DETECCION DE FRAUDE EN TELEFONIA CELULAR USANDO REDES NEURONALES

Grosser, H.[1], Britos, P. [2,3], Sicre, J.[2], Servetto, A. [4,3], García-Martínez, R.[2,1] y Perichinsky. G.[4,3]

1.- Laboratorio de Sistemas Inteligentes
Facultad de Ingeniería.
Universidad de Buenos Aires.
Paseo Colón 850 4to Piso. Ala Sur. (1063) Capital Federal

2.- Centro de Ingeniería del Software e Ingeniería del Conocimiento (CAPIS)
Instituto Tecnológico de Buenos Aires
Av. Madero 399. (1106) Capital Federal.

3.- Programa de Doctorado en Ciencias Informáticas
Facultad de Informática
Universidad Nacional de La Plata.
Buenos Aires

4.- Laboratorio de Sistemas Operativos y Bases de Datos.
Facultad de Ingeniería.
Universidad de Buenos Aires
Paseo Colón 850 4to Piso. Ala Sur. (1063) Capital Federal

Resumen: En este trabajo se aborda el problema de la detección de cambios de consumo de usuarios de telefonía celular fuera de lo normal, la correspondiente construcción de estructuras de datos que representen el comportamiento reciente e histórico de cada uno de los usuarios, teniendo en cuenta la información que contiene una llamada y lo complejo de la construcción de una función con tantas variables de entrada paramentrización no siempre conocida.

Palabras Clave: Detección de Fraude, Redes Neuronales, Telefonía Celular

INTRODUCCION

Cuando se inicia una llamada de celular, las celdas o switches registran que la misma se está realizando y producen información referida a este evento. Estos registros de datos son comúnmente llamados CDR's (Call Detail Records). Los CDR's contienen importante información sobre la llamada para que luego ésta pueda ser cobrada a quien corresponda. Estos registros también pueden ser usados para detectar actividad fraudulenta considerando indicadores de fraude bien estudiados. Es decir, procesando una cantidad de CDR's recientes y comparando una función de los diferentes campos tales como IMSI (International Mobile Subscriber Identity, que identifica unívocamente un usuario en una red de telefonía celular), fecha de la llamada, hora de la llamada, duración, tipo de llamada con un cierto criterio determinado [Moreau-Preneel]. Si esta función devuelve un valor que se considera fuera de los límites normales se activa una alarma, que debe ser tomada en cuenta por los analistas de fraude para constatar si realmente hubo o no actividad de mala fé. Para poder procesar estos CDR's, es necesario realizar previamente un proceso conocido en telecomunicaciones como *mediación*, en el cual se lee la información con el formato de registro en el que vienen los CDR's (el mismo puede ser de longitud variable dependiendo del tipo de llamada y del proveedor del switch) y se codifica en un nuevo formato de registro entendible por el sistema de fraude en este caso. Los sistemas existentes de detección de fraude intentan consultar secuencia de CDR's comparando alguna función de los campos con criterios fijos conocidos como *Triggers*. Un *trigger*, si es activado, envía una alarma que lleva a la investigación por parte de los analistas de fraude. Estos sistemas realizan lo que se conoce como *Análisis absoluto de CDR's* y son utilizados para detectar los extremos de la actividad fraudulenta. Para realizar un *Análisis diferencial*, se monitorean patrones de comportamiento del teléfono celular comparando sus actividades más recientes con la historia de uso del mismo; un cambio en el patrón de comportamiento es una característica sospechosa de ser un escenario fraudulento [APeCT 1996].

DESCRIPCION DEL PROBLEMA

Para poder construir un sistema de detección de fraude basado en un *Análisis diferencial* es necesario tener en cuenta varias problemáticas que se presentan, que deben ser cuidadosamente trabajadas, ellas son:

a) El problema de la construcción y mantenimiento de "perfiles de usuario"
La mayoría de los indicadores de fraude no se analizan utilizando un único CDR. Gran parte de los fraudes pueden ser solo detectados utilizando una secuencia de los mismos. En un sistema de detección de fraude diferencial; se necesita información acerca de la historia, sumado a muestras de su actividad más reciente. Un intento inicial para solucionar el problema, puede ser el extractar y codificar la información de los CDR's y almacenarla en un formato de registro determinado; para ello se necesitan dos tipos de registro: Uno que almacene la información más reciente, al que llamaremos CUP (Current User Profile) y otro con la información histórica al que llamaremos UPH (User Profile History) [Burge & Taylor]. Cuando un nuevo CDR de un determinado usuario llega para ser procesado, la entrada más vieja del registro UPH debería ser descartada y la más vieja del CUP debería ingresar al UPH. Entonces este nuevo registro codificado, debería ingresar al CUP. Esta información debe ser almacenada en una forma compacta y fácil de analizar luego por el sistema de detección de fraude. Teniendo en cuenta la cantidad de información que contiene un CDR es necesario encontrar una forma de "clasificar" estas llamadas en grupos o prototipos donde cada una debe pertenecer a un único grupo. Es decir que aquí se nos plantean varias preguntas importantes que debemos resolver: (a) ¿Qué estructura deben tener los registros CUP y UPH?, (b) ¿Cuántos grupos o prototipos deben tener los registros CUP y UPH para tener la información necesaria?, (c) ¿Cómo se puede clasificar a las llamadas en los diferentes prototipos definidos? y (d) ¿Cómo codificar las llamadas para que estas puedan "prototiparse"?

