Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Science and Engineering

Elixir Comp. Sci. & Engg. 60 (2013) 16146-16151

ISSN: 2229-712X

Hybridized Intelligent Data Analysis Model for Fraud Detection in Mobile Communication Networks

Aranuwa Felix Ola

Department of Computer Science, Adekunle Ajasin University, Akungba – Akoko, Ondo State, Nigeria.

ARTICLE INFO

Article history:

Received: 6 May 2013;

Received in revised form:

17 June 2013;

Accepted: 3 July 2013;

Keywords

Intelligent data,
Mobile communication,
Pattern recognition,
Fraud detection,
User profiling.

ABSTRACT

The development of intelligent data analysis techniques for fraud detection can be well motivated from an economic point of view. Following the definition of fraud, it is easy to state the losses caused by fraud as primary motivation for fraud detection mechanism. Fraud in communication networks can be described and characterized as determined unobserved intentions to illegitimately use the communication networks in order to avoid service charges or gain unjust advantage. Efforts in this work are directed at fraud detection in post-paid organizational mobile communications networks. Two different complementary approaches are used: (differential and absolute, user profiling and classification approaches). It is observed that fraudulent intentions are reflected in the observed call data, which was subsequently used in describing behavioural patterns of users. Relevant user groups based on call data were identified and users are assigned to a relevant group to model the fraud detection mechanism. In the task the call data was used to learn models of calling behaviour so that these models make inferences about users' intentions. From the analysis and model detectability experiment carried out in this scientific research work. It was discovered that the model detects over 89% of the fraudsters in the testing set (i.e fraud with certainty factor of 0.89). With the bias proportion of 0.0 and Mean Absolute Error (MAE) of (2.71) generated in the fraud detection. The model of course shows a good performance.

Introduction

Generally, the problem of the fraudulent use of mobile phones is a common thing to communication service providers. By definition, fraud in communication networks can be defined as the illegal access to the network and the use of its services with the intention to avoid service charges or gain unjust or undue advantage, while fraud detection is referred to as an attempt to detect illegitimate usage of a communication network, [4], [7]. Following the definition of fraud, it is easy to state the losses caused by fraud as primary motivation for fraud detection mechanism. In addition to financial losses, fraud may cause distress, loss of service, loss of customer confidence, reputation of network operators may even suffer from an increasing number of fraud cases. Apart from the facts stated above, user profiling effort is also motivated by the need to understand the behaviour of customers to enable provision of matching services and improve operations.

Therefore, it is believed that with the help of improved fraud detection models, fraudulent activities in mobile communication networks can be revealed and this would be beneficial to the network operator, who may lose several percent of revenue to fraud, since the service charges from the fraudulent activity remain uncollected, [2]. Hence, the need for improved fraud detection model, which is the focus of this work.

Fraud in Telecommunications Networks

Fraud in telecommunications networks can be characterized by fraud scenarios, which essentially describe how the fraudster gained the illegitimate access to the network. According to Oluwagbemi, (2008). Two major fraud scenarios were identified: Subscription and super-imposed fraud. In subscription fraud, fraudsters obtain a phone account without having any intention to pay the bill. In such cases, abnormal usage occurs throughout the active period of the account. Such account is usually used for call selling or intensive self usage. Also into this category falls the case of bad debt, where customers who do not necessarily have fraudulent intentions, but never pay a single bill, but later drop the line for another. In superimposed fraud, here fraudsters "take over" a legitimate account. In such a case, the abnormal usage is superimposed upon the normal usage of the legitimate customers. Examples of such include handset theft [8].

