Predictive SIM Box Fraud Detection Model for ethio telecom

Feven Fesseha
HiLCoE, Software Engineering Programme, Ethiopia
feven.fesseha@ethiotelecom.et

Mesfin Kifle
HiLCoE, Ethiopia
Department of Computer Science, Addis Ababa University, Ethiopia
kiflemestir95@gmail.com

Abstract

Fraud in communication has been increasing dramatically due to the new modern technologies and the global super highways of communication, resulting in loss of revenue and quality of service in telecommunication providers. One of the dominant types of fraud is SIM box bypass fraud whereby the SIM cards in the SIM box are used to channel international calls away from mobile operators and deliver as local calls. The major focus of this paper is to design a predictive SIM box fraud detection model by taking a total of 15,000 records from a telecom service provider company in Ethiopia. The research also use 12 selected attributes from a Customer Database Record (CDR). Four classification techniques from WEKA tool have been used to conduct the experiments. These are Artificial Neural Network (ANN), Decision tree learning, K nearest neighbor and function based Support Vector Machine (SVM). The classification algorithms use 12 selected features of data extracted from Customer Database Record (CDR). Comparison of the performance of each algorithm is made to select the algorithm with the best performance. As a result of the experiments, it is found that BFTree algorithm from decision tree model gives higher accuracy compared to the others with a classification accuracy of 99.9%. Using the pattern/output of this algorithm, a prototype is developed.

Keywords: SIM box fraud; Artificial Neural Network; Support Vector Machine; Decision Tree Learning; K Nearest Neighbor Classification; CRISP

1. Introduction

Nowadays, there has been a significant growth in the development of universal cellular network technology. All countries across the world use mobile phones with subscriber identity module (SIM) cards for communication among individuals. However, this phenomenal growth has brought an increase in variety and complexity of fraudulent activities on mobile networks. Some of the different types of telecom frauds that may cause gigantic loss of company's financial revenue are [2]:

- Interconnect Bypass Fraud (SIM box fraud),
- International Revenue Sharing Fraud (IRSF),
- False Answer Fraud,
- "A" Number Pass-through/Interconnect Agreement Compliance Testing,
- Superimposed Fraud, and
- Subscription Fraud.

The primary goal of any telecom company is to make sure that all calls that are directly or indirectly routed through the SIM must be legal and authorized. Especially, in developing countries, it is very easy for fraudsters to make unauthorized access and use the company asset. The reasons why fraudsters commit frauds are:

- Failure to understand the complexity of new technologies,
- Dissatisfaction of employees due to lack of experience with new technology,
- Weaknesses in systems (some operational systems are prone to fraudsters to commit frauds),
- International call termination fee is very expensive compared to local call,
- Political and ideological difference, and
- The fast growth in technology and the fraudsters have varieties of ways in committing frauds.

There are different fraud detection technologies proposed in ethio telecom to detect or reduce the losses caused by fraud but the problem has continued to pose significant revenue loss to the company.

Fraud is wrongful or criminal deception intended to result in financial or personal gain. Fraud can happen in most industrial sectors including health, telecommunication and banking institutes. In companies where there is more than one service provider, the actions of the fraudsters cause lots of inconveniences to subscribers and might encourage customers to switch to another competing provider. Again, large number of revenue loses in a country is a cause of fraud. These fraudsters may have different motives in order to commit fraud.

Different researchers define fraud in different ways. According to Levi and Burrows [1], fraud is "a mechanism through which the fraudster gains an unlawful advantage or causes unlawful loss".

According to the Cambridge Advanced Learner's Dictionary, fraud is "an intentional deception or cheating intended to gain an advantage.

One of the most severe threats to revenue and quality of service in telecom providers is fraud. The arrival of new technologies has provided fraudsters new techniques to commit fraud. SIM box fraud is one of such fraud that has emerged with the use of Voice over IP (VoIP) technologies.

SIM box or international bypass fraud is the most popular fraud type. This fraud also causes large number of revenue loss in a telecom company. SIM box fraud occurs when an international call is terminated with a local call tariff. Mostly SIM box fraud takes place when the cost of terminating domestic or international calls exceeds the cost of a local mobile-to-mobile call in a particular region or 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. As a result, the calls become local at the destination network [3].

