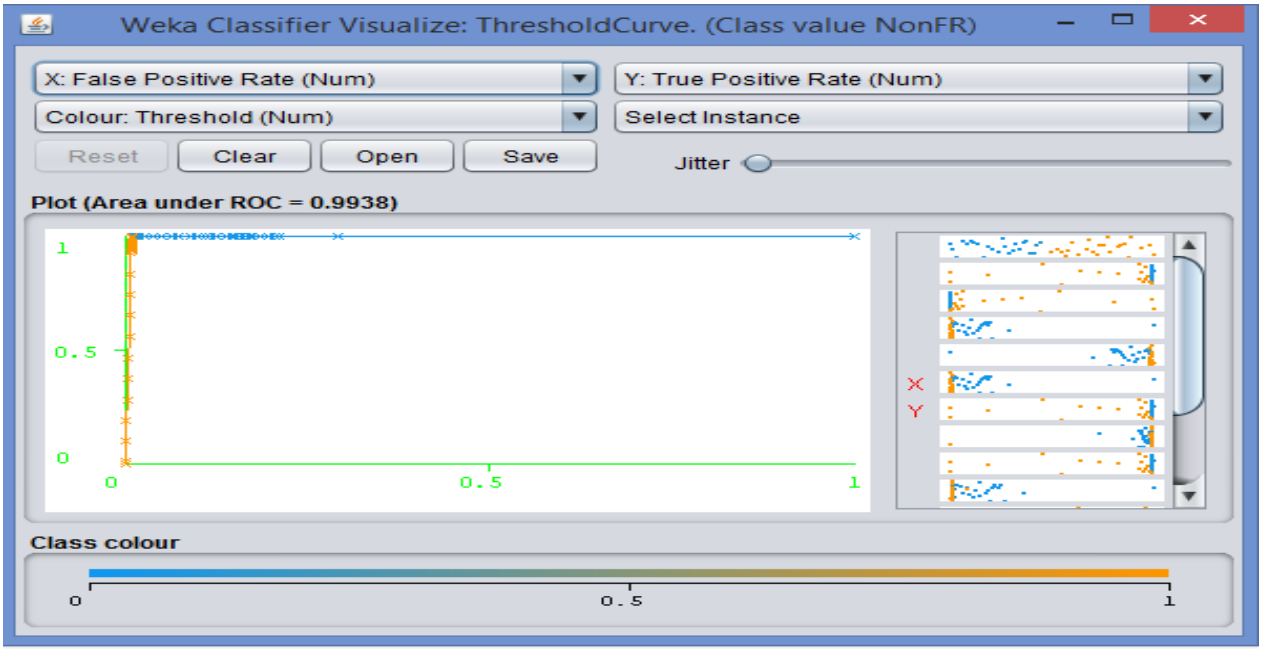

Figure 10: ROC curve with J48 for class value NonFR (Non fraudulent subscriber)

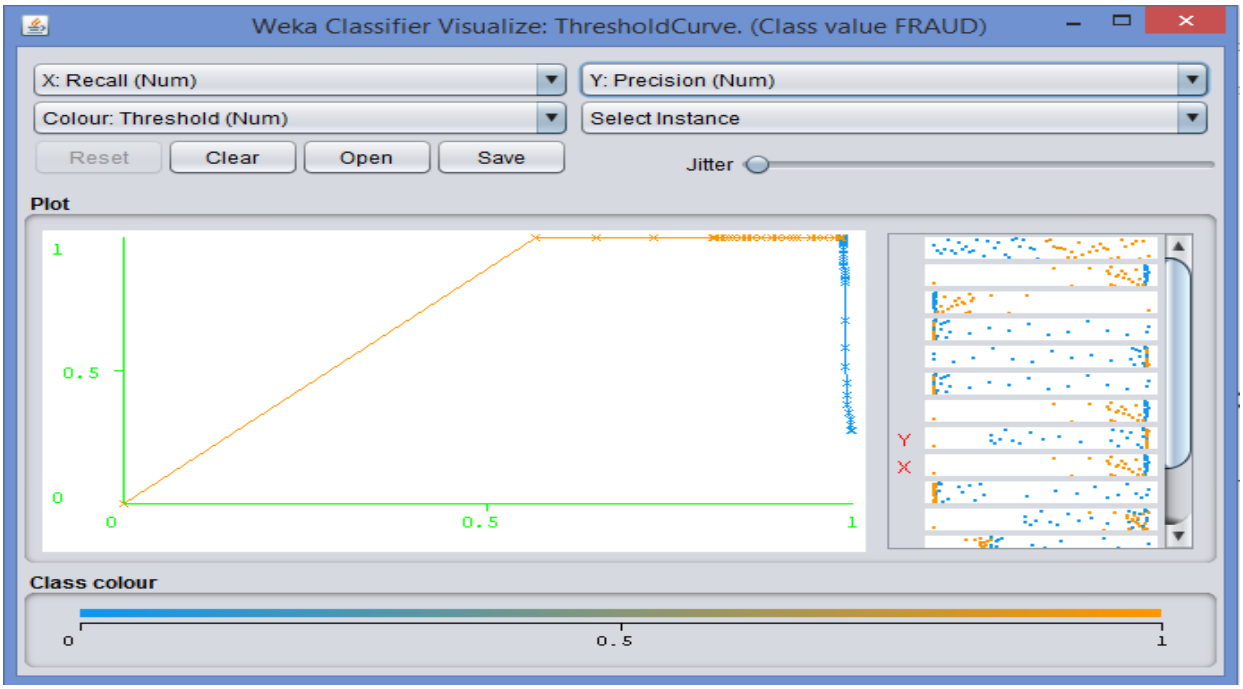

Figure 11: ROC curve with J48 for class value FRAUD (fraudulent Subscriber)

5.3 PART Algorithm

The second classification algorithm used in this thesis is PART rule induction algorithm. PART is a rule generator that uses J48 to generate pruned decision trees from which rules are extracted. And PART rule induction algorithm have the ability to produce accurate classification rules that help to predict fraudulent numbers and their behavior. Input attributes can be categorical and numerical. In this Thesis, ten attributes and 9030 datasets are used to test part rule induction algorithm with cross- validation test mode.

