Detecting Unusual Changes of Users Consumption

Paola Britos[1], Hernan Grosser[2], Dario Rodríguez[3], and Ramon Garcia-Martinez[4]

Abstract The points being approached in this paper are: the problem of detecting unusual changes of consumption in mobile phone users, the corresponding building of data structures which represent the recent and historic users' behavior bearing in mind the information included in a call, and the complexity of the construction of a function with so many variables where the parameterization is not always known.

1. Introduction

When a mobile call is started, the cells or switches record that it is being made and they produce information referring to this event. These records are commonly called CDR's (Call Detail Records). CDR's contain useful information about the call so that it can be properly charged to whom it may correspond [1]. They can also be used to detect any fraudulent activity considering well-studied fraud indicators. That is, processing an amount of recent CDR's and comparing a function of the different fields such as, IMSI (International Mobile Subscriber Identity, which univocally identifies a user in a mobile phone network), date of call, time of call, duration, type of call (with a specific criteria). If this function retrieves a value that is considered beyond normal limits, an alarm is set off. This alarm must be taken into account by fraud analysts in order to determine if there

[1] Paola Britos
PhD Program, Computer Science School, La Plata University. CAPIS-ITBA. pbritos@itba.edu.ar
[2] Hernan Grosser
Intelligent Systems Lab. School of Engineering. University of Buenos Aires. hgrosser@fi.uba.ar
[3] Dario Rodríguez
Software & Knowledge Engineering Center (CAPIS), ITBA. drodrigu@itba.edu.ar
[4] Ramon Garcia-Martinez
Software & Knowledge Engineering Center (CAPIS), ITBA. rgm@itba.edu.ar

Please use the following format when citing this chapter:

Britos, P., Grosser, H., Rodríguez, D. and Garcia-Martinez, R., 2008, in IFIP International Federation for Information Processing, Volume 276; *Artificial Intelligence and Practice II*; Max Bramer; (Boston: Springer), pp. 297306.

has been any activity in bad faith or not. To be able to process these CDR's, it is necessary to make previously a process known in telecommunications as mediation, in which the information is read with the format of record in which CDR's come and then it is encoded in a new format of record which is understood by the fraud system.

The existing systems of fraud detection try to consult sequences of CDR's by comparing any field function with fixed criteria known as Triggers. A trigger, when activated, sends an alarm which leads to fraud analysts' investigation. These systems make what is known as a CDR's absolute analysis and they are used to detect the extremes of fraudulent activity. To make a differential analysis, patterns of behavior of the mobile phone are monitored by comparing the most recent activities to the historic use of the phone; a change in the pattern of behavior is a suspicious characteristic of a fraudulent act. [1]

2. Description of the problem

In order to build a system of fraud detection based on a differential analysis it is necessary to bear in mind different problems that arise and must be carefully worked on. These are:

2.1. The problem of building and maintaining "users' profiles"

The majority of fraud indicators are not analyzed by using a unique CDR. In a system of differential fraud detection, information about the history together with samples of the most recent activities is necessary. An initial attempt to solve the problem could be to extract and encode CDR's information and store it in a given format of record. To do this, two types of records are needed; one, which we shall call CUP (Current User Profile) to store the most recent information, and another, to be called UPH (User Profile History) with the historic information [2], [3] and [2]. When a new CDR of a certain user arrives in order to be processed, the oldest arrival of the UPH record should be discarded and the oldest arrival of the CUP should enter the UPH. Therefore, this new, encoded record should enter CUP. This information should be stored in a compact form so it is easy to analyze later on by the system of fraud detection. Considering the amount of information that a CDR contains it is necessary to find a way to "classify" these calls into groups or prototypes where each call must belong to a unique group. This raises several important questions to deal with: (a) What structure must CUP and UPH records have?, (b) How many groups or prototypes must CUP and UPH records have in order to take the necessary information?, (c) How can calls be classified in the different, pre-defined prototypes? and (d) How to encode calls so that they can be "prototyped".

2.2. The problem of detecting changes in behavior

Once the encoded image of the recent and historic consumption of each user is built, it is necessary to find the way to analyze this information so that it detects any anomaly in the consumption and so triggers the corresponding alarm. It is here that the most important question of the whole paper arises and it is: How can the changes in a user's pattern of behavior may be detected? Our problem, then, is focused not only on the detection of abnormal changes in consumption, but also and fundamentally on building the data structures that represent the recent and historic behavior of each user considering the great amount of information that a call takes and the complexity of building a function with so many variables of input, complex and unknown.