Fraudsters gain some profit by providing low international call tariff at the time they buy bulk SIM cards and put those SIM cards in hardware devices which are SIM box in the recipient country. Moreover, every time there is a new international call arrival, they redirect the international call to the SIM cards in the SIM box using voice over IP. Finally the call reaches to the destination route with local numbers with local call tariff. Due to this, the operator loses the revenue that it should have received from international calls.

2. Related Work

Several researches have been done in the domain of fraud detection. They include fraud detection in credit cards, telecommunications, money laundering, and intrusion detection. This section reviews some prominent works related to fraud detection methodologies in the telecommunication industry. Most of the researches follow different Data mining approach to detect fraud. This techniques include machine learning algorithms and rule based systems.

The researchers in [4] analyze fraudulent SIM box traffic based on communication data from one of the major tier-1 network operators in the United States. The SIM boxes operate with a large number of SIM cards from foreign and national operators. If an operator detects and shuts down a fraudulent account, the researchers made their observation and do some analysis on some fields of the CDR. Also, they noticed that fraudulent SIM boxes have almost static physical locations and generate disproportionately large number of outgoing calls (100 times as many as incoming calls). Based on these observations, they introduced three classifiers for fraudulent SIM boxes and combine their outputs into a classification rule, which has a high detection rate and correctly filters out mobile network probes with traffic patterns similar to those of SIM boxes.

A comparison is made between SVM and ANN technologies to detect SIM box fraud in [5]. In this work, two classification techniques, namely, Artificial Neural Network (ANN) and Support

Vector Machine (SVM) were used to detect this type of fraud. These classifications use nine selected features of data extracted from CDR. In the experiments, the dataset used consists of 2126 fraud subscribers, 4289 normal, and 9 fields from CDR records and it is found that SVM gives higher accuracy compared to ANN with a classification accuracy of 99.06% compared to ANN which has 98.71% accuracy. Besides better accuracy, SVM also requires less computational time compared to ANN since it takes lesser amount of time in model building and training.

Six features extracted from CDR data are utilized to build decision tree that can be used to distinguish between legitimate and SIM box (fraud) accounts in [6]. The features include the total number of outgoing and incoming calls, the total number of SMS originating and SMS terminating, the total number of hand over and the total number of different locations. The proposed decision tree algorithm has shown accuracy of 97.95% when it was tested using sample data from Almadar Aljadid company.

In [7], a total of nine features found to be useful in identifying SIM box fraud subscriber are derived from the attributes of CDR. Artificial Neural Networks (ANN) have shown promising results in classification problems due to their generalization capabilities. Therefore, supervised learning method was applied using Multilayer perception (MLP) as a classifier. Dataset obtained from a mobile communication company was used for the experiments. ANN showed classification accuracy of 98.71 %.

In [8], a new fraud detection technique is proposed. The proposed technique depends on user profiling and using fuzzy logic. Fuzzy logic was used in decision making process by utilizing fuzzy logic membership function. A database from a mobile operator company (Almadar Aljadid) in Libya is used in this investigation. Five features are extracted and employed as detection patterns in the proposed technique. The five features or detection patterns are, subscriber's mobility, incoming to outgoing calls ratio, suspicious cell activity, irregular calls, and service type.

After reviewing the above listed related literature we found that CDR data is mostly used and it is a relevant source for analyzing and determining fraudulent calls and also it is known that different features (attributes) from the CDR data can be used. We also found different data mining techniques can be used to detect fraudulent activities by analyzing call patterns.

The limitation of the above listed works is that there is no way of updating the existing patterns which makes fraudsters to know the existing patterns and come up with a committing fraud. This paper tries to address this gap and find a way of updating the existing patterns by integrating pattern based fraud detection method with a test call generation method.

3. Findings and the Proposed Model

Four classification algorithms selected from WEKA tool are used in this research. These are Decision tree learning, Neural networks, K nearest neighbor and function base techniques. A total of eight experiments have been done and Table 1 shows the comparison of each result from the experiments.

Table 1: Comparison of Different Models

Algorithm	*Using Cross Validation Testing*			*Using Separate Test Set Testing*		
	Time Taken to Build Model (Sec)	*Correctly Classified Instance %*	*Incorrectly Classified Instance %*	*Correctly Classified Instance %*	*Incorrectly Classified Instance %*	*Average False Positive Rate*
BFTree	1	100%	0%	99.99%	0.08%	0
SMO	9.44	86.38%	13.61%	86.54%	11.54%	0.555
MLP	27.6	93.13%	6.866%	92.026%	7.97%	0.222
IBK	0.01	78.447%	22.552%	88.72%	11.27%	0.136

Using four different algorithms, 8 experiments have been done and among those the classification accuracy level of BFTree is very high. It correctly classifies 99.92%, so the pattern generated from BFTree is used to do the deployment. The rules are listed below.