Table 4 shows PART algorithm with 10-fold cross validation scored an accuracy of 99.4906% with 17 rules. This result shows that out of 9030 training datasets 8984 (99.4906%) instances are correctly classified, while 46 (0.5094 %)of the instances are incorrectly classified. Table 5 indicates PART classification model accuracy, True positive and false negative rate, precision, recall, f-measure values of both fraudulent and non-fraudulent classes, to measure the performance of PART classification model and values of time taken to build PART classification model. PART

algorithm have 0.0206% better classification accuracy performance as compared to the previous J48 algorithm.

Table 4: PART confusion matrix and details of classification model

Confusion Matrix			No of Instances	No of Rules	Correctly classified instances	Incorrectly Classified instances	Test Mode
a	b	Classified as					
6578	19	a= Non fraudulent	9030	17	8984	46	Cross-validation
27	2406	b=fraudulent					

Table 5: PART classification accuracy

PART Algorithm Performance							
Class	TP Rate	FP Rate	Precision	Recall	F-measure	Accuracy	Time
NonFR	0.997	0.011	0.996	0.997	0.997	99.4906%	1.14 sec
FRAUD	0.989	0.003	0.992	0.989	0.991		

Figure 12 and 13 shows the ROC curve for both data set of fraudulent subscribers and non-fraudulent subscribers result of PART classification model. An ROC curve is thus a two-dimensional graph that visually depicts the relative trade-offs between (false positives rates) and (true positives rates). The following ROC curves plots with the area under ROC value of 0.9952 for both fraudulent and non-fraudulent classes. The closer the curve follows the left hand border

and then the top border of the ROC spaces shows the more accurate. The closer curve comes to the 45-degree diagonal of the ROC space shows the less accurate test.

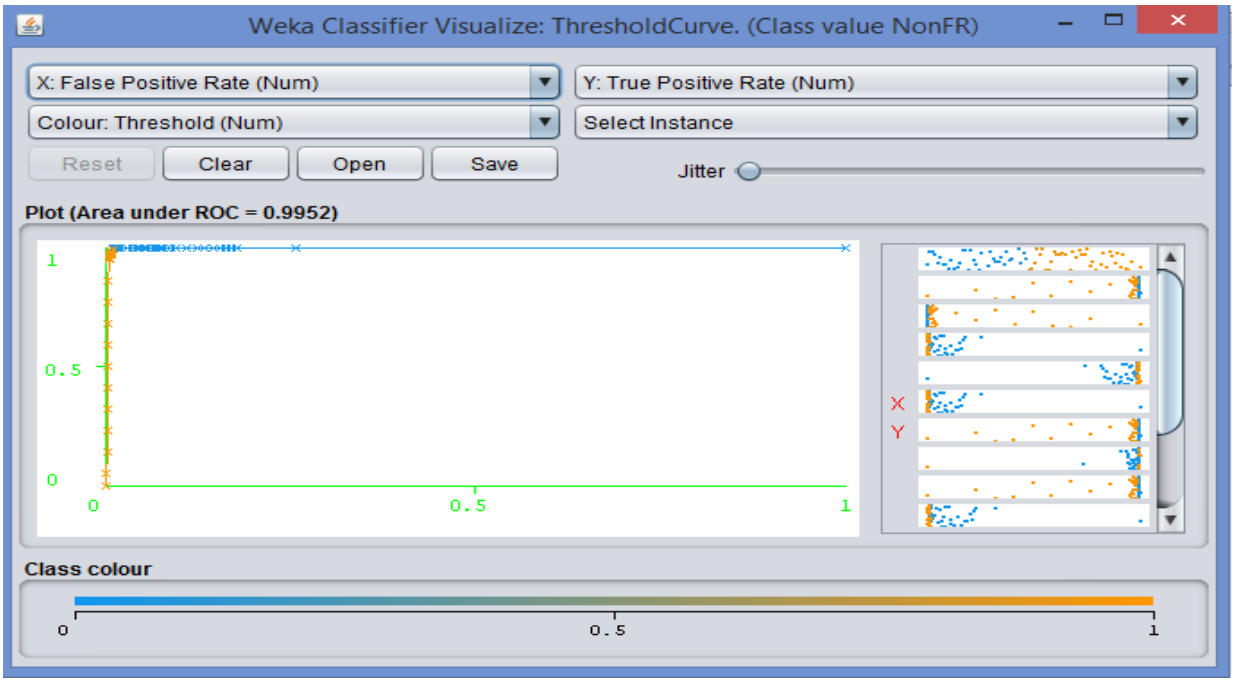

Figure 12: ROC curve with PART for class value NonFR (Non fraudulent subscriber)