3. Description of the suggested solution

The solution that has been developed has taken into account each and every question mentioned before, attempting to solve them in the most effectual and effectively possible way. Below is the presentation of each answer to the questions met in the analysis of the problem. In order to be able start processing the CDR's, a new format of record (mediation process output) must be created containing the following information: IMSI, date of call in YYYYMMDD format, time of the call in HH24MISS format, duration of call in 00000 format and type of call classified as LOC (local call), NAT (national call) and INT (international call). With this information together with the necessary data, it is possible to start solving the following and most important questions by using as input data the output of mediation process.

3.1. User's profiles construction and maintenance Solution

The first point to solve is to determine how to make the CUP and UPH profiles. This means fixing the patterns that will make up each of the profiles. The patterns must have information about the user's consumption, separating LOC consumption (local calls), NAT (national calls) and INT (international calls) respectively. An interesting way to build these patterns is using neural networks so as to determine the space of all users' calls generating a space of patterns which represent the consumption of all users, and then generating a distribution of frequencies by user in which the probability of a user making calls following this pattern is represented [2]. To sum up, when a user's profile is built, the representation of the distribution of frequency in which a certain user makes a certain call is made. This data structure shows the user's pattern of consumption. Among other advantages, neural networks have the capacity to classify the information in certain patterns. Especially, SOM (Self Organizing Map) networks can take this information and build these patterns in a way which is not supervised

by similarity criteria, and without knowing anything a priori about the data [3] and [4]. In our case, all the calls made by all users can be processed so that the networks, depending of the quantity of calls there are of each type, generate the patterns (creating resemblance groups) that represent all of them. To avoid noise in the data, three neural networks are used to generate patterns to represent LOC, NAT, and INT calls respectively. The user's profile is built using all three patterns generated by the three networks. The data used to represent a pattern are the time of the call and its duration. We know that if we represent, in a Cartesian axis, the time of all calls and their corresponding duration, we will obtain a rectangle full of points. The idea is to obtain a graph in which only the most representative points of the whole space will appear; that is the neural network task. Once the patterns that will be used to represent the user's profile are obtained, it is necessary to start filling them with information. The procedure consists of taking the call to be analyzed, encoding it and letting the neural network decide which pattern it resembles. After getting this information, the CUP user profile must be adapted in such a way that the distribution of frequency shows that the user now has a higher chance of making this type of calls. Knowing that a user's profile has K patterns that are made up of L patterns LOC, N patterns NAT and I patterns INT, we can build a profile that is representative of the processed call and then adapt the CUP profile to that call. If the call is LOC, the N patterns NAT and the I patterns INT will have a distribution of frequency equal to 0, and the K patterns LOC will have a distribution of frequency given by the equation [Burge & Shawe-Taylor, 1997a].

$$v_i = \frac{e^{-\|x-q_i\|}}{\sum_{j=1}^{L} e^{-\|x-q_i\|}}$$

where:
X: encoded call to be processed
v : probability that X call could be i pattern
Qi: pattern i generated by the neural LOC network.

Notice that: $\sum_{j=1}^{K} v_j = 1$

If the call were NAT, then L must be replaced by N and the distribution of LOC and INT frequencies will be 0; if the call were INT, then L must be replaced by I and the distribution of LOC and NAT frequencies will be 0.

Then, we can define the vector which represents V call, of K dimension, as

Vi = vi, with $1 \leq i \leq L$

Vi = 0, with $L+1 \leq i \leq K$, when the call is LOC.

Vi = vi, with $L+1 \leq i \leq L+N$

Vi = 0 with $1 \leq i \leq L$ y $L+N \leq i \leq K$, when the call is NAT.

Vi = vi, with $L+N+1 \leq i \leq K$

Vi = 0, with $1 \leq i \leq L+N$, when the call is INT.

Now that we have V vector, we can adapt CUP vector with the information of the processed call:

$CUP_i = \alpha_{LOC} CUP_i - (1- \alpha_{LOC}) V_i$, with $1 \leq i \leq K$, when the call is LOC,

$CUP_i = \alpha_{NAT} CUP_i - (1- \alpha_{NAT}) V_i$, with $1 \leq i \leq K$, when the call is NAT,

$CUP_i = \alpha_{INT} CUP_i - (1- \alpha_{INT}) V_i$, with $1 \leq i \leq K$, when the call is INT, where:

α_{LOC}: adaptability rate applied when call X is incorporated to CUP, if X corresponds to a local call.