RULE 1: if incoming to outgoing call ratio < 0.49603: Fraud

RULE 2: if Data usage in MB < 2.0: Fraud

RULE 3: if number of outgoing SMS < 5.5: Fraud

RULE 4: if number of incoming SMS <10.0: Fraud

RULE 5: if number of incoming SMS >= 10.0: Normal

RULE 6: if number of outgoing SMS >= 5.5: Normal

RULE 7: if Data usage in MB >= 2.0: Normal

RULE 8: if incoming to outgoing call ratio >= 0.49603: Normal

RULE 9: if Number of frequently called numbers <=5: Fraud

RULE10: if Number of frequently called numbers >=6: Normal

After analyzing the patterns generated from BFTree, a prototype is developed. This prototype will help ethiot elecom to predict SIM box Fraud. Figure1 shows the entire model.

4. Deployment

In the deployment stage a general procedure has been identified to create the relevant model(s).

The interface is developed using Microsoft Visual Studio 2012, ASPX.NET with C#, Developer Express and from the database side Microsoft SQL Server 2012 is used. Using this interface one can check a weather given number is fraud or not. This interface can be integrated with ethio telecom's CDR database and the data can be synchronized between both systems. This interface considers the patterns generated from BFTree algorithm.

The homepage shown in Figure 2 contains of search functionality. The user can insert an input and click the search button. To facilitate the user interaction, the system provides an advanced search facility. The advanced search facility allows users to retrieve data based on their desire with single or multiple fields. To use the advanced search facility, the user must select the field in which s/he would like to retrieve the data. After selecting the field the user should have to select the operator in which s/he would like to retrieve the data and insert the search criteria on the field provided. To use multiple field search functionality the user has to select the connector operator (Logical operator like AND, OR, NOT) that would be applied between the fields.

This new system can also be integrated with the current ethio telecom's fraud detection system which is TCG and the call patterns from the numbers which are found from the TCG are also registered in order to update the patterns of this system.

The advanced search functionality also allows users to filter records based on different criteria (like call number, call type, etc.) and different operator (like equal to, not equal to, less than, greater than, etc.).