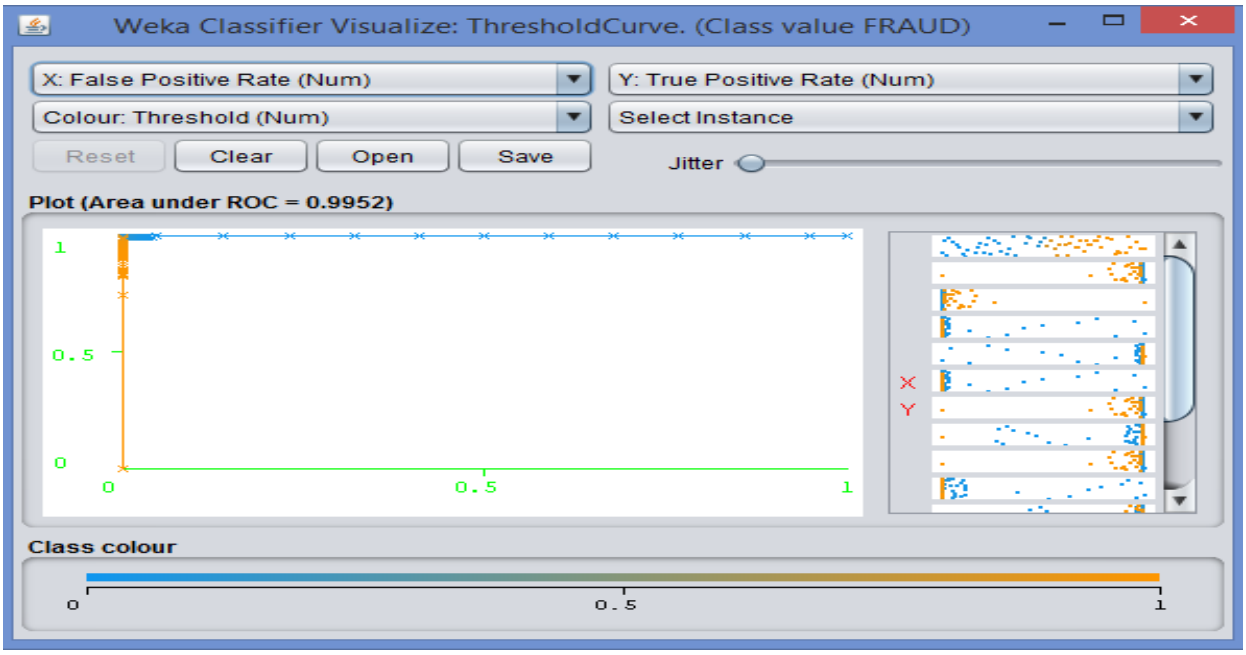

Figure 13: ROC curve with PART for class value FRAUD (fraudulent Subscriber)

5.4 MLP (Multi-layer perceptron) Algorithm

Multi-layer networks use a variety of learning techniques, the most popular being back-propagation. Here the output values are compared with the correct answer to compute the value of some predefined error-function. By various techniques the error is then feedback through the network. Using this information, the algorithm adjusts the weights of each connection in order to reduce the value of the error function by small amount. After repeating this process for a sufficiently large number of training cycles the network will usually converge to some state where the error of the calculations is small. In this case one says that the network has learned a certain target function.

Table 6 shows multilayer perceptron algorithm with 10-fold cross validation scored an accuracy of 98.5491%. This result shows that out of 9029 training datasets 8898 (98.5491%) instances are correctly classified, while 131 (1.4509 %)of the instances are incorrectly classified, when compared with PART and J48 classification model, which is less accurate. The time to build the model is 7.26 sec which means it takes more time. Generally, Table 7 indicates multilayer perceptron algorithm accuracy, True positive and false negative rate, precision, recall, f-measure values of both fraudulent and non-fraudulent classes, it uses to measure the performance of Multilayer perceptron classification algorithm. The previous two Algorithms J48 and PART algorithms are better than multilayer perceptron algorithm interims of accuracy and processing time. And with different performance metrics multilayer perceptron algorithm a little inferior than the two. Multilayer perceptron algorithm is selected from artificial neural network and one of its drawback is processing time. When multilayer perceptron algorithm is compared with other algorithms, it takes more processing time.

Table 6: Multilayer perceptron Confusion matrix and details of classification model.

Confusion Matrix			No of Instances	Correctly classified instances	Incorrectly Classified instances	Test Mode
a	b	Classified as				
6531	65	a= Non fraudulent	9029	8898	131	Cross-validation
66	2367	b=fraudulent				

Table 7: Multilayer perceptron classification accuracy

Multilayer perceptron Algorithm Performance							
Class	TP Rate	FP Rate	Precision	Recall	F-measure	Accuracy	Time
NonFR	0.990	0.027	0.990	0.990	0.990	98.5491%	7.26 sec
FRAUD	0.973	0.010	0.973	0.973	0.973		

Figure 14 and 15 shows the ROC curve for data set of both fraudulent subscribers and non-fraudulent subscribers result of Multilayer perceptron classification model. The ROC curves below with Multilayer perceptron for class value of both fraudulent and non-fraudulent plot with the area value 0.9877. The closer it follows the left hand border and then the top border of the ROC spaces, the more accurate. The closer the curve comes to the 45-degree diagonal side of the ROC space, the less accurate the test.

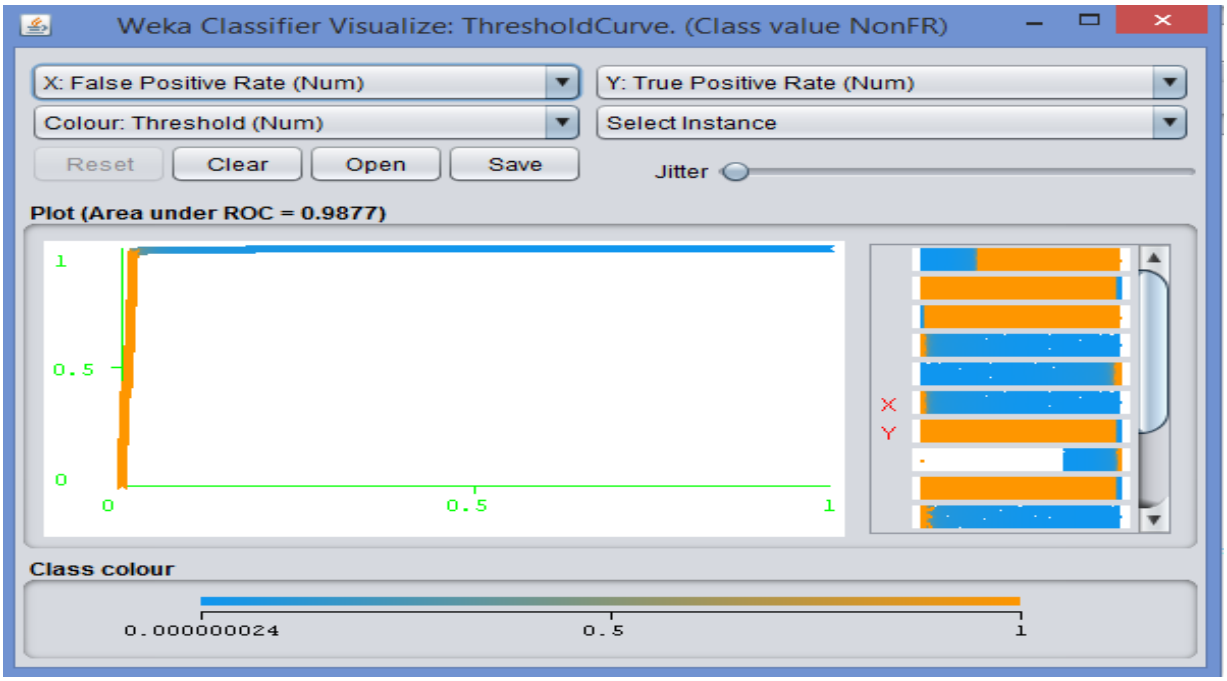

Figure 14: ROC curve with Multilayer perceptron for class value NonFR (Non fraudulent subscriber)