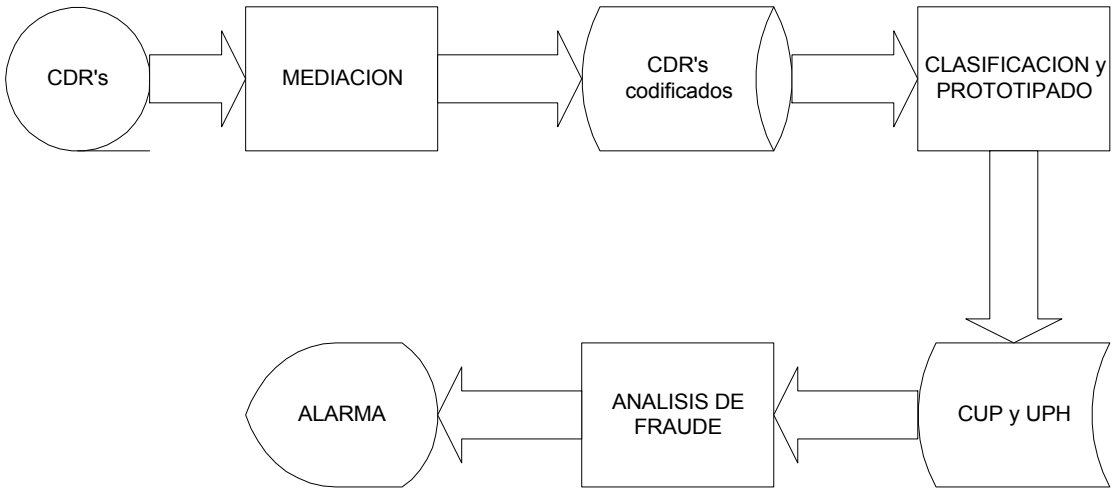
b) El problema de la detección de cambios de comportamiento
Una vez que se ha logrado construir una imagen codificada del consumo reciente e histórico de cada usuario, es necesario, encontrar la forma de analizar esta información para que detecte alguna anomalía en el consumo y dispare la alarma correspondiente. Es entonces que aquí se plantea la pregunta más importante de este trabajo: ¿Cómo se detecta el cambio en el patrón de consumo de un usuario?

c) El problema de la performance
La performance en este tipo de sistemas es de vital importancia para el éxito de la detección de fraudes. Se debe tener en cuenta que se procesarán millones de llamadas por día y que las mismas serán utilizadas para construir los perfiles de cada uno de los usuarios de la compañía, que dependiendo del tamaño de la misma, puede variar de cientos de miles a un par de millones. Es decir, que la cantidad de información a almacenar y la forma de acceder a la misma son puntos tan importantes como el sistema de detección de fraude en sí. También lo es la velocidad de procesamiento, por la cantidad de información que debe analizarse. Un sistema de detección de fraude que tarde días en analizar una poca cantidad de llamadas es totalmente obsoleto y no cumple con los objetivos y tiempos de las compañías.

Nuestro problema se enfoca entonces, no solo en la detección de cambios de consumo fuera de lo normal, sino que también con una importancia fundamental, en la construcción de estructuras de datos que representen el comportamiento reciente e histórico de cada uno de los usuarios, teniendo en cuenta la gran cantidad de información que contiene una llamada y lo complejo de la

construcción de una función con tantas variables de entrada y obviamente muy compleja y no conocida.

Representación gráfica del problema a resolver
Habiendo analizado todos los problemas que se nos plantean, podemos hacer un resumen gráfico de lo que debería ser el sistema de detección de fraude.



DESCRIPCION DE LA SOLUCION PROPUESTA

La solución que se ha propuesto y desarrollado ha tenido en cuenta cada una de las cuestiones planteadas anteriormente, intentando resolverlas del modo más eficaz y eficiente posible. A continuación se presentan cada una de las respuestas a las preguntas que surgieron en el análisis del problema. Para poder comenzar a procesar los CDR's se debe crear un formato de registro (salida de la mediación) con la siguiente información: IMSI (identifica al usuario), fecha de la llamada en formato AAAAMMDD, hora de la llamada en formato HH24MISS, duración de la llamada en formato 00000 y tipo de llamada clasificada en LOC (llamada local), NAT (llamada DDN o nacional) e INT (llamada DDI o internacional). Con esta información ya acotada a los datos necesarios, se pueden comenzar a resolver las siguientes y más importantes cuestiones utilizando como datos de entrada la salida de la mediación.

Solución a la construcción y mantenimiento de "perfiles de usuario"
La primer cuestión a resolver es determinar cómo construir los perfiles CUP y UPH. Es decir que se deben determinar los *patrones* que compondrán cada uno de estos perfiles. Los patrones deberán tener información del consumo del usuario, separando el consumo LOC (llamadas locales), NAT (llamadas nacionales o DDN) e INT (llamadas internacionales o DDI) respectivamente. Una forma interesante de construir estos patrones es utilizando redes neuronales, para discretizar el espacio de todas las llamadas de los usuarios, generando un espacio de n patrones que representen el consumo de todos los usuarios y luego generando una distribución de frecuencias por cada usuario en la cuál se represente qué probabilidad de hacer llamadas de ese patrón tiene un usuario. En resumen, cuando se construya el perfil de usuario se estará representando la distribución de frecuencia en la cuál un determinado usuario realiza un tipo de llamada determinada, mostrando esta estructura de datos el *patrón de consumo* del mismo. Las redes neuronales, entre otras ventajas, tienen la capacidad de clasificar la información en determinados patrones. En especial, las redes SOM (Self