Literature Review

Researchers have developed growing interests in analyzing fraud scenarios and creating fraud detection engines. However, according to Hollm'en and Tresm, (1999) and Moreau et al, (1997), earliest detection methodologies designed for one specific scenario are likely to miss plenty of the others. Therefore, the interest is increasing on devising advanced and intelligent detection mechanisms based on the behavioural modelling of calling activity, which is the focus of this research work and publication. Michiaki et al, (2000) described fraud detection as the attempt to detect illegitimate usage of a communication network. They presented three methods to detect fraud in their work [5]. Firstly, a feed-forward neural network based on supervised learning was used to learn a discriminative function to classify subscribers using summary statistics. Secondly, Gaussian mixture model was used to model the probability density of subscribers' past behavior so that the probability of current behaviour can be calculated to detect any abnormalities from the past behaviour and lastly, bayesian

Tele: +2347031341911, +60102944150
E-mail addresses: felixaranuwa@gmail.com

networks was used to describe the statistics of a particular user and the statistics of different fraud scenarios. They conclude that Bayesian networks can be used to infer the probability of fraud given the subscribers' behaviour. Their experiments show that the methods detect over 85 % of the fraudsters in their testing set without causing false alarms.

Wang et al, (2004) proposed a feature extraction method called GPCA based on IG (information gain) and PCA (principal component analysis). They analyzed the data on CDR (call detail record), customer information, paying and arrear information etc. in mobile communication networks. The data was then used by the SVM classifier to build the fraud detection model. In the work, GPCA outperforms some of the most popular feature extraction methods such as BS (bivariate statistics), IG and PCA in predicting accuracy and training time. To get the higher predicting accuracy, a binary SVM using RBF (radial basis function) kernel was used. The experiments show that the classifier with GPCA has fine predicting accuracy.

Panigrahi et, (2007), introduced a framework for fraud detection in mobile communication network. The proposed fraud detection system (FDS) consists of four components, namely, rule-based deviation detector, Dempster-Shafer component, call history database and Bayesian learning. In the rule-based component, they determine the suspicion level of each incoming call based on the extent to which it deviates from expected call patterns. Dempster-Shafer's theory was used to combine multiple evidences from the rule-based component to compute an overall suspicion score. A call is classified as normal, abnormal, or suspicious depending on this suspicion score. The results of their experimental show that the method is promising in detecting fraudulent behaviour without raising too many false alarms.

Andrej et al., (2009), presented a novel bi-ANN-based approach for generic mobile-phone fraud detection capable of detecting fraud in real time. The analyses were accomplished using real-life CDR data, obtained from a Slovenian mobile operator. The overall finding of the study is that the bi-ANN is capable of predicting time series, resulting in 90% success rate in optimal configuration.

In the work of Ogwueleka, (2010), a fraud detection tool utilizing both rule-based and neural network technologies on the record of network subscribers and network traffic was developed. The performance evaluation showing the percentage of correctly identified fraudsters versus the percentage of new subscribers raising alarms are optimized using the Receiver-Operating Characteristic (ROC) curve for both the training and test sets, resulting in very low false positive rates.

In this work, Two different complementary approaches was combined: (differential and absolute, user profiling and classification approaches) to model a detector, while neural networks was employed in learning the usage patterns from the subscriber's call data.

Materials and Methods

Data Collection:

The method employed in this research work included, data collection by survey from the telecommunication industry, data entry, data training and testing using neural network model embedded in a neural network software called Neuro-Solutions. To access the performance of the method, reports on the experimental results obtained are display by relevant graphs as shown in the figures 3 and 4. The data used was based on toll tickets, which are call records stored for billing purposes. The toll tickets created for each phone call made included information like identification of the caller, stating time of the call, duration of the call, and the call party number to mention a few.

Billing System in Telecommunication Industry

Telecommunication companies need an effective and accurate billing system to be able to assure their revenue. Billing systems process the usage of network equipment that is used during the service usage into a single Call Detail Record (CDR). The billing process involves receiving billing records from various networks, determining the billing rates associated with it, calculating the cost for each billing record, aggregating these records periodically to generate invoices, sending invoices to the customer, and collecting payments received from the customer, [2]. The billing system provides service to the user, collect user usage records, and generate invoices of each credit expire, each billing cycle depends on the billing type, collect payments and adjust customers' balances as shown in figure 1:

