Sistema para la detección de patrones de fraude en las redes de telecomunicaciones

Por MsC. Carlos A. Rodríguez López, Ing. Samuel Montejo Sánchez
Profesores Universidad Central de Las Villas
crodrigz@fie.uclv.edu.cu, montejo@fie.uclv.edu.cu

Introducción

El fraude en las telecomunicaciones causa significativas pérdidas financieras a las compañías telefónicas, afecta la esfera de la mercadotecnia y deteriora, en muchos casos, los servicios. Estas empresas realizan grandes esfuerzos por controlarlo, pero su carácter dinámico imposibilita su erradicación total. Por esta razón, un mecanismo que permita la detección de patrones de fraude constituye un requisito indispensable para cualquier operador de telecomunicaciones.

La descripción de un sistema que permita detectar patrones de fraude, en un tiempo relativamente corto, y que sea capaz de aprender de las nuevas reglas que le introduzcan los expertos y las nuevas bases de casos a las que se enfrente, es el principal objetivo de este trabajo.

El fraude en las telecomunicaciones

Fraude es el acto deliberado de usar servicios y recursos evadiendo el pago de los mismos. Cualquier interacción donde un actor presta un servicio y otro actor paga por este, puede ser una fuente potencial de fraude[1].

Este fenómeno causa pérdidas millonarias y se produce en todas las administraciones telefónicas del mundo. En 1998, se estimaban pérdidas en las telecomunicaciones a nivel mundial de 12 mil millones de dólares por año; de ellos, 3 mil millones correspondían a telefonía celular[2]. Ya en el año 2001, para los operadores de GSM representaban del 3 al 5 por ciento de sus ingresos anuales y las pérdidas globales estimadas para el 2002 alcanzaban los 30 mil millones de dólares[3]. En ese mismo año las pérdidas evaluadas en Cuba demostraron que nuestro país no está exento de estas afectaciones.

El fraude se manifiesta de diferentes maneras, desde el empleado que utiliza para su provecho los servicios de la empresa para la cual labora, hasta las bandas internacionales que hacen de las estafas su profesión.

A pesar de los ingentes esfuerzos de las empresas no ha sido posible controlarlo totalmente, pues el problema evoluciona aceleradamente. Los defraudadores no sólo se motivan por dinero, sino también por la necesidad del anonimato para enmascarar otros crímenes, y a veces sólo por desafiar el sistema.

Conocer la naturaleza del fraude y el modo en que actúan los defraudadores es la clave para combatirlo. Existen numerosos métodos para cometer fraude, pero pueden agruparse en dependencia de los rasgos que los clasifican en tipos o modalidades específicas de fraude. El conocimiento de dichos rasgos constituye un primer paso para poder detectarlos.

Los tipos de fraude principales pueden clasificarse en:

Quienes lo cometen: Fraude interno Fraude externo	**Motivos:** Fraude con ingresos Fraude sin ingresos
Métodos genéricos: Suscripción *Surfing* Fantasma Arreglo de cuentas Abuso de la información *Hacking*	**Características técnicas:** Desvío de líneas En teléfonos públicos *Call back* Clonación *By pass* *Refilling*

Son muchos los aspectos relacionados con el fraude, pero basta con operar una red de telecomunicaciones o usar un servicio, para convertirse en una víctima. El fraude prevalece en las redes fijas y móviles de todas las tecnologías. Típicamente el servicio más avanzado es susceptible a las nuevas modalidades del fraude, así unas vulnerabilidades son cerradas por los operadores de la red, mientras otras nuevas son encontradas por los defraudadores.

¿Entonces qué hacer? ¿Puede evitarse el fraude? La cuestión es discutible, pero es un hecho que las personas siempre intentan aprovecharse de los puntos débiles de una red de telecomunicaciones. Sin embargo, con la presencia de un sistema de gestión de fraude que permita la detección de anomalías de la forma más cercana posible al momento de ocurrencia, el fenómeno puede ser controlado.

Sistema propuesto

La fuente de información, las características del procesamiento y el empleo de perfiles de usuario son puntos imprescindibles del diseño. Con este sistema se pretende, a través de una o varias fuentes de información, prevenir y contener los fraudes o un gran por ciento de los mismos. Estas fuentes de información pueden ser CDRs —*Call Detailed Records*—, flujos de señalización No.7, y registro de intercambios asociados a *roaming*.