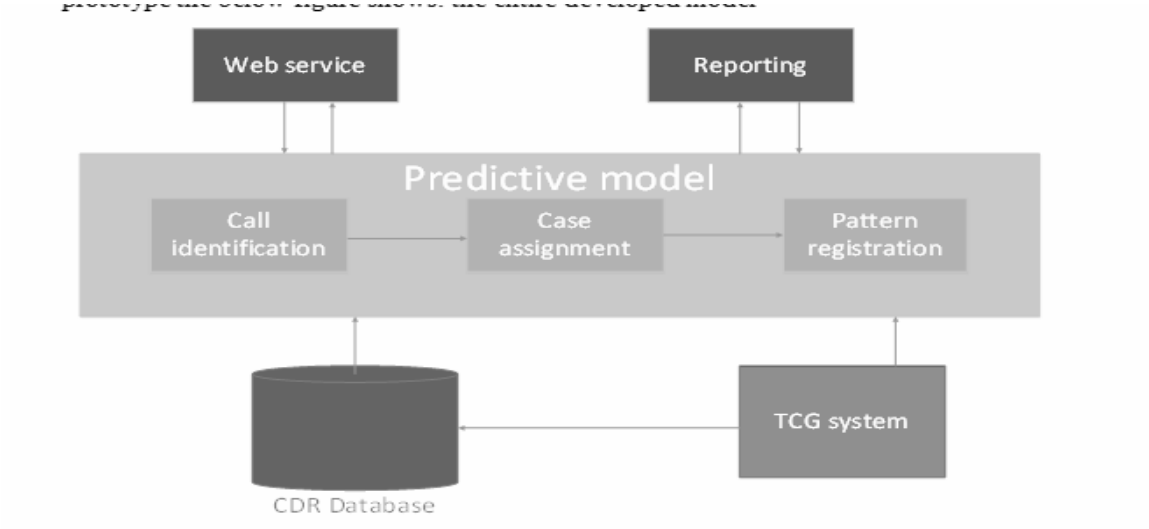

Figure 1: Architectural Design of Predictive SIM Box Fraud Detection Model

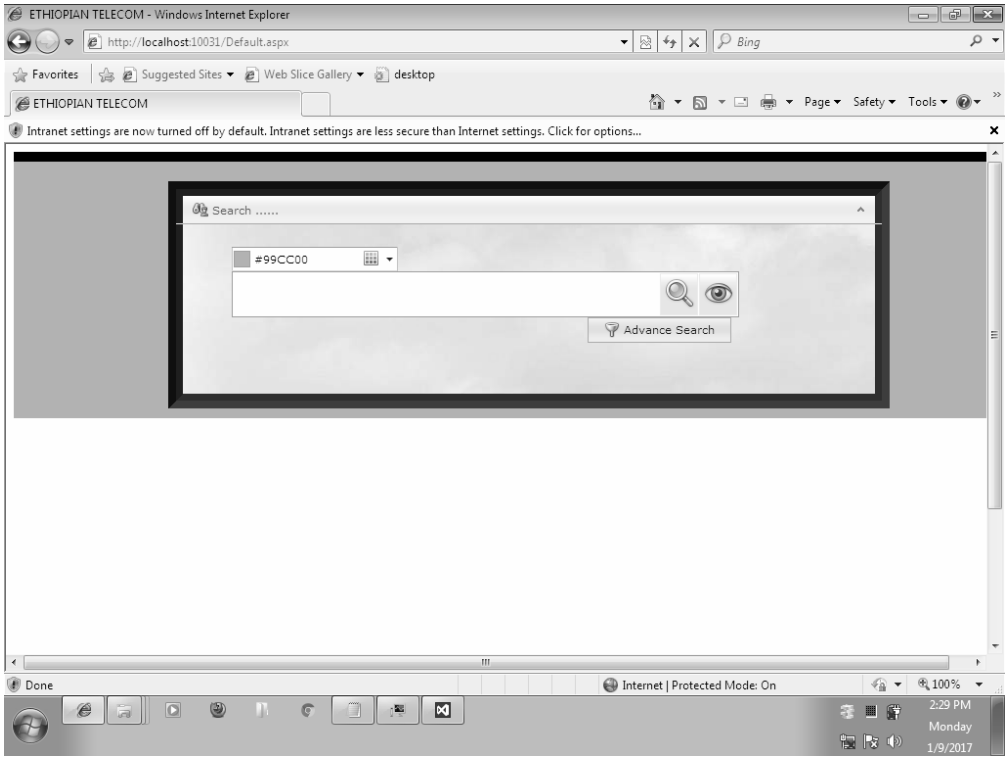

Figure 2: Starter Home Page

Figure 3 shows the advanced search screen. A user can search using different search criteria. The user can filter records by first selecting search criteria (i.e., the filed name that the user intended to filter records). Secondly, the user must select the operator that is to be applied on the search criteria. The operator may be Equal, Not Equal, Greater than, Less than, between, is not between, etc. At last the user have to enter the search term on the provided text box. The user can input the Plus sign to insert more than one line search criteria. In this case the user must select the operand (i.e., the operator that connect the two search criteria).