α_{NAT}: adaptability rate applied when call X is incorporated to CUP, if X corresponds to a national call.

α_{INT}: adaptability rate applied when call X is incorporated to CUP, if X corresponds to an international call.

Once the CUP profile is adapted, it is compared with the UPH profile and then it is decided whether there has been a significant change in behavior (engine of detection of changes in behavior). After this, the UPH is adapted with the CUP information, only if the number of calls necessary to change the historic patterns has been processed,

$$UPH_i = \beta UPH_i + (1-\beta)CUP_i$$

With $1 \leq i \leq K$, where β: adaptability rate applied when CUP is incorporated to UPH.

3.2. Solution to the detection of changes in behavior

In order to settle whether there have been changes in the pattern of behavior or not, it is necessary to compare, somehow, the CUP and UPH profiles and decide if the difference between them is big enough so as to set an alarm off. Because both the CUP and the UPH are vectors that represent frequency distributions, a vectorial distance can be used to compare how different they are. For this, the Hellinger distance (H) can be used; it indicates the difference between two distributions of frequency [1]. This distance will always be somewhere between zero and two, where zero is for equal distributions and two represents orthogonally. The value of H will establish how different must CUP and UPH frequency distributions be, in order to set an alarm going. By changing this value, there will be more or fewer alarms set off.

3.3. Limitations of the solution

This solution is focused, as we described, on the analysis of the user's differential consumption. One case that may not be detected would be that in which the user always makes a lot of calls of the same type with a high consumption, as his pattern of behavior would never change. That is why there should always be a combination of several solutions in order to have a system of fraud detection that can detect different types of fraud. In this case, the absolute analysis would be a good solution. The other limitation centers in that the patterns are static, so that if the way in which the company users consume changes completely, it will be necessary to train again the neural networks to establish new patterns that represent the total space of calls and to re- build the CUP and UPH profiles as from the new distributions.

4. Experimentation

4.1. Methodology used

The experiments were divided in two parts: the first was focused on the training of the neural network and the generation of patterns to build the user's profiles later on; the second was aimed at the analysis of the calls made by high-consumption users and the corresponding analysis and detection of alarms. The second part of the test was divided again into two different experiences: 1) updating of UPH profile with each call (f= 1 call) and low Hellinger threshold (H) for the setting off of alarms of change of behavior; 2) updating of UPH profile once a day (f= 1 day) and high Hellinger threshold (H).

4.2. Experiments on the generation of patterns

Three SOM were built for the generation of patterns for LOC, NAT and INT calls respectively. Each of the networks was trained with an amount of calls that was representative of the consumption that company users made during a couple of days at all times. The calls were introduced to the network in a disorderly manner so that the patterns that were generated were not representative only of the time and duration of the last calls. The result of this experience defined the patterns to build the users' profiles. The patterns are made up of the time of the call and its duration in minutes, which managed to build a discrete space composed of all the types of called made by any user in a fixed quantity representative of that space.

4.3. Experiments on the construction of profiles and detection of behaviors

Once the patterns that define the space of all calls are obtained, tests have been carried out on the construction of user's profiles through the development of a distribution of frequencies of each of the patterns for each profile (CUP and UPH) and the corresponding detection of alarms. The process was based on the introduction to the system of calls made within a period of three months by users reported as "high-consumption user". With each call the CUP user profile was updated, it was then compared with the UPH profile, thus, obtaining the Hellinger distance between them. If it surpassed the fixed threshold, an alarm was set off. Depending on the parameter of updating frequency of UPH profile (f), the UPH was updated with the corresponding contribution of the CUP. At the moment of inputting a user's first call, all CUP and UPH patterns were initialized with the same distribution of frequency, assuming a priori that the user had the same

tendency to make any type of call, without any information. Moreover, this experience was carried out twice; the first one updating the UPH with each call, therefore, with a low Hellinger threshold (H) for the detection of alarms. This was because the difference that may arise between the CUP and UPH profiles was too small if updating the historic profile with each call, due to the fact that the historic profile tended to be the same as the current profile. The second experience was made by updating the UPH once a day and a high Hellinger (H) threshold to detect important differences that can be considered as changes in behavior.