De todas estas fuentes una de las más completas es la explotación del sistema de señalización No. 7, que brinda una visión centralizada y totalizadora de la red, pues constituye el centro de los servicios proporcionados por las redes de telecomunicaciones actuales.

A través de sondas se adquiere la información necesaria de los flujos de SS7 y se almacena en bases de datos. Luego esta información es filtrada considerando algunos atributos que permiten clasificarla en flujos. Estos flujos de información son procesados por un sistema experto con el empleo de reglas predefinidas para el procesamiento y establecimiento de umbrales, y arroja alertas —en caso de detectarse situaciones que indiquen posibles fraudes o situaciones atípicas—. El sistema ofrece registros (CDRs) más completos en los cuales quedan almacenados los detalles de los fraudes detectados y la incorporación de listas negras donde se guardan los números sospechosos y los descubiertos, así como la creación y actualización de perfiles de usuarios (PU), que permitirán estudiar y conocer a los clientes, y detectar a tiempo cualquier cambio en su comportamiento.

Al sistema lo caracterizan funcionalidad, escalabilidad, fiabilidad, rendimiento, costo, repercusión y beneficios, modularidad y velocidad de procesamiento.

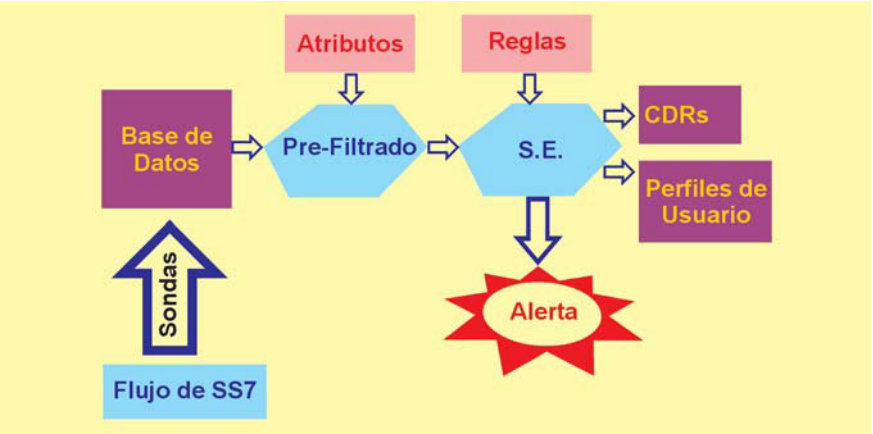

Figura 1 Estructura del sistema

El prefiltrado y los atributos

Como todos los datos no son empleados en el procesamiento, es necesaria una etapa para seleccionar, clasificar y organizar la información en flujos, para lo cual se considera la presencia de algunos atributos. Esto ayuda al uso eficiente de la memoria y se beneficia la velocidad de procesamiento.

Los atributos permiten clasificar la información, pues todos los análisis de los tipos de fraude no incluyen las mismas variables. De este modo, cada análisis sólo debe recibir como entradas punteros a los valores de los atributos que intervienen.