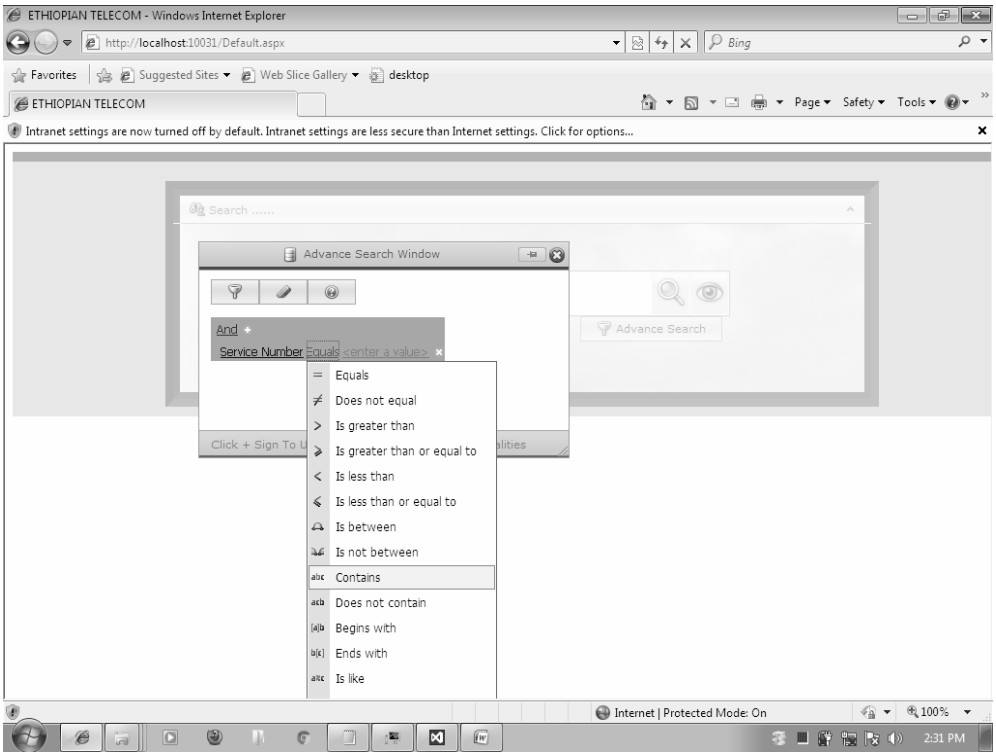

Figure 3: Advanced Search Screen

Figure 4 shows the page where a user can extract and view the report after specifying the searching criteria. On the report viewer, the system also incorporates a filtering mechanism that helps a user to filter and view the desired result based on all columns.

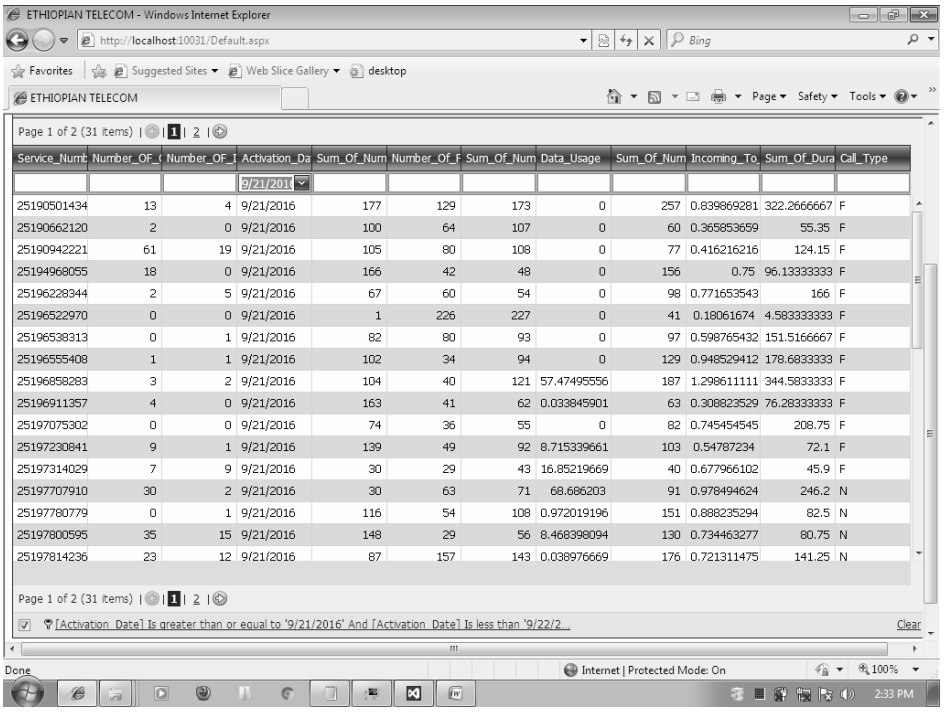

Figure 4: Report Viewer

Figure 5 shows the final page. After extracting the report the user can export it to Excel and view the detail. Here the user can open directly the report via Excel or save the result.