Organizing Map) pueden tomar esta información y construir estos patrones de manera *no supervisada* por criterios de semejanza, sin saber nada a priori de los datos [Hollmen, 1996]. En nuestro caso, se pueden procesar todas las llamadas realizadas por todos los usuarios para que las redes, según la cantidad que hay de cada tipo, genere los patrones (creando grupos de semejanza) que representen a todas ellas. Para evitar ruidos en los datos, se utilizan 3 redes neuronales que generen patrones para representar a las llamadas LOC, NAT e INT respectivamente. El perfil de usuario se construye utilizando todos los patrones generados por las 3 redes. Los datos que se utilizan para representar un patrón son la hora de la llamada y la duración de la misma. Sabemos que si representamos en un eje cartesiano la hora de todas las llamadas y la duración correspondiente, obtendremos un rectángulo prácticamente lleno de puntos; la idea es obtener un gráfico en el que sólo aparezcan los puntos más representativos de todo el espacio en cuestión, esa es la tarea de las redes neuronales. Una vez obtenidos los patrones que se utilizarán para representar los perfiles de usuario, es necesario comenzar a *llenar de información* de los mismos. El procedimiento consiste en tomar la llamada a analizar, codificarla y que la red neuronal determine a qué patrón se parece la misma. Una vez obtenida esta información, se debe adaptar el perfil de usuario CUP de manera que la distribución de frecuencia muestre que el usuario tiene ahora una probabilidad mayor de realizar este tipo de llamadas. Sabiendo que el perfil de usuario tiene *K* patrones que se componen de *L* patrones LOC, *N* patrones NAT e *I* patrones INT, podemos construir un perfil representativo de la llamada procesada y luego *adaptar* el perfil CUP con dicha llamada. Si la llamada es LOC, los *N* patrones NAT y los *I* patrones INT tendrán una distribución de frecuencia igual a 0, y los *K* patrones LOC tendrán una distribución de frecuencia dada por la ecuación [Burge & Taylor, 2000]:

$$v_i = \frac{e^{-\|X-Q_j\|}}{\sum_{j=1}^{L} e^{-\|X-Q_j\|}}, \text{ dónde}$$

X: llamada a procesar codificada
v_i: probabilidad que la llamada X sea del patrón i
Q_i: patrón i generado por la red neuronal LOC.

Nótese que $\sum_{j=1}^{K} v_j = 1$.

Si la llamada fuese NAT, entonces se debe reemplazar L por *N* y la distribución de frecuencias LOC e INT serán 0; Si la llamada fuese INT, entonces se debe reemplazar *L* por *I* y la distribución de frecuencias LOC e NAT serán 0.
Entonces, podemos definir el vector representativo de la llamada *V*, de dimensión *K* como:

$V_i = v_i$, con $1 \le i \le L$
$V_i = 0$, con $L+1 \le i \le K$, cuando la llamada es LOC.

$V_i = v_i$, con $L+1 \le i \le L+N$
$V_i = 0$, con $1 \le i \le L$ y $L+N \le i \le K$, cuando la llamada es NAT.

$V_i = v_i$, con $L+N+1 \le i \le K$
$V_i = 0$, con $1 \le i \le L+N$, cuando la llamada es INT.

Ahora que tenemos el vector *V*, podemos adaptar el vector *CUP* con la información de la llamada procesada:

$CUP_i = a_{LOC} CUP_i - (1 - a_{LOC} V_i)$, con 1 ≤ i ≤ K, cuando la llamada es LOC,

$CUP_i = a_{NAT} CUP_i - (1 - a_{NAT} V_i)$, con 1 ≤ i ≤ K, cuando la llamada es NAT,

$CUP_i = a_{INT} CUP_i - (1 - a_{INT} V_i)$, con 1 ≤ i ≤ K, cuando la llamada es INT, dónde

α_{LOC}: tasa de adaptabilidad aplicada cuando la llamada X se incorpora al CUP, si X corresponde a una llamada local.
α_{NAT}: tasa de adaptabilidad aplicada cuando la llamada X se incorpora al CUP, si X corresponde a una llamada nacional.
α_{INT}: tasa de adaptabilidad aplicada cuando la llamada X se incorpora al CUP, si X corresponde a una llamada internacional.

Una vez adaptado el perfil CUP, se compara con el perfil UPH y se determina si ha habido un cambio significativo de comportamiento (motor de detección de cambios de comportamiento). Una vez realizada esta tarea, se adapta el UPH con la información del CUP, solamente si la cantidad de llamadas necesarias para cambiar el patrón histórico se han procesado:

$UPH_i = bUPH_i + (1 - b)CUP_i$, con 1 ≤ i ≤ K, dónde

β: tasa de adaptabilidad aplicada cuando el CUP se incorpora al UPH.

Solución a la detección de cambios de comportamiento
Para determinar si hubo o no cambios en el patrón de comportamiento, es necesario comparar los perfiles CUP y UPH de alguna manera y decidir si la diferencia entre los mismos es lo suficientemente grande como para lanzar una alarma. Debido a que el CUP y el UPH son dos vectores que representan distribuciones de frecuencia, se puede utilizar una distancia vectorial para comparar qué tan diferentes son. Para ello se puede utilizar la distancia Hellinger (H) cuyo valor indica la diferencia entre dos distribuciones de frecuencia [Burge, 2000]. La distancia siempre será un valor entre cero y dos, donde cero es para distribuciones iguales y dos representa ortogonalidad. El valor de H determinará qué tan diferentes deben ser las distribuciones de frecuencia CUP y UPH para lanzar una alarma. Variando este valor, habrá más o menos alarmas.

$$H = \sum_{i=1}^{K} \sqrt{CUP_i} - \sqrt{UPH_i}$$

Solución a los problemas de performance
La performance dependerá directamente del hardware donde corra el sistema de detección de fraude y cambios de comportamiento. Pero se pueden tener en cuenta varios puntos, desde el punto de vista del software para poder mejorarla. En principio, se debe trabajar lo menos posible con bases de datos relacionales y tratar de hacer todo el procesamiento utilizando archivos planos de datos, con la mínima cantidad de escrituras y lecturas de disco. Será importante la compresión de los mismos, ya que el espacio es otra restricción que se debe tener en cuenta. En nuestra solución solo se trabaja con archivos planos y se almacena un archivo por usuario con la información de las distribuciones

CUP y UPH, así como también la última llamada procesada y la cantidad total de llamadas procesadas por el sistema.