5. Results

5.1. Generation of patterns

In this section results are presented after the training of the three SOM (See Fig. 1 to 3). The results show each of the patterns that the networks fixed as most representative of the space of all the users' calls. Three graphs are represented (one for each network) to show the patterns that were generated. On axis X, the time of the call is shown and on axis Y, the duration expressed in minutes is illustrated. Each of the points represented corresponds to a pattern being chosen by the network as representative of the sample. In the local neural network graph, 144 patterns are shown, in the NAT network, 64 and in the INT network, 36.

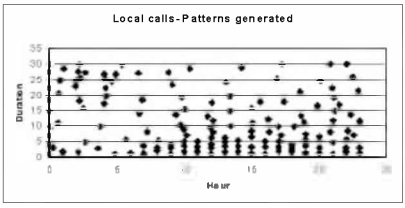

Fig. 1. Patterns generated after the training corresponding to local calls

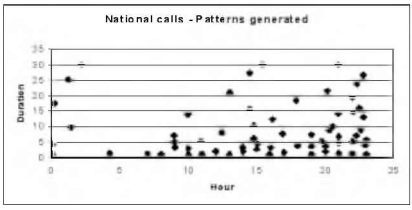

Fig. 2. Patterns generated after the training corresponding to national calls

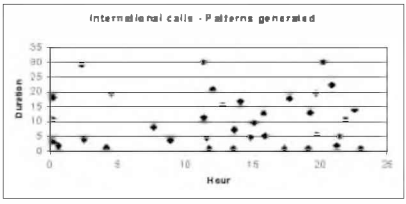

Fig. 3. Patterns generated after the training corresponding to international calls

The graph (Fig. 1) shows the 144 patterns generated after the training of the neural network corresponding to local calls. At simple sight, it is easy to notice that there is a greater concentration of patterns in the time range between 8h and 20h and duration of about 0–5 minutes. This denotes that most of the local calls made by this company's customers occur at these hours with the average durations indicated. The graph (Fig. 2) shows the 64 patterns generated after the training of the national calls neural network. Here, also, a concentration of patterns can be seen, but this time more towards the time range of 15h to 22h with durations that vary between 0 and 7 minutes. It also shows that there are practically no patterns generated for dawn, which may lead to conclude that most users of the company being analyzed do not make any NAT calls during early hours. The graph (Fig. 3) shows the 36 patterns after the training of the international calls neural network. Here the distribution is a little more aleatory, but the duration of the calls "chosen" as patterns tends to have a longer duration (between 7 and 10 minutes).

5.2. Profiles construction and changes detection in behavior

In this section, the results presented were obtained after the construction (from the company records) of the profiles and the detection of the corresponding alarms for each of the two experiences made. The graphs show the CUP and UPH profiles at the moment an alarm was set off. On axis X, the 244 patterns (144 LOC, 64 NAT and 36 INT) are shown and on axis Y the distribution of frequencies of each of the patterns for the user being analyzed at the moment the alarm was set off.

5.2.1. Experience 1 (Updating UPH with each call, high sensitivity with low Hellinger Threshold)

The graph (Fig. 4) shows a user's CUP at the moment an alarm was set off. It can be observed that the distribution of frequencies indicates a major tendency to make NAT calls (patterns 145 to 208). The graph (Fig. 5) shows the same user's UPH at the moment the alarm was set off. It can also be observed that the distribution of frequencies indicates a major tendency to make local calls (patterns 1 to 144). Hence, the difference between both distributions of frequencies defined by Hellinger distance (H) equals 0,30081. By analyzing the detail of this user's calls from dates previous to the triggering of the alarm to the day it was set off, there is evidence that the alarm responded to the user's making his first NAT call since his calls were processed.

This means, his historic pattern of behavior did not make it evident that this user would make such a call. However, when these calls were made, the system detected the change and generated the corresponding alarm. These results also show that, having made the experience with such high sensitivity, one single different call can indicate a change in behavior that leads to an alarm. The total number of alarms that were set off after analyzing the 60 users was 88, out of which 33 correspond to different cases.

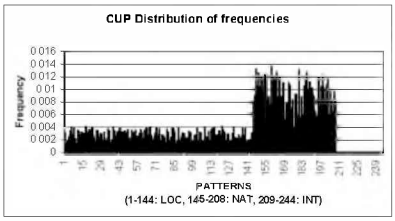

Fig. 4. User's CUP at the moment an alarm was set off

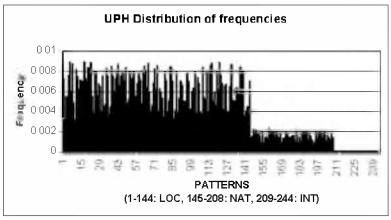

Fig. 5. User's UPH at the moment an alarm was set off

This is due to the fact that once an alarm for a user is set off, the following calls keep on setting off alarms till the UPH definitely adapts to the change in behavior. Most of the calls follow the pattern of the case in the graph in which a call that is different from the normal pattern of behavior is enough for the system to define the user as suspicious.