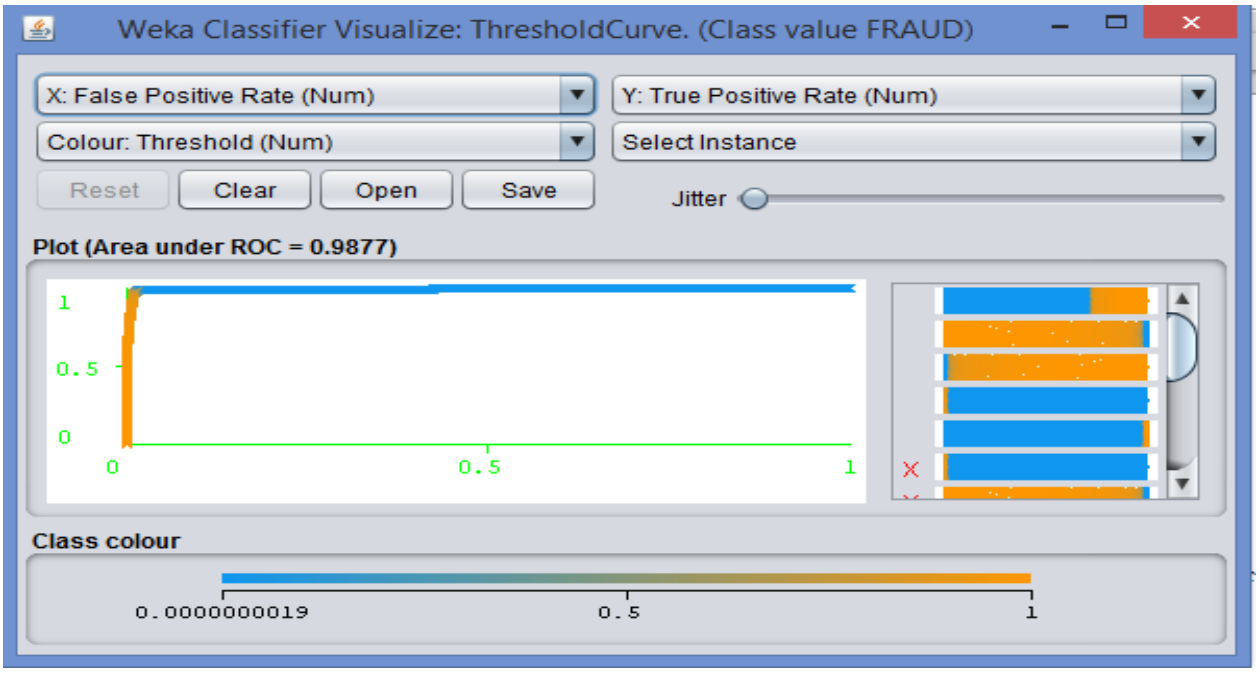

Figure 15: ROC curve with Multilayer perceptron for class value FRAUD (fraudulent subscriber)

5.5 Hybrid Algorithm

In this study, hybrid algorithm is used to identify SIM Box fraudulent numbers with better accuracy by using PART algorithm from rule based induction and J48 from decision tree. J48 and PART algorithms are capable of handling datasets that may have errors and handling datasets that may have missing values. Both algorithms can quickly express complex alternatives clearly and have a capability to generate best splitting decision trees and rules. Results of empirical comparisons of existing algorithms illustrate that each algorithm has certain selective superiority. It is best for some, but not all tasks. [25] [26] researches recommend those two algorithms in fraud detection like telecommunication fraud especially for (SIM box fraud detection) and credit card fraud detection. The two algorithms hybrid by using the Average combination rule and voting method. Voting is the most common method used to combine classifiers, the strategy is motivated by the Bayesian learning theory which stipulates that in order to maximize the predictive accuracy, instead of using just a single learning model. When we use hybrid algorithm, the learning algorithm run several times, each time using a different distribution of the training examples, then the prediction accuracy will increase. But it has also a disadvantage because when we combine multiple algorithms, it may occur overfeeding problem. Many research papers have stated that when we combine multiple algorithms or classifiers in order to classify instances from dataset, at that time there is a risk to have multiple classification algorithm problems in one task.

Table 8 shows Hybrid (PART and J48) algorithm with 10-fold cross validation scored an accuracy of 99.4795%. This result shows that out of 9030 training datasets 8983 (99.479%) instances are correctly classified, while 49 (0.5205 %)of the instances are incorrectly classified. When we compare with, Multilayer and J48 classification model which is more accurate. But the PART classification model have more accuracy form this hybrid algorithm. The time to build the model is 1.16 sec which means it takes good time. Table 9 indicates hybrid classification model accuracy, True positive and false negative rate, precision, recall, f-measure values of both fraudulent and non-fraudulent classes, it uses to measure the performance of PART classification model. Hybrid algorithm have a better performance than multilayer and J48 algorithm but it is a little inferior to PART algorithm.