Limitaciones de la solución
Esta solución se enfoca, tal cual describimos en el análisis diferencial del consumo del usuario. Un caso que no sería detectado es aquel en el cual el usuario *siempre realiza muchas llamadas del mismo tipo con un alto consumo*, ya que su patrón de comportamiento nunca cambiaría. Es por eso que siempre se deben combinar varias soluciones para tener un sistema de detección de fraude que detecte diferentes tipos de fraude. En este caso, el análisis absoluto sería una buena solución. La otra gran limitación se centra en que los patrones son estáticos, con lo que si la forma de consumo de los usuarios de la empresa cambia completamente, será necesario re-entrenar a las redes neuronales para que determinen nuevos patrones que representen el espacio total de llamadas y volver a construir los perfiles CUP y UPH a partir de las nuevas distribuciones.

EXPERIMENTACION

Metodología utilizada
Los experimentos se dividieron en dos partes: la primera se enfocó en el entrenamiento de la red y la generación de los patrones para construir posteriormente los perfiles de usuario; la segunda prueba se enfocó en el análisis de las llamadas de los usuarios con alto consumo y el correspondiente análisis y detección de alarmas. La segunda parte de la prueba se dividió a su vez en dos experiencias diferentes: 1) actualización del perfil UPH con cada llamada (f = 1 llamada) y bajo umbral Hellinger (H) para el lanzamiento de alarmas de cambio de comportamiento; 2) actualización del perfil UPH una vez por día (f = 1 día) y alto umbral Hellinger (H).

Experimentos de generación de patrones
Se construyeron 3 redes neuronales Self Organizing Map (SOM) para la generación de los patrones para las llamadas locales (LOC), DDN (NAT) y DDI (INT) respectivamente. Cada una de las redes fue entrenada con una cantidad de llamadas representativa del consumo de los usuarios de la empresa que los mismo realizaron durante unos días en todos los horarios. Las llamadas se presentaron a las redes de manera desordenada de manera que los patrones que se generaron no fueran solamente representativos de los horarios y duraciones de las últimas llamadas. El resultado de esta experiencia definió los patrones para construir los perfiles de los usuarios. Los patrones se componen de la hora de la llamada y la duración en minutos de la misma, que lograron discretizar el espacio compuesto por todos los tipos de llamada realizadas por cualquier usuario en una cantidad fija representativa del mismo.

Experimentos de construcción de perfiles y detección de comportamientos
Una vez obtenidos los patrones que definen el espacio de todas las llamadas, se realizaron las pruebas de construcción de los perfiles de usuario a través del desarrollo de una distribución de frecuencias de cada uno de los patrones para cada perfil (CUP y UPH) y la correspondiente detección de alarmas. El proceso se basó en presentar al sistema las llamadas realizadas en un período de 3 meses por los usuarios reportados como “alto consumo”. Con cada llamada se actualizaba el perfil CUP del usuario, se comparaba con el perfil UPH obteniendo la distancia Hellinger (H) entre ambos, y si la misma superaba el umbral fijado, se lanzaba una alarma. Dependiendo del parámetro de frecuencia de actualización del perfil UPH (f), se actualizaba el UPH con el aporte del CUP según corresponda. Vale aclarar, que el proceso de construcción y actualización se hizo desde la primer llamada del usuario; en cambio la comparación y

correspondiente detección de la alarma se realizó solamente luego que la cantidad de llamadas analizadas para el usuario pasara la cantidad mínima para construir un perfil (Q_L) con la suficiente información del usuario. En el momento de ingresar la primer llamada de un usuario, se inicializaba a todos los patrones del CUP y UPH con la misma distribución de frecuencia, asumiendo que el usuario tenía la misma tendencia a realizar cualquier tipo de llamada a priori, sin información. A su vez esta experiencia se realizó dos veces: la primera actualizando el UPH con cada llamada y por consiguiente con un bajo umbral Hellinger (H) para la detección de alarmas debido a que la diferencia que se pudiera presentar entre los perfiles CUP y UPH era muy pequeña actualizando el perfil histórico con cada llamada, ya que el mismo tendía a ser igual al perfil actual. La segunda experiencia se realizó actualizando el UPH una vez por día y un umbral Hellinger (H) alto para detectar diferencias importantes que puedan ser consideradas como cambios de comportamiento.

Parámetros utilizados para la generación de patrones

Los valores utilizados para la generación de los perfiles fueron los siguientes:

- *Dimensión de la red neuronal para clasificar llamadas locales* ($N_L x M_L$) = 12x12
- *Dimensión de la red neuronal para clasificar llamadas nacionales* ($N_N x M_N$) = 8x8
- *Dimensión de la red neuronal para clasificar llamadas internac.* ($N_I x M_I$) = 6x6
- *Tasa de aprendizaje estática* (α) = 0,6
- *Distancia máxima de neurona "vecina" afectada* (D_{VMAX}) = 10

Los mismos definen la dimensión de los perfiles CUP y UPH:

- *Cantidad de patrones para clasificar las llamadas locales* (P_L) = 144
- *Cantidad de patrones para clasificar las llamadas DDN* (P_N) = 64
- *Cantidad de patrones para clasificar las llamadas internacionales* (P_I) = 36
- *Dimensión de los perfiles CUP y UPH* (K) = 244

Parámetros utilizados para la construcción de perfiles y detección de cambios de comportamiento

Los valores utilizados para la construcción de perfiles y detección de alarmas fueron los siguientes:

Experiencia 1:

- *Factor de adaptabilidad de llamadas locales en el CUP* (α_{LOC}) = 0,8
- *Factor de adaptabilidad de llamadas locales en el CUP* (α_{NAT}) = 0,8
- *Factor de adaptabilidad de llamadas locales en el CUP* (α_{INT}) = 0,8
- *Factor de adaptabilidad de UPH* (β) = 0,9
- *Sensibilidad del sistema - Umbral Hellinger* (H) = 0,3
- *Frecuencia de actualización del UPH* (f) = 1 llamada.