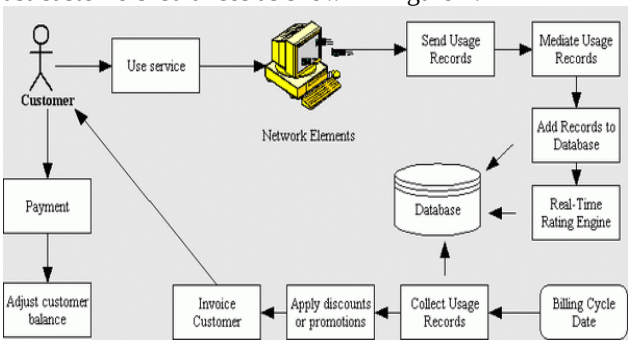


Figure 1 – Billing System Components

Billing Types: Billing types indicate whether and how the users pay to get and use the services [2]. The two main billing types in telecommunication discussed in this work are postpaid and prepaid billing types. In postpaid billing the customer may pay an insurance payment in advance, and he may pay the installation or setup fees, and in each billing cycle he will be invoiced (receive a bill) to pay for his usage of the service. In prepaid billing the customer buys a given amount of credits (duration, volume, number of events) and is then allowed to use the corresponding network resources as long as their account is in credit. Billing system receives customer usage records from the network elements and adjusts the customer credits. When user's credit is used up, network usage will be restricted. Prepaid corresponds to a real-time process, because transactions are only allowed if the user account is in credit, and this has to be checked in real-time.

Sample Representation

In order to develop models of normal and fraudulent behaviour and to be able to assess the diagnostic accuracy of the models, call data exhibiting both kinds of behaviour is needed. The data used for the purpose of this experiment was obtained by survey from telecommunication industry. The significance of the choice of this data can be seen in the important features it contains which helped to facilitate the discovery and detection of fraud occurrence in the usage patterns. The data set is a call record describing the daily usage of the network during the observed time period. The information about the call activity (call pattern) on the network is encoded in the toll tickets of all the calls placed on network. The data representation used in this research work is the features through aggregation in time coupled with absolute and differential analysis of user behaviours. (see Table 1 and Table 2 for details):

Table 1. - Sample Data from the Telecoms Industry

Time of call	Duration of call (min)	Destination No	Type of call (National or International)
8.24am	1	8059226458	national
8.40am	5	8024499351	national
8.53am	5	7031341911	national
9.01am	1	7031996511	national
9.32am	2	8059226458	national
10.02am	1	7036798193	national
10.30am	10	22672030193	International
10.45am	5	7064510884	national
11.06am	2	7069378895	national
11.32am	3	8133014443	national
11.51am	4	8024499351	national
12.02pm	2	8130585141	national
12.26pm	5	7060652263	national
1.15pm	7	22672030193	International
1.35pm	9	8024499351	national
2.05pm	4	7031341911	national
2.34pm	7	7031095668	national
2.57pm	12	8131996511	national
3.06pm	3	7032845881	national
3.27pm	2	7036583529	national
3.46pm	2	8133014443	national
4.04pm	6	22780967845	International
4.11pm	3	7031341919	national
5.05pm	1	8024499351	national
5.17pm	5	80371369499	national
5.30pm	3	9229335538	International
8.45pm	24	9442038357	International
10.06pm	15	13134670179	International
10.35pm	37	7038564686	national
11.15pm	5	8060693258	national

Data Pre-processing (Scaling technique)

In pattern recognition applications, the usual way to create input data for the model is through feature extraction. In feature extraction, descriptors or statistics of the domain are calculated from raw data, The data collected by survey was scaled for it to properly fit into the neural network software used for this research work. The unit of aggregation in time is one day reflecting the daily usage of an account. The typical features used in this application were time of calls, (where calls made during business hours, late evening or night hours were regarded as different categories), the other feature used was summary statistics of call duration and fraud occurrence value (FOV), during the observed period.