Table 8: Hybrid algorithm Confusion matrix and details of classification model

Confusion Matrix			No of Instances	Correctly classified instances	Incorrectly Classified instances	Test Mode
a	b	Classified as				
6580	17	a= Non fraudulent	9030	8983	47	Cross-validation
30	2403	b=fraudulent				

Table 9: Hybrid algorithm confusion Matrix and classification accuracy

Hybrid Algorithm Performance							
Class	TP Rate	FP Rate	Precision	Recall	F-measure	Accuracy	Time
NonFR	0.997	0.012	0.995	0.997	0.996	99.4795%	1.16 sec
FRAUD	0.988	0.003	0.993	0.988	0.990		

Figure 16 and 17 shows that the ROC curve for both training set and test set for fraudulent subscribers and non-fraudulent subscribers result of hybrid (J48 and PART) classification model. On the x axis plots the false positive rate and on the y axis plots true positive rate. ROC curves with Hybrid algorithm for class value of non-fraudulent subscriber shows (0.012 False positive rate value with 0.997 True positive rate value). And for fraudulent subscriber the ROC curve shows (0.003 False positive rate value with 0.988 True positive rate value). The ROC curve plot on the area under ROC value of 0. 9959.The closer follows the left hand border and then the top border

of the ROC spaces shows the more accurate. The closer curve comes to the 45-degree diagonal of the ROC space shows the less accurate test.

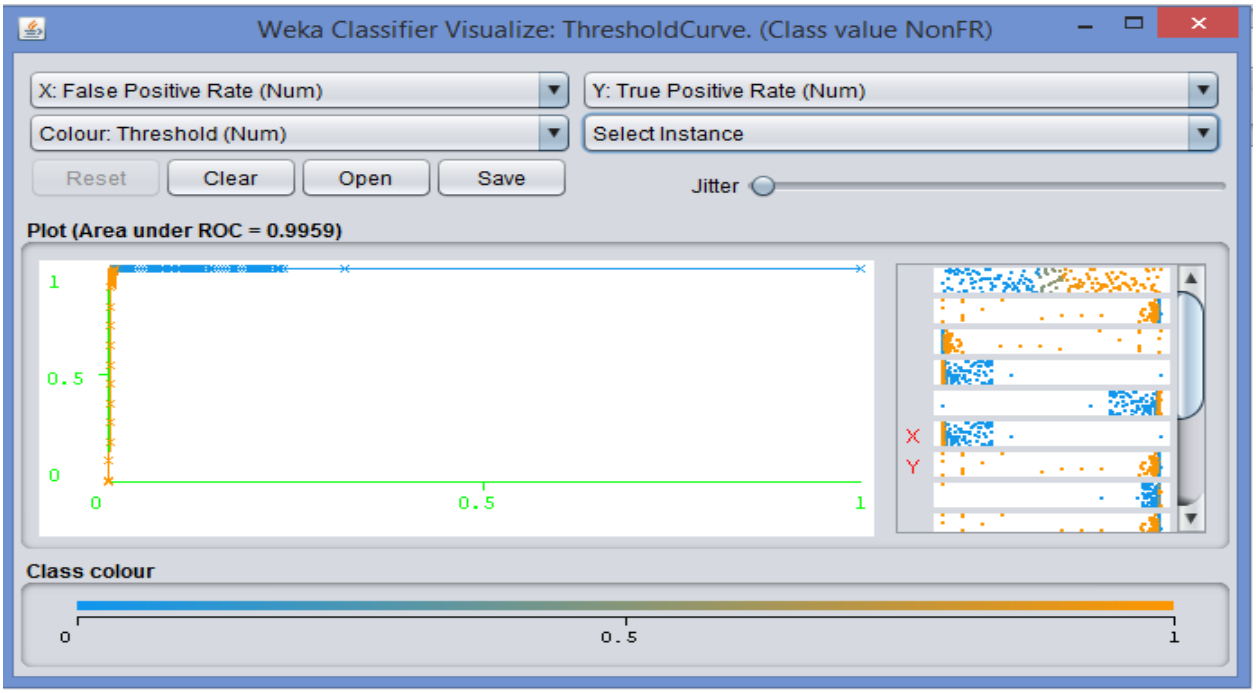

Figure 16: ROC curve with hybrid algorithm for class value NonFR (Non fraudulent subscriber)

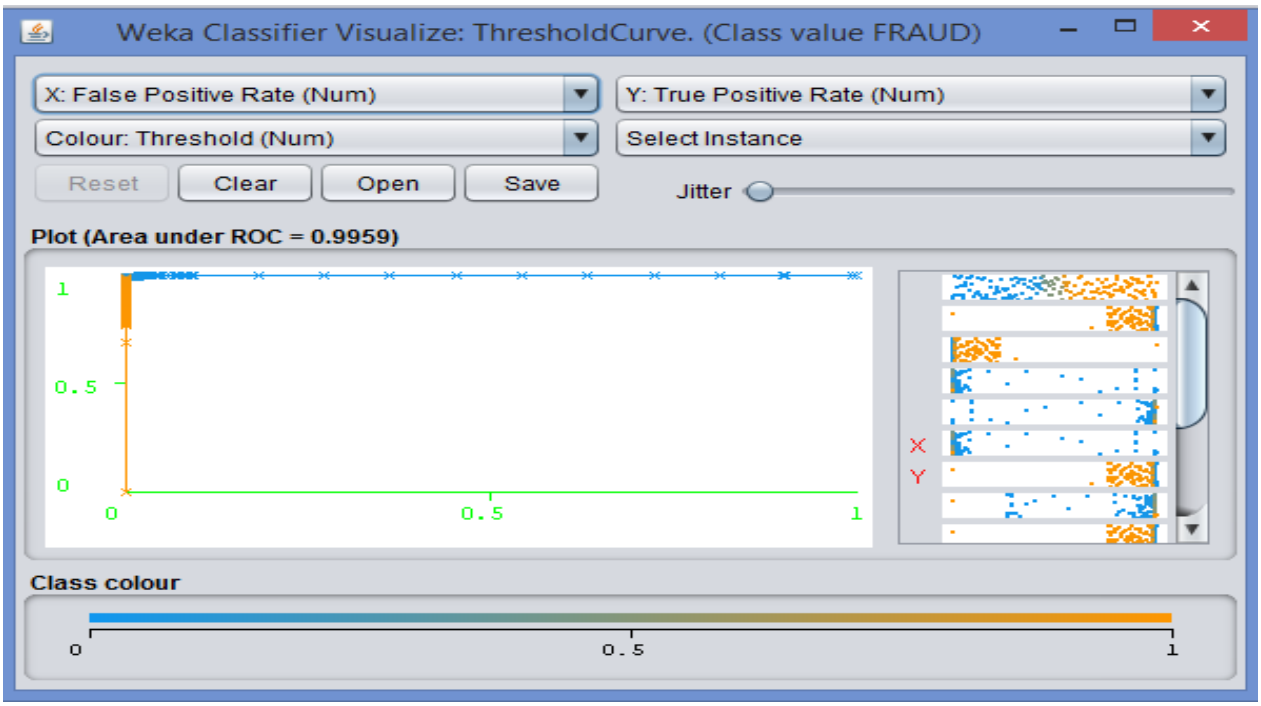

Figure 17: ROC curve with hybrid algorithm for class value FRAUD (fraudulent subscriber)