Experiencia 2:

- *Factor de adaptabilidad de llamadas locales en el CUP* (α_{LOC}) = 0,8
- *Factor de adaptabilidad de llamadas locales en el CUP* (α_{NAT}) = 0,9
- *Factor de adaptabilidad de llamadas locales en el CUP* (α_{INT}) = 0,9
- *Factor de adaptabilidad de UPH* (β) = 0,6
- *Sensibilidad del sistema - Umbral Hellinger* (H) = 0,75
- *Frecuencia de actualización del UPH* (f) = 1 día.

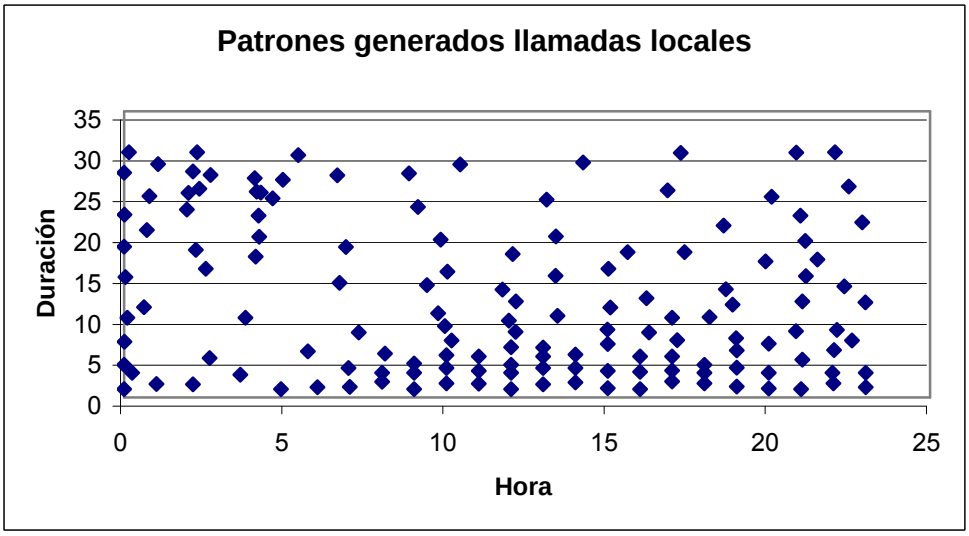
Valores comunes a ambas pruebas:

- *Cantidad mínima de llamadas antes de comparar perfiles* (Q_L) = 100 llamadas.

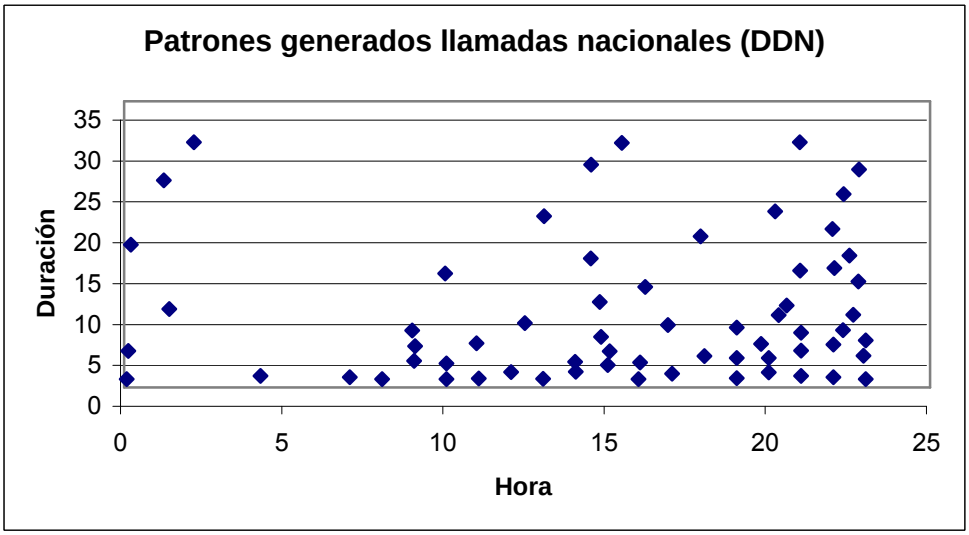
Resultados

Generación de patrones

En esta sección se presentan los resultados obtenidos luego del entrenamiento de las 3 redes neuronales. Es decir, que los resultados muestran cada uno de los patrones que las redes determinaron como más representativos del espacio de todas las llamadas de todos los usuarios. Se presentan 3 gráficos (uno por cada red) en el que se muestra los patrones generados. En el eje X se muestra la hora de la llamada y en el eje Y la duración expresada en minutos. Cada uno de los puntos representados corresponde a un patrón elegido por la red como representativo de la muestra. En el gráfico de la red neuronal local, se muestran 144 patrones; en el de la red DDN, 64 y en el de la red DDI 36.