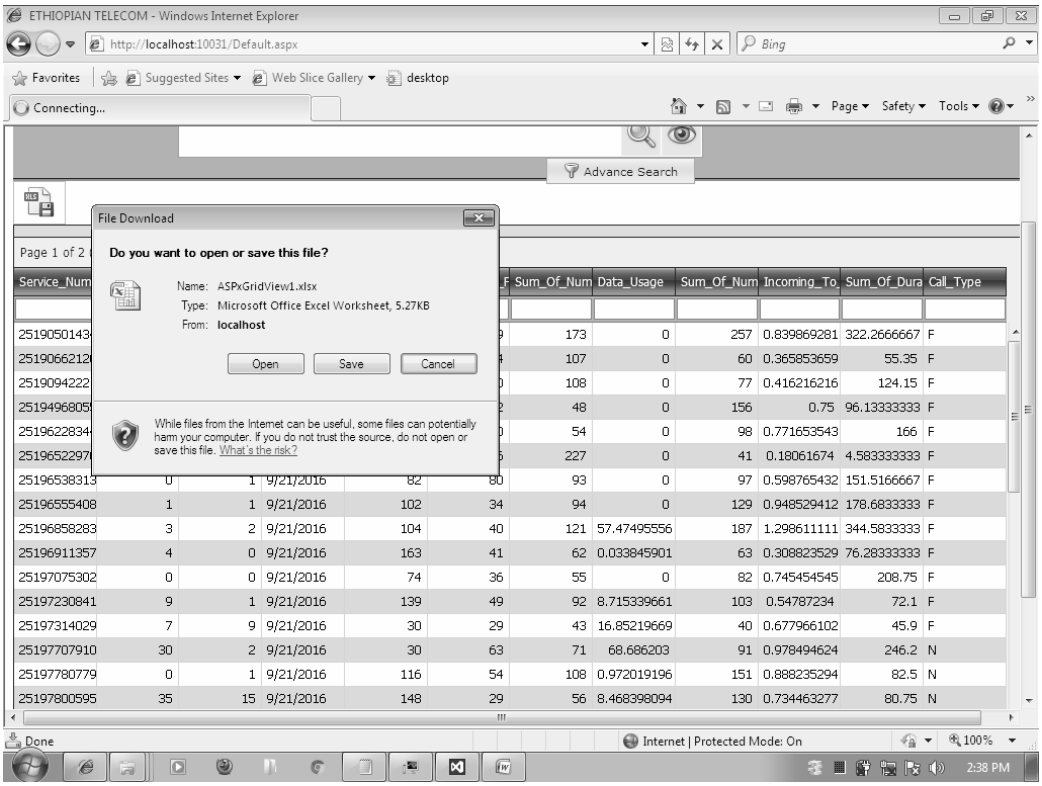

Figure 5: Export to Excel

5. Conclusion and Future Works

This research designs and implements a predictive model using data mining technology to detect SIM box or bypass fraud cases targeting ethio telecom. To achieve the objective, we used Call Detail Record (CDR) data. We selected a total of 15,000 customers based on the recharge amount (more than Birr 3000). 12 attributes have been selected.

CRISP-DM process model was used while undertaking the experimentation. The study was conducted using WEKA software version 3.6.3 and four data mining algorithms for classification were used. The rule generated from BFTree algorithm is chosen to develop the prototype.

References

[1] M. Levi and M. Burrows, "Measuring the Impact of Fraud in the UK: A Conceptual and Empirical Journey", British Journal of Criminology, 48 (3), pp. 293–318, 2008.

[2] Ogundile O. O., "Fraud Analysis in Nigeria's Mobile Telecommunication Industry".

[3] David Michaux, CEO Scant, "Telecom Fraud".

[4] Roselina Sallehuddin, Subariah Ibrahim, Azlan MohdZain, and Abdikarim Hussein Elmi, "Detecting SIM Box Fraud Using Support Vector Machine and Artificial Neural Network", 2015.

[5] Ahmed Aljarray and Abdulla Abouda Almadar Aljadid, "Analysis and Detection of Fraud in International Calls Using Decision Tree".

[6] Hussein M. Marah, Osama Mohamed Elrajubi, and Abdulla A. Aboud, "Fraud Detection in International Calls Using Fuzzy Logic".

[7] Abouda Abdulla A. Marah, Hussein Elrajubi, Osama Mohamed,"Fraud Detection in International Calls Using Fuzzy Logic".

[8] Kalpana Rangra and K. L. Bansal, "Comparative Study of Data Mining Tools", International Journal of Advanced Research in Computer Science and Software Engineering; June 2014.