5.2.2. Experience 2 (Updating UPH once a day, moderate sensitivity with Hellinger threshold)

The graph (Fig. 6) shows a user's CUP at the moment an alarm was set off. It can be observed that the distribution of frequencies indicates a tendency to make local calls (patterns 1 to 144) and International calls (patterns 209 to 244). The graph (Fig. 7) shows the same user's UPH at the moment the alarm was set off.

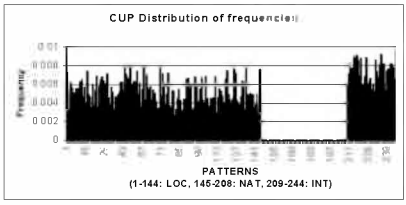

Fig. 6. User's CUP at the moment an alarm was set off

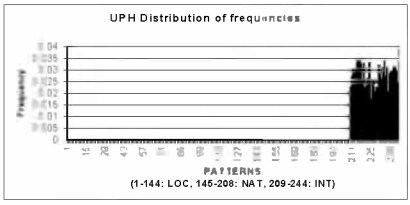

Fig. 7. User's UPH at the moment an alarm was set off

It can also be observed that the distribution of frequencies indicates a major tendency to make INT calls only (patterns 209 to 244). Therefore, the difference between both distributions of frequencies defined by Hellinger distance (H) equals 0,82815. By analyzing the detail of this user's calls from dates previous to the triggering of the alarm to the day it was set off, there is evidence that the alarm responded to the user's making only international calls till the moment that he started making local calls. When the number of local calls modified the CUP in the way illustrated by the graph, the alarm was triggered. This is a particular case as; surely, this alarm is not an indicator of fraud if the user pays his invoice for international calls. But it is an indicator of a sensitive change of behavior in the pattern of consumption, and that is exactly what this system searches. The total

number of alarms that were set off after analyzing the 60 users was 64, out of which 14 correspond to different cases. This is due to the fact that once an alarm for a user is set off, the following calls keep on setting off alarms till the UPH definitely adapts to the change in behavior. This phenomenon is emphasized here because it is only after calls of the next day are processed that the UPH is updated. The majority of the calls follow the pattern of the case in the graph in which there must be several calls outside the pattern of behavior for the system to find the user suspicious. This is much more satisfactory than what was obtained in experience 1 in which the high sensitivity showed users as suspicious simply for having made one single different call.

6. Conclusions

The results that were obtained were satisfactory in the sense that they were able to establish changes in the behavior of the users analyzed. Though the change in behavior does not necessarily imply fraudulent activity, it manages to restrict fraud analysts' investigation to this users' group. By using then other types of techniques [5], it is possible to obtain, with a high degree of certainty, a list of users who are using their mobile phone in an "unloyal" way. Besides, the experiences have helped to find users who have effectively changed their behavior, but in an inverse way, i.e. , they were users with high INT consumption and then they started making local calls. Commercially speaking, it could be an interesting tip to evaluate this type of consumers, since, for a certain reason they decided not to use their mobile phones to make international calls any more and it could help draw conclusions and create new rate plans based on these situations. It is also proven, with the experiences carried out, that the differential analysis provides with much more information than the absolute analysis, which can only detect peaks of consumption and cannot describe the user in question. As a final conclusion, neural networks can be said to be excellent tools for the classification of calls and the construction of users' profiles as they represent their behavior in a faithful and efficient manner.

References

1. ASPeCT, *Definition of Fraud Detection Concepts*, Deliverable D06. 47 pages. (1996).
2. Burge P, Shawe-Taylor J. *Fraud Detection and Management in Mobile Telecommunications networks*, Department of Computer Science Royal Holloway, University of London. Vodafone, England. Siemens A. G. (1997).
3. Kohonen, T. *Self-Organizing Maps*. Springer Series in Information Sciences. (2000).
4. Hollmen J. *Process Modeling using the Self-Organizing Map*, Helsinki University of Technology Department of Computer Science. (1996).
5. ASPeCT, *Fraud Management tools: First Prototype*, Deliverable D08. 31 pages. (1997).