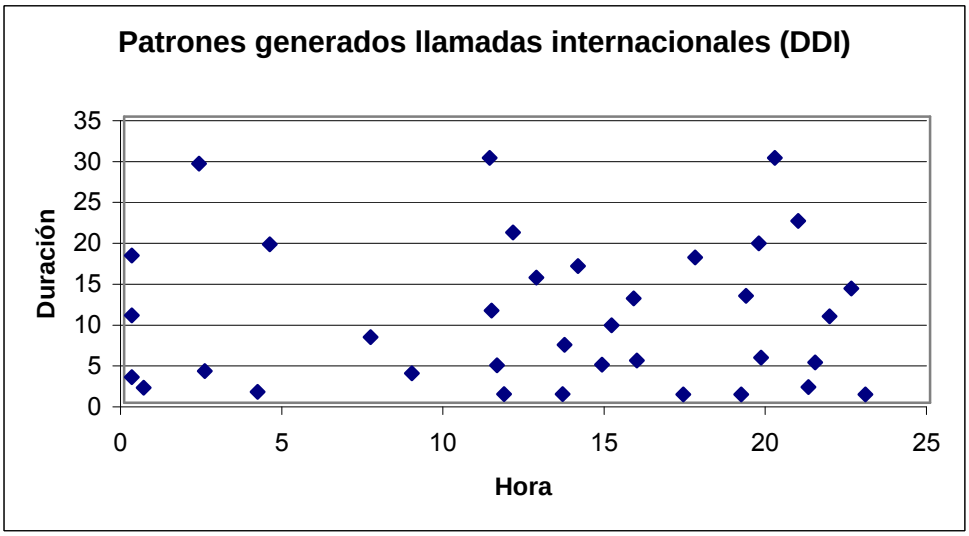


Se observa en el gráfico los 144 patrones generados luego del entrenamiento de la red neuronal de llamadas locales. A simple vista se puede notar que hay una concentración mayor de patrones en la banda horaria de las 8 hs. a las 20 hs y una duración entre 0 y 5 minutos. Esto denota que la mayoría de las llamadas locales realizadas por los clientes de esta empresa ocurren en estos horarios con los promedios de duración indicados.



Se observa en el gráfico los 64 patrones generados luego del entrenamiento de la red neuronal de llamadas nacionales. Aquí también se observa una concentración de patrones, pero más desplazada hacia la banda horaria de las 15 a las 22 con duraciones que oscilan entre los 0 y 7 minutos. También se observa que prácticamente no hay patrones generados para la madrugada, con lo cual se

puede concluir que la mayoría de los usuarios de la empresa analizada no realizan llamadas DDN en horas muy tempranas.



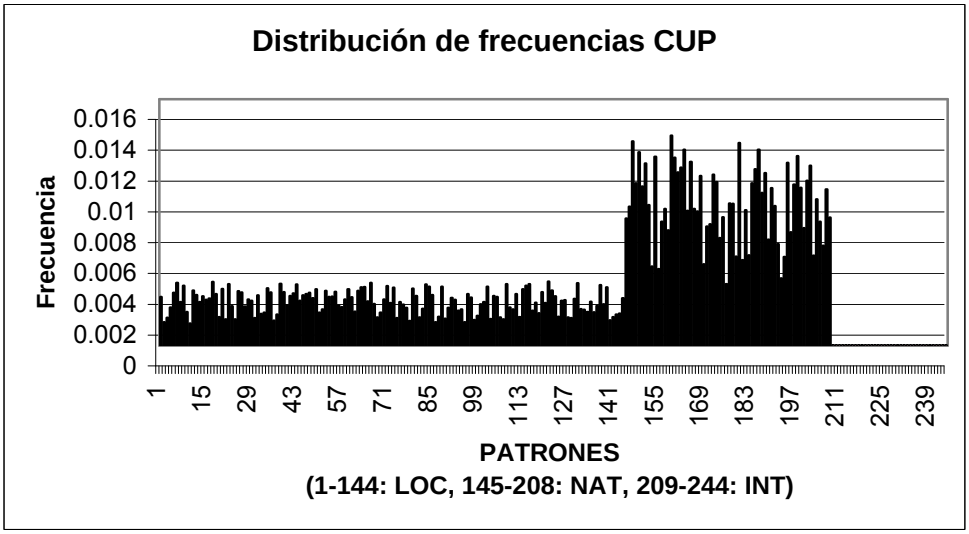
Se observa en el gráfico los 36 patrones generados luego del entrenamiento de la red neuronal de llamadas internacionales. Aquí la distribución es un poco más aleatoria, pero la duración de las llamadas "elegidas" como patrones tienden a tener una duración mayor (entre 7 y 10 minutos).

Construcción de perfiles y detección de cambios de comportamiento

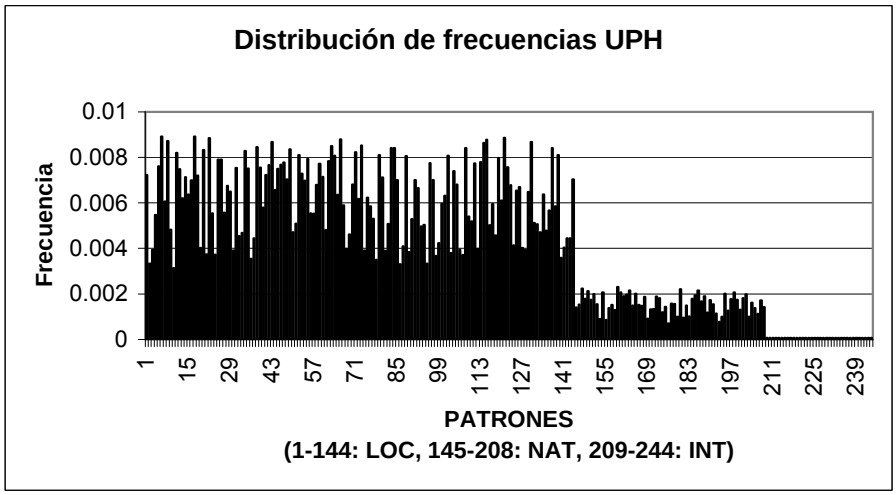
En esta sección se presentan los resultados obtenidos luego de la construcción de los perfiles y la detección de las correspondientes alarmas para cada una de las 2 experiencias realizadas. Se muestran gráficos con una descripción de los perfiles CUP y UPH de uno algunos casos en el momento que se lanzó una alarma.