The Hybridized Intelligent Data Analysis (HIDA) Model

The architecture of the hybridized intelligent data (HIDA) model is depicted in the figure 2:

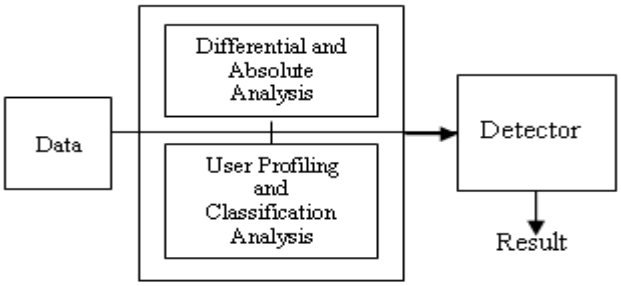


Figure 2: The HIDA Model Architecture

Data Analysis and Detector Constructor Components

The two different complementary approaches presented were used in this work to model the detector mechanism while neural networks was employed in learning the usage patterns from the call data. In the differential approach, a model of recent behaviour is used in quantifying novelty found in the future call data so as to detect abrupt changes in the calling behaviour, which may be a consequence of fraud. In the absolute approach, models typifying fraudulent and normal behaviour are used to determine the most likely mode. The user profiling is the process of modelling characteristic aspects of user behaviour. In user classification; users are assigned to distinctive groups. User profiling and classification are important tasks in data intensive environments where the behaviour of a heterogeneous mass of users is to be understood or where computer assisted decision making is sought for. Fraud detection is a prime example of this kind of problem. In an absolute and analysis, a user is classified as a fraudster based on features derived from daily statistics summarizing the call pattern such as the time of calls. In a differential analysis, the detection is based on measures describing the changes in those features capturing the transition from a normal use to fraud.

Detector Constructor

The approaches used in this work are focused on the pattern matching in which abnormal patterns are identified from the normality. Detector constructor framework is built to help in the data analysis and it is applied to the account history of subscriber's call accounts. Using differential analysis methods typically alarm deviations from the established patterns of usage. When current behaviour of a user differs from the established model of behaviour, alarm is raised. For example, in the case study, a user's calling patterns for business calls is between 8.00 am and 5.00pm and evening calls between 6.00pm and 9.00pm, any call in the night say 12.00 midnight, indicates that the call is very possible to be fraudulent.

Classification Rule Learning

This is firstly based on the given history of an account; calls are analyzed and labelled as fraudulent and legitimate (non-fraudulent) calls using the approach described above. The local set of rules for the account is searched. For example, for one specific account, the following classification rule is devised:

(Time-of-Day = Night) ⟶ Fraud with certainty factor = 0.89.

The certainty factor is defined as a simple frequency-based probability estimate. This rule means that a call is made at night can be considered fraudulent with 89% of the probability.

Implementation:

Data obtained from the post-paid subscriber's record were subject to a neural networks software called Neuro-Solutions. The phone data was keyed into the model for detection purpose and the following results were obtained as shown in Table 2, Figure 3 and Figure 4.

Table 2: Activities by Region

Time of call	**Duration** of call (mins)	**Destination** No	**Type of call** (National/International)	**Fraud Occurrence Value** (FOV)	**xi-x**	**(xi-x^2)**	
8.24am	1	8059226458	National	100	-9	81	
8.40am	5	8024499351	National	100	-5	25	
8.53am	5	7031341911	National	100	-5	25	
9.01am	1	7031996511	National	115	-9	81	
9.32am	2	8059226458	National	115	-8	64	
10.02am	1	7036798193	National	120	-9	81	
10.30am	10	22672030193	International	120	0	0	
10.45am	5	7064510884	National	120	-5	25	
11.06am	2	7069378895	National	125	-8	64	
11.32am	3	8133014443	National	125	-7	49	
11.51am	4	8024499351	National	125	-6	36	
12.02pm	2	8130585141	National	130	-8	64	
12.26pm	5	7060652263	National	130	-5	25	
1.15pm	7	22672030193	International	135	-3	9	
1.35pm	9	8024499351	National	135	-1	1	
2.05pm	4	7031341911	National	140	-6	36	NON FRAUD REGION
2.34pm	7	7031095668	National	140	-3	9	
2.57pm	12	8131996511	National	140	2	4	
3.06pm	3	7032845881	National	145	-7	49	
3.27pm	2	7036583529	National	145	-8	64	
3.46pm	2	8133014443	National	145	-8	64	
4.04pm	6	22780967845	International	150	-4	16	
4.11pm	3	7031341919	National	150	-7	49	
5.05pm	1	8024499351	National	155	-9	81	
5.17pm	5	80371369499	National	155	-5	25	
5.30pm	3	9229335538	International	155	-7	49	
8.45pm	24	9442038357	International	165	14	196	
10.06pm	15	13134670179	International	175	5	25	
10.35pm	37	7038564686	National	180	27	729	FRAUD REGION
11.15pm	5	8060693258	National	185	-5	25	
11.28pm	30	13134095125	International	185	20	400	
12.36am	20	8034336480	National	185	10	100	
1.00am	35	13134522934	International	190	25	625	
1.43am	13	8030517635	National	190	3	9	
2.10am	32	8051113022	National	200	22	484	
2.52am	19	8076919223	National	200	9	81	
3.33am	13	8059225014	National	205	3	9	
	353			5,390		3759	
MEAN	9.540540541		MEAN	147.972973	VARIANCE	101.5946	
STD	10.20782541		STD	30.15	STD	10.36	