Table 10 shows that four algorithms performances used in this thesis and detail value of performance metrics like false positive rate, True positive rate, precision, recall, F-measure, Roc area. PART rule based induction algorithm have a better accuracy from J48, MLP and hybrid algorithm. And hybrid algorithm has a better accuracy than J48 and multilayer algorithm and a little inferior than PART algorithm. J48 is also have a better accuracy than multilayer perceptron and inferior than PART and Hybrid algorithm. Multilayer perceptron has lower accuracy when we compare to the other three algorithms.

Table 10: Summery of J48, PART, multilayer perceptron and hybrid algorithms performance

Features	J48	PART	MLP	Hybrid
Accuracy	99.47%	99.4906 %	98.5491 %	99.4795 %
Time to build	1.42 sec	1.14 sec	7.26 sec	1.16sec
TP	0.995	0.995	0.985	0.995
FP	0.009	0.009	0.022	0.010
Precision	0.995	0.995	0.985	0.995
Recall	0.995	0.995	0.985	0.995
F-Measure	0.995	0.995	0.985	0.995
Roc Area	0.994	0.995	0.991	0.995

Chapter Six

6. Conclusion and Future Direction

The main focus of this thesis was to come up with a set of features that can be used to identify calls originating from SIM box devices and an algorithm that classify subscribers with high accuracy and less processing time. The algorithms used to detect and predict SIM box fraudulent numbers are PART, J48, multilayer perceptron and hybrid algorithms. By using voting combination mechanism and by taking the strength of J48 decision tree and PART from rule based classification algorithms built hybrid algorithm. Classification algorithms (PART, J48, multilayer perceptron and hybrid algorithms) are efficient to solve the problem of SIM box fraud and Data mining tools and techniques were used. Nine features extracted from CDR data utilized to build a model that can be used to distinguish between legitimate and SIM box fraud calls. The feature includes calling number, called number, call fee, call duration, date and time, location number, total number of calls per day, SMS and GPRS call detail records.

PART classification from rule based induction resulted 99.4906% accuracy that performs better than multilayer perceptron, J48 and Hybrid algorithm (J48 and PART). The hybrid algorithm performs 99.4795% accuracy, it performs better than J48 and multilayer perceptron and a little inferior than PART.

Generally, the rules and techniques generated from the classification algorithms can help ethio telecom or other telecommunication companies to identify the SIM box fraudulent calls behavior. This thesis can help for future work on SIM box fraud detection and prevention area.

6.1 Recommendations

Now a days ethio telecom start to register IMEI (international mobile equipment identity) number. If the company can consider and include (IMEI) in Call detail records, it helps to identify mobile numbers under a specific IMEI, take some kind of action on the devices. Different researchers used this attribute of CDR data and it makes easy to identify the fraudulent numbers when we have international mobile equipment identity number. For fraudsters losing the SIM cards is not a big loss as such but the SIM box device. Once the device is disabled from the network, it is no more useful in that country.

This thesis is limited on two months sample CDR data and use only pre-paid mobile numbers. One can conduct similar researches by using more sample CDR data.

It is also possible to conduct similar researches on the other types of Telecommunication fraud in ethio telecom case especially subscription fraud and call and SMS spamming fraud.

Reference

[1] Shawe-Taylor, J., Howker, K., & Burge, P. (1999). Detection of fraud in mobile Telecommunications. Information Security Technical Report, 4(1), 16-28.

[2]. Bolton, R.J., & Hand, D.J. (2002). Statistical fraud detection: A review. Statistical Science,

17(3), 235-249.

[3] Fayemiwo Michael and Olasoji Babatunde (2014). Fraud detection in mobile telecommunication. Department of Mathematical Sciences, College of Natural and Applied Science, Oduduwa University.

[4] Igor Ruiz-Agundez, Yoseba K. Penya, and Pablo Garcia Bringas. Fraud Detection for Voice over IP Services on Next-Generation Networks. University of Deusto Bilbao, Basque Country.

[5] Ilona Murynets, Michael Zabarankin , Roger Piqueras Jover and Adam Panagia., Analysis and Detection of SIMbox Fraud in Mobility Networks. AT&T Security Research Center, New York, NY.

[6] Walid Moudani and Fadi Chakik, Fraud Detection in Mobile Telecommunication. Lecture Notes on Software Engineering, Vol. 1, No. 1, February 2013.

[7] Tesfay (2013), Predictive Model to Subscription Fraud Detection using Data Mining Techniques , Addis Abeba University, Department of Information System.

[8] Daya Gupta, Payal Pahwa and Rajiv Arora (2014), An Analysis of Telecommunication Fraud using Outlier Detection Model based on Similar Coefficient Sum. Bhagwan Parshuram Institute of Technology, New Delhi, India.

[9] Krenker A., Volk, M., Sedlar, U., Bester, J. and Kos, A. 2009. Bidirectional Artificial Neural Networks for Mobile-phone Fraud Detection. Etri Journal. 31(1): 92–94.

[10] Farvaresh, H. and Sepehri, M. M. 2011. A Data Mining Framework for Detecting Subscription Fraud in Telecommunication. Engineering Applications of Artificial Intelligence. 24(1): 182–194.