En el eje X se presentan los 244 patrones (144 LOC, 64 NAT y 36 INT) y en el eje Y la distribución de frecuencias de cada uno de los patrones para el usuario analizado en el momento que fue lanzada la alarma (la sumatoria de todas está normalizada a 1). También se realizará una explicación general del por qué de alarmas lanzadas por el sistema y un análisis de la confiabilidad y veracidad de las mismas basadas en el detalle de llamadas de cada usuario.

Experiencia 1 (Actualización UPH con cada llamada, alta sensibilidad con bajo Umbral Hellinger):



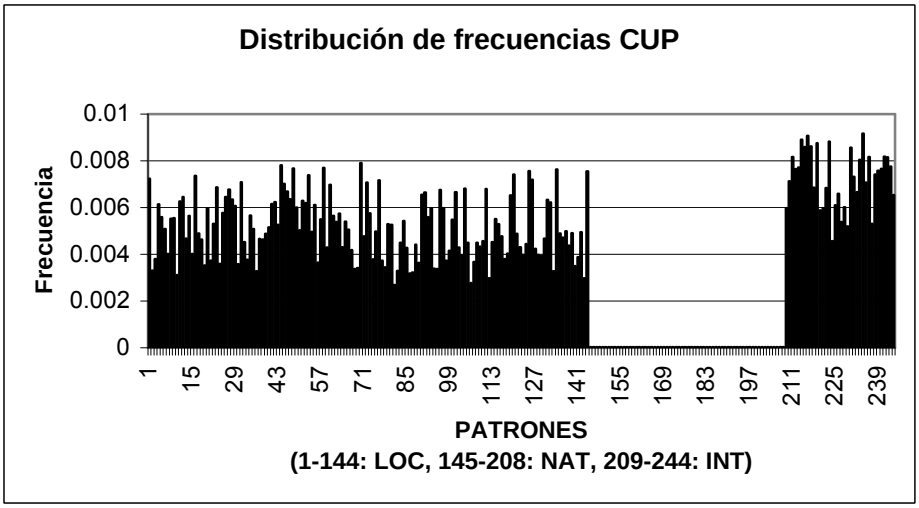
El gráfico muestra el CUP de un usuario en el momento que se lanzó una alarma. Se puede observar en el mismo que la distribución de frecuencias indica una mayor tendencia a realizar llamadas DDN (patrones 145 a 208).



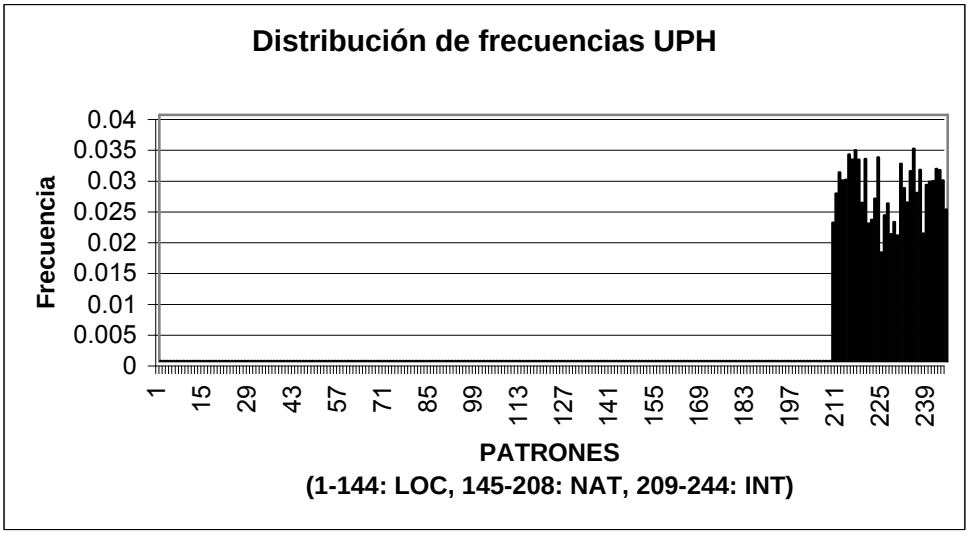
El gráfico muestra el UPH del mismo usuario en el momento que se lanzó la alarma. Se puede observar en el mismo que la distribución de frecuencias indica una mayor tendencia a realizar llamadas locales (patrones 1 a 144).

En consecuencia la diferencia entre ambas distribuciones de frecuencias definida por la distancia Hellinger (H) es igual a: 0,30081. Analizando el detalle de llamadas de este usuario desde fechas anteriores al lanzamiento de la alarma hasta la aparición de la misma, se observa que la alarma se produjo debido a que el usuario hizo una llamada DDN por primera vez desde que se procesaron sus llamadas. Es decir, que su patrón de comportamiento histórico no hacía creer que iba a realizar llamadas de este tipo. Sin embargo al realizarlas, el sistema detectó este cambio y generó la correspondiente alarma. También muestran estos resultados que al haber realizado la experiencia con tan alta sensibilidad, una llamada diferente puede indicar un cambio de comportamiento que conduce a una alarma. El total de alarmas lanzadas luego de analizar los 60 usuarios fue de 88, de las cuales 33 corresponden a diferentes casos. Esto se debe a que una vez lanzada una alarma para un usuario, las siguientes llamadas del mismo vuelven a lanzar alarmas hasta tanto el UPH no se adapta definitivamente al cambio de comportamiento. La mayoría de las mismas siguen el patrón del caso graficado en el cual una llamada diferente al patrón normal de comportamiento alcanza para que el sistema defina al usuario como sospechoso.

Experiencia 2 (Actualización UPH una vez por día, moderada sensibilidad con Umbral Hellinger):



El gráfico muestra el CUP de un usuario en el momento que se lanzó una alarma. Se puede observar en el mismo que la distribución de frecuencias indica una tendencia a realizar llamadas locales (patrones 1 a 144) e internacionales (patrones 209 a 244).