- Regions of fraud occurrence take values between *175 - 205 at random*
- Regions of Non-fraud occurrence take values between 100 - 165 at random

Simulation results and model performance report

Modelling Region for Fraud and Non -fraud

Report on the experimental results obtained and the display of relevant graphs are shown in the figures 2 below: It shows the region of fraud occurrence value (FOV) for Non-fraud starting from 100 - 165 occurring from item 1 to 27 while fraud occurrence value for fraudulent region starts from 175 – 205 between item 28 to 38.

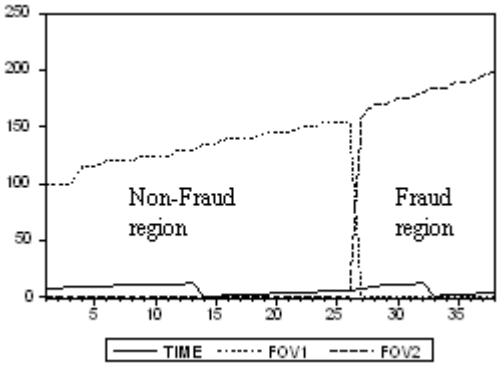
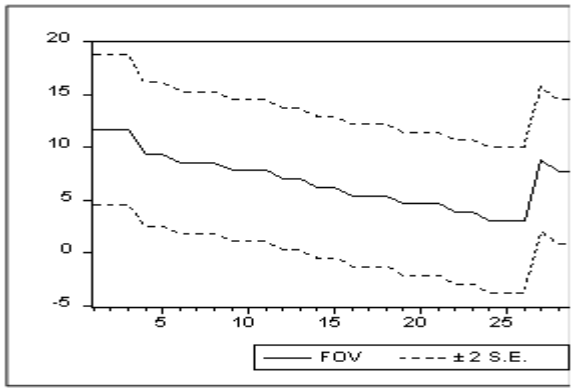


Figure 3 - Region of fraud and Non fraud

Forecast Bias Proportion and Mean Absolute Error

From the analysis and model detectability experiment carried out in this scientific research work. It was discovered that the model had the bias proportion of 0.0 and Mean Absolute Error of (2.711609) generated in its fraud detection capability which of course shows a good performance and that there is no bias proportion in the capability of this model.



Forecast: FOV	
Actual: TIME	
Forecast sample: 1 38	
Included observations: 38	
Root Mean Squared Error	3.145132
Mean Absolute Error	2.711609
Mean Abs. Percent Error	82.34218
Theil Inequality Coefficient	0.213024
Bias Proportion	0.000000
Variance Proportion	0.260842
Covariance Proportion	0.739158

Figure 4 - Forecast Bias Proportion and Mean Absolute Error

Actual Fraud Detection Analysis

The analysis of the experiment carried out to detect actual fraud shows the changes in the call pattern usually occurs after the office hours calls (i.e after 9.00pm) which depict the actual time of fraud. The performance in terms of the error generated in this fraud detection experiment showed that its NMSE (Normalized mean squared error) for the fraud detection was 3.145132 and the mean absolute error (MAE = 2.711609) which shows a good performance as well.