[11] Abidogun, O. A. (2005). Data mining, fraud detection and mobile telecommunications: Call pattern analysis with unsupervised neural networks. University of the Western Cape.

[12] Asfaw, N. (2006). Challenges Facing International Telecom Business and the Way Forward, Ethiopian Telecommunication Corporation's Perspectives. Masters Thesis (Telecom MBA), College of Telecommunication and Information Technology, Management Department.

[13] Gebremeskal, G. (2006). Data Mining Application in Supporting Fraud Detection on Ethio-Mobile Services. Masters Thesis, AAU, Faculty of Informatics, Department of Information Science, 67.

[14] Jember, G. (2005). Data Mining Application in Supporting Fraud Detection on Mobile Communication: The Case of Ethio-Mobile. Master's Thesis, AAU, Informatics Faculty, Department of Information Science, 98.

[15] Melaku, G. (2009). Application of Data Mining Techniques to Customer Relationship Management (CRM): The case of Ethiopian Telecommunications Corporation's (ETC) Code Division Multiple Access (CDMA) Telephone Service. AAU, Faculty of Informatics, Department of Information Science.

[16] Yigzaw, M., Hill, S., Banser, A., & Lessa, L. (2010). Using Data Mining to Combat Infrastructure Inefficiencies: The Case of Predicting Non-payment for Ethiopian Telecom. Paper presented at the 2010 AAAI Spring Symposium Series

[17] Abdikarim Hussein and Roselina Sallehuddin. (2014) classification of SIM Box fraud Detection using support vector machine and Artfical neural network. University of Technology Malaysia.

[18] Bulent Kusaksizogiu. (2006). Fraud Detection in Mobile communication networks using Data mining, University of Bahcesehir Department of Computer Engineering.

[19] Mhd Redwan ., (2016).Comparing Data Mining classification algorithms in Detection of SIM Box Fraud , ST cloud state University .

[20] N2B Risk Management, http://www.zira.com.ba/products/risk-managemet/n2b-fraud-management-system/sim-box.

[21] Shearer, C, 2000, The CRISP-DM model: the new blueprint for data mining. Journal of Data Warehousing l5 (4), 13 –19.

[22] Sumathi, S., & Sivanandam, S. (2006). Introduction to data mining and its applications (Vol. 29): Springer.

[23] Two-Crows. (1999). Introduction to Data Mining and Knowledge Discovery (3rd edition ed.): Two Crows Corporation.

[24] Barson P., Field S., Davey N., McAskie G. and Frank R. 1996. The Detection of Fraud in Mobile Phone Networks. Neural Network World. 6(4): 477–484.

[25] Ganji,V.R. & Mannem,S.N (2012), Credit card fraud detection using anti-k nearest neighbour algorithm, International Journal on Computer Science and Engineering,1035-1039.

[26] Sahin, Y., Bulkan, S. & Duman, E. (2013). A cost sensitive decision tree approach for fraud detection Expert System with Application 40, 5916-5923.

[27] Pandya,S.S (2013). Mobile Fraud Detection, International Journal of IT, Engineering and Applied Sciences Research, 2, 15-18.

Appendixes

Best Rules from J48 Decision Tree Algorithm

Attributes: 9 (A_NBR, B_NBR, SMS, GPRS, DURATION, CALL_FEE, CELL_A, TOTAL CALL, CLASS).

Test mode: 10-fold cross-validation

Rule 1: TOTAL CALL <= 51 AND A_NBR <= 251983xxxxxx AND A_NBR <= 251983xxxxxx: NonFR (445.0/6.0)

Rule 2: A_NBR > 2519831xxxx AND A_NBR <= 251983xxxx: NonFR (14.0)

Rule 3: A_NBR > 251983xxxx: FRAUD (14.0/1.0)

Rule 4: A_NBR> 251970xxxx AND SMS>=0 AND TOTAL CALL <51 : NonFR (6075.0/19.0)

Rule 5: TOTAL CALL > 51 AND CELL_A <= 636011500614216 AND SMS <= 0 AND TOTAL CALL <= 65 AND A_NBR <= 251987xxxxxx: FRAUD (257.0/6.0)

Rule 6: A_NBR > 251987xxxxxx AND A_NBR <= 251988xxxxxx AND A_NBR <= 251987xxxxxx: NonFR (6.0)

Rule 7: A_NBR > 251987xxxxxx: FRAUD (12.0)

Rule 8: A_NBR > 251988xxxxxx: NonFR (7.0)

Rule 9: TOTAL CALL > 65: FRAUD (1819.0)

Rule 10: SMS > 0 AND A_NBR <= 251973xxxxxx: FRAUD (23.0)

Rule 11: A_NBR > 251973xxxxxx AND DURATION <= 41: FRAUD (6.0)

Rule 12: DURATION > 41: NonFR (6.0)

Rule 13: CELL_A > 636011500614216 AND CELL_A <= 636012412042676 AND A_NBR <= 251970xxxxxx: FRAUD (104.0/1.0)

Rule 14: A_NBR > 251970xxxxxx AND CALL_FEE <= 1380: NonFR (58.0/1.0)

Rule 15:CALL_FEE > 1380: FRAUD (9.0/1.0)

Rule 16: CELL_A > 636012412042676: FRAUD (175.0/3.0)

Best Rules from PART Rule induction algorithm

Attributes: 9 (A_NBR, B_NBR, SMS, GPRS, DURATION, CALL_FEE, CELL_A, TOTAL CALL, CLASS).