El gráfico muestra el UPH del mismo usuario en el momento que se lanzó la alarma. Se puede observar en el mismo que la distribución de frecuencias indica una tendencia a realizar llamadas internacionales solamente (patrones 209 a 244). En consecuencia la diferencia entre ambas distribuciones de frecuencias definida por la distancia Hellinger (H) es igual a: 0,82815. Analizando el detalle de llamadas de este usuario desde fechas anteriores al lanzamiento de la alarma hasta la aparición de la misma, se observa que la alarma se produjo debido a que el usuario solamente realizaba llamadas internacionales, pero un momento dado comenzó a realizar llamadas locales. Cuando la cantidad de llamadas locales modificó el CUP de la manera que se muestra en el gráfico, se lanzó la alarma. Este es un caso curioso ya que seguramente esta alarma no es indicadora de fraude si el usuario paga su factura de llamadas internacionales. Pero sí es indicadora de un sensible cambio de comportamiento en su patrón de consumo, y este sistema busca exactamente eso. El total de alarmas lanzadas luego de analizar los 60 usuarios fue de 64, de las cuales 14 corresponden a diferentes casos. Esto se debe a que una vez lanzada una alarma para un usuario, las siguientes llamadas del mismo vuelven a lanzar alarmas hasta tanto el UPH no se adapta definitivamente al cambio de comportamiento. Aquí este fenómeno se acentúa debido a que recién cuando se procesan llamadas del día siguiente se actualiza el UPH. La mayoría de las mismas siguen el patrón del caso graficado en el cual debe haber varias llamadas fuera del patrón de comportamiento para que el sistema encuentre al usuario sospechoso. Esto es mucho más satisfactorio que lo obtenido en la experiencia 1 en la cual la alta sensibilidad mostraba usuarios como sospechosos simplemente por el hecho de haber realizado *una sola llamada* diferente.

CONCLUSIONES

Los resultados obtenidos fueron satisfactorios, en el sentido que pudieron determinar cambios de comportamiento en los usuarios analizados. Si bien el cambio de comportamiento no implica necesariamente actividad fraudulenta, logra acotar la investigación de los analistas de fraude a este grupo de usuarios. Utilizando luego otro tipo de técnicas [ASPeCT, 1997] se puede llegar a obtener con un alto grado de certeza, usuarios que estén utilizando sus teléfonos celulares “deslealmente”. Además, las experiencias han servido para encontrar usuarios que efectivamente cambiaron su comportamiento, pero de manera inversa, es decir, que eran usuarios con alto consumo INT y luego comenzaron a realizar llamadas locales. Comercialmente puede ser un dato interesante evaluar a

este tipo de usuarios ya que por algún motivo en particular decidieron no utilizar más su teléfono celular para realizar llamadas internacionales y puede servir para sacar conclusiones y crear nuevos planes de tarifa basado en estas situaciones. También queda demostrado con las experiencias realizadas que el análisis diferencial provee mucha más información que el análisis absoluto, el cual solo puede detectar picos de consumo y no puede describir al usuario en cuestión. Como última conclusión se puede decir que las redes neuronales han demostrado ser una excelente herramienta para la clasificación de las llamadas y construcción de perfiles de usuario ya que representaron fielmente y eficazmente el comportamiento de los mismos.

BIBLIOGRAFIA

ASPeCT, 1996. *Definition of Fraud Detection Concepts, Deliverable D06*. 47 páginas.

ASPeCT, 1997. *Fraud Management tools: First Prototype, Deliverable D08.* 31 páginas.

Burge P, Shawe-Taylor J. *Frameworks for Fraud Detection in Mobile Telecommunications Networks,* Department of Computer Science Royal Holloway, University of London.

Burge P, Shawe-Taylor J. *Detecting Cellular Fraud Using Adaptive Prototypes,* Department of Computer Science Royal Holloway, University of London.

Hilera J. R., Martínez V. 2000. *Redes Neuronales Artificiales: Fundamentos, modelos y aplicaciones,* RA-MA Editorial, Madrid.

Hollmén J. 2000. *User profiling and classification for fraud detection in mobile communications network,* Helsinki University of Technology Department of Computer Science and Engineering Laboratory of Computer and Information Science.

Fawcett T., Provost F. 1997. *Adaptive Fraud Detection,* NYNEX Science and Technology.

Frank R. J., Hunt S. P., Davey N. *Aplications of Neural Networks to Telecommunications Systems,* Department of Computer Science, University of Hertfordshire.

Moreau Y., Vandewalle J., 1997. *Fraud Detection in Mobile Communications using Supervised Neural Networks,* Departement Elektrotechniek Katholieke Universiteit Leuven.

Moreau Y., Vandewalle J., 1997. *Fraud Detection in Mobile Communications using Supervised Neural Networks: A First Prototype,* Departement Elektrotechniek Katholieke Universiteit Leuven.

Moreau Y., Burge P.. *Novel Techniques for Fraud Detection in Mobile Telecommunication Networks*.

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.

Hollmen J., Tresp V. *Call-based Fraud Detection in Mobile Communication Networks using a Hierarchical Regime-Switching Model,* Helsinki University of Technology Department of Computer Science and Engineering Laboratory of Computer and Information Science. Siemens A. G., Corporate Technology.

Weiss G., Eddy J., Weiss S, 1998. *Intelligent Telecommunication technologies,* Network & Computing Services. AT & T Labs. AT & T Corporation. United States.

Hollmen J, 1996. *Process Modeling using the Self-Organizing Map*, Helsinki University of Technology Department of Computer Science.

Beveridge M., 1996. *Self Organizing Maps*. http://www.dcs.napier.ac.uk/hci/martin/msc/node6.html