Conclusion

Intelligent data analysis is an important task in data intensive environments where the behaviour of a heterogeneous mass of users is to be understood or where computer assisted decision making is sought for. Fraud detection is a prime example of this kind of problem. In this work, the fraud detection system presented attempts to discover illegitimate behaviour of the users in the data used for the work. The system worked on the actions of he users, that is, their calling behaviour (calling patterns) in order to make plausible decisions about fraud occurring. The approach provides an economic feature to learning how to detect fraud in mobile communication networks. The techniques presented in this work can be seen as solutions to a specific user profiling problem in fraud detection in post-paid organizational mobile communication network. The technique is shown to be effective in detecting fraudulent behaviour in mobile communication network by testing the method with data from real mobile communications networks. However, several similar problems exist, for instance fraud detection in other communication networks and other type of frauds. Therefore, it is recommend that more research can still be carry out in order to boost this work further. To cover such fraud as subscription or velocity trap and comparing this method with other models.

References

[1] Andrej, K,, Mojca, V., Urban, S., Janez, B, and Andrej, Kos (2009). Bidirectional Artificial Neural Networks for Mobile-Phone Fraud Detection. ETRI Journal, Volume 31, Number 1, February 2009.

[2]. Aranuwa, F.O., Longe, O.B., and Ukpe, K. (2011). Post-Paid Mobile Communication Networks Call Data Modelling For Fraud Detection on GSM Platforms. Proceedings of International Conference on ICT for Africa, 23rd -26th March, 2011. Covenant University & Bells University of Technology, Ota, Nigeria.

[3] Hollm´en, J. and V. Tresp (1999). Call-based fraud detection in mobile communication networks using a hierarchical regime-switching model. In M. Kearns, S. Solla, and D. Cohn (Eds.), Advances in Neural Information Processing Systems 11: Proceedings of the 1998 Conference (NIPS'11), pp. 889–895. MIT Press.

[4]. Jaakko Hollmén (1997). Novelty filters for fraud detection in Mobile Communications Networks. Technical Report A48, Helsinki University of Technology, Laboratory of Computer and Information Science.

[5]. Michiaki Taniguchi, Michael Haft, Jaakko Hollmén, Volker Tresp (2000). Fraud detection in communications networks using neural and probabilistic methods. Siemens AG, Corporate Technology Department Information and Communications D-81730 Munich, Germany. e-mail: Michiaki. Taniguchi@mchp.siemens.de.

[6]. Moreau, Y., Verrelst, H. and Vandewalle, J. (1997). Detection of mobile phone fraud using supervised neural networks: A first prototype. In International Conference on Artificial Neural Networks Proceedings (ICANN'97), pp. 1065–1070.

[7]. Ogwueleka, F.N,(2010): Fraud Detection in Mobile Communications Using Rule-Based and Neural Network System. The IUP Journal of Science & Technology, Vol. 6, No. 4, pp. 21-34, December 2010.

[8]. Oluwagbemi O.O (2008), Predicting Fraud in Mobile Phone Usage Using Artificial Neural Networks. Journal of Applied Sciences Research, 4(6): 707-715, 2008.

[9]. Panigrahi, S., Kundu, A., Sural, S., Majumdar A. K (2007). Use of Dempster-Shafer Theory and Bayesian Inferencing for Fraud Detection in Mobile Communication Networks Information Security and Privacy Lecture Notes in Computer Science. 12th Australasian Conference, ACISP 2007, Townsville, Australia, July 2-4, 2007. Proceedings Volume 4586, 2007, pp 446-460 .

[10] Wang, D., Wang, Q., Zhan S., Li, F (2004). A feature extraction method for fraud detection in mobile communication networks. A conference proceedings of Intelligent Control and Automation, WCICA. Fifth World Congress on Volume:2, 15-19 June 2004.