Test mode: 10-fold cross-validation

Rule 1:- TOTAL CALL <= 51 AND TOTAL CALL <= 49 AND B_NBR <= 251988xxxxxx AND

A_NBR > 251983xxxxxx: NonFR (5940.0/5.0)

Rule 2:-TOTAL CALL > 51 AND CELL_A <= 636011500614216 AND SMS <= 0 AND TOTAL CALL > 65: FRAUD (1819.0)

Rule 3:-TOTAL CALL <= 51 AND A_NBR <= 251983xxxxxx AND GPRS > 0 AND B_NBR <= 251972xxxxxx: NonFR (314.0)

Rule 4:-TOTAL CALL <= 51 AND CELL_A > 636011500211072 AND GPRS <= 0 AND TOTAL CALL <= 50: NonFR (80.0)

Rule 5:-TOTAL CALL > 51 AND A_NBR <= 251970xxxxxx: FRAUD (258.0/2.0)

Rule 6:-A_NBR <= 251970xxxxxx AND B_NBR > 251915xxxxxx AND SMS > 0 AND GPRS > 0: NonFR (141.0/13.0)

Rule 7:-A_NBR <= 251970xxxxxx AND A_NBR > 251918xxxxxx: NonFR (45.0)

Rule 8:-TOTAL CALL > 51 AND A_NBR <= 251972xxxxxx: FRAUD (97.0)

Rule 9:-TOTAL CALL > 51 AND CELL_A > 636012234420167: FRAUD (92.0/3.0)

Rule 10:-CELL_A > 636011600518766 AND CELL_A <= 636012434840959: NonFR (34.0)

Rule 11:-CELL_A > 636011301019381 AND A_NBR <= 251987xxxxxx AND TOTAL CALL > 51 AND TOTAL CALL <= 213 AND A_NBR > 251974xxxxxx: FRAUD (72.0)

Rule 12:-CELL_A > 636011500210827: FRAUD (56.0/5.0)

Rule 13:-CELL_A <= 636011400110195: NonFR (17.0)

Rule 14:-CELL_A > 636011400317833 AND TOTAL CALL <= 219 AND SMS > 0: NonFR (10.0)

Rule 15:-A_NBR <= 251988xxxxxx AND CALL_FEE <= 9200 AND TOTAL CALL > 53: FRAUD (29.0/2.0)

Rule 16:-SMS <= 0 AND CELL_A <= 636011400718091: NonFR (18.0)

Rule 17:-B_NBR <= 251941xxxxxx: FRAUD (4.0)

Best rules from hybrid Algorithms

Attributes: 8 (A_NBR, SMS, GPRS, DURATION, CALL_FEE, CELL_ATOTAL CALL, CLASS)

Rule 1: TOTAL CALL <= 51 AND TOTAL CALL <= 49 ANDA_NBR > 251983xxxxxx: NonFR (6038.0/14.0)

Rule 2:A_NBR <= 251983xxxxxx AND GPRS > 0: NonFR (389.0/4.0)

Rule 3:A_NBR <= 251983xxxxxx AND CELL_A > 636011400519063: NonFR (58.0)

Rule 4: TOTAL CALL > 63 AND CELL_A <= 636011802437803 AND TOTAL CALL <= 207 AND A_NBR <= 251988xxxxxx: FRAUD (1837.0)

Rule 5: TOTAL CALL > 51 AND A_NBR <= 251970xxxxxx: FRAUD (228.0/2.0)

Rule 6: A_NBR > 251970xxxxxx AND TOTAL CALL > 51 AND A_NBR > 251972xxxxxx AND A_NBR > 251973xxxxxx AND SMS <= 0 AND TOTAL CALL > 65: FRAUD (96.0)

Rule 7: A_NBR <= 251970xxxxxx AND SMS <= 0: NonFR (42.0)

Rule 8: TOTAL CALL > 51 AND A_NBR <= 251972xxxxxx: FRAUD (130.0/4.0)

Rule 9: A_NBR > 251973xxxxxx AND A_NBR <= 251988xxxxxx AND TOTAL CALL <= 213 AND TOTAL CALL > 51 AND CALL_FEE <= 5520 AND TOTAL CALL <= 62: FRAUD (79.0)

Rule 10: CALL_FEE <= 800 AND A_NBR <= 251974xxxxxx: NonFR (43.0/1.0)

Rule 11: A_NBR > 251987xxxxxx: NonFR (16.0/1.0) CELL_A > 636011301019381 AND GPRS <= 0 AND DURATION <= 46: FRAUD (15.0)

Rule 12: CELL_A > 636011301019381 AND TOTAL CALL <= 139 AND SMS > 0: FRAUD (13.0) SMS > 0: NonFR (15.0)

Rule 13: TOTAL CALL > 50 AND CELL_A <= 636011500614213: FRAUD (12.0)

Rule 14: TOTAL CALL <= 50: NonFR (6.0)

Rule 15: GPRS <= 0 AND A_NBR <= 251979xxxxxx AND CELL_A <= 636011500614216 AND DURATION > 385: FRAUD (3.0)

Rule 16: GPRS <= 0 AND DURATION <= 180: FRAUD (5.0/1.0)

Rule 17: TOTAL CALL <= 51, A_NBR <= 251983xxxxxx, A_NBR <= 251983xxxxxx: NonFR (445.0/6.0)

Rule 18: A_NBR > 251983xxxxxx, A_NBR <= 251983xxxxxx: NonFR (15.0), A_NBR > 251983xxxxxx: FRAUD (13.0), A_NBR > 251983xxxxxx: NonFR (6075.0/19.0)

Rule 19: TOTAL CALL > 51, CELL_A <= 636011500614216, SMS <= 0, TOTAL CALL <= 65, A_NBR <= 251987xxxxxx: FRAUD (257.0/6.0)

Rule 20:A_NBR > 251987xxxxxx A_NBR <= 251988xxxxxx A_NBR <= 251987xxxxxx: NonFR (6.0)

Rule 21: A_NBR > 251987xxxxxx: FRAUD (12.0), A_NBR > 251988xxxxxx: NonFR (7.0), TOTAL CALL > 65: FRAUD (1819.0)

Rule 22: SMS > 0, A_NBR <= 251973xxxxxx: FRAUD (23.0), A_NBR > 251973xxxxxx, DURATION <= 41: NonFR (6.0)

Rule 23: CELL_A > 636011500614216, CELL_A <= 636012412042676, A_NBR <= 251970xxxxxx: FRAUD (104.0/1.0)

Rule 24: A_NBR > 251970xxxxxx, CALL_FEE <= 1380: NonFR (58.0/1.0), CALL_FEE > 1380: FRAUD (9.0/1.0)

Rule 25:CELL_A > 636012412042676: FRAUD (175.0/3.0)