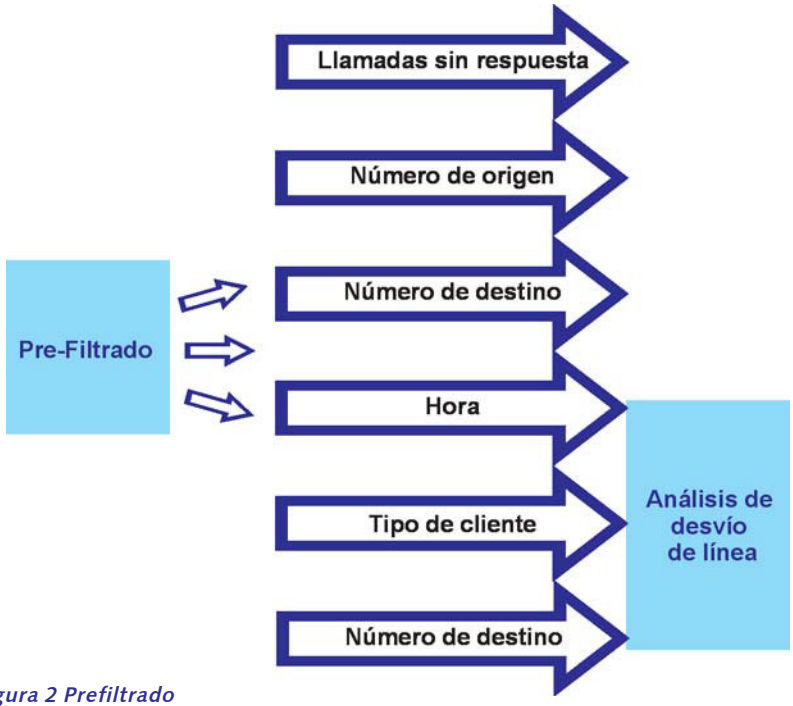

Figura 2 Prefiltrado

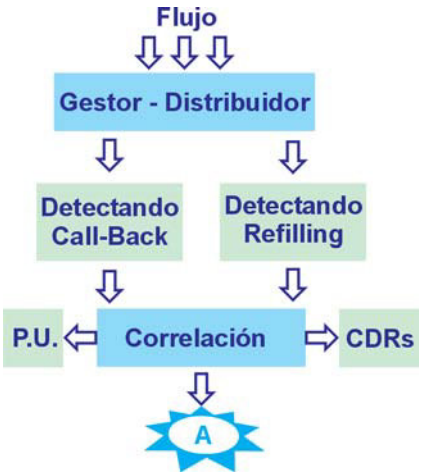

Figura 3 Arquitectura distribuida

Análisis de la potencia de cálculo

Para realizar el análisis de la potencia de cálculo, es importante mencionar que el sistema propuesto recibe y genera gran cantidad de información con elevadas exigencias de velocidad de procesamiento y de precisión en los resultados, esto puede desbordar la capacidad de análisis y procesamiento de un PC común.

La solución tradicional a este problema ha sido acudir a las supercomputadoras —máquinas de compleja gestión y extremadamente caras—. El empleo de *Clusters* de computadoras podría ser la solución. Sin embargo, a continuación se propone una variante más eficiente, económica y específica para el problema en cuestión.

Un servidor central, que distribuye los flujos de información necesarios a cada uno de los ordenadores que se encargan de realizar los análisis independientes —en cada análisis sólo influyen algunas de las variables—, y algunos son realizados con datos previamente almacenados, mientras que otros se realizan en tiempo real, por ejemplo, la detección de *refilling* se basa principalmente en los PU, mientras que la de *call back* en la aparición de muchas llamadas sin completamiento a un mismo destino. Por lo tanto, un ordenador puede encargarse de la actualización y procesamiento de los PU y entonces detectar *refilling*; otro del filtrado de grupo y detectar *by pass*; otro del filtrado individual y detectar llamadas excesivamente largas, o desde oficinas a altas horas de la noche.

Posteriormente, el servidor máster correlaciona los resultados recibidos de los ordenadores subordinados, a través del empleo del Id de llamada. Este paso es muy importante pues una llamada detectada como sospechosa por su larga duración y la hora de realización, es correlacionada con el historial del número de origen en el que se determina que es un usuario conectado a Internet.

Muchas llamadas sospechosas son detectadas por sistemas menos complejos y exigentes, sin embargo, la detección de estas irregularidades ocurre después de que se ha cometido el fraude y, por lo tanto, las pérdidas ya se han producido. La efectividad de un sistema que combate el fraude se basa en la oportuna rapidez con que pueda descubrirse el hecho. Por lo tanto, se imponen las interrogantes ¿cómo conseguir que el sistema tenga la inteligencia necesaria para detectar estos patrones? y ¿cómo conservar los conocimientos de los especialistas?

Sistema de inferencia borroso

Se desea proporcionar la tecnología analítica que permita a los operadores aprender del pasado para comprender qué es lo que sucede en el presente y de esta manera anticipar el futuro. Mientras que los operadores varían, comparten la necesidad de analizar

gran flujo de información y de poseer un sistema que posibilite que perdure su experiencia.

Los sistemas encargados de la detección de fraude analizan datos en tiempo real con la finalidad de advertir las anomalías, predecir las tendencias y controlar la corrección. El monitoreo de posibles fraudes no requiere de minuciosa observación de los datos por parte de los expertos. La tecnología y la experiencia están disponibles para implementar soluciones de monitoreo de situaciones anómalas, desarrollar diagnóstico y diseñar reglas para combatir dicha situación.

Análisis absoluto

La evolución de los sistemas de detección de fraudes telefónicos se inicia con la comparación de la información contenida en los registros CDR con criterios fijos conocidos como niveles o umbrales. Este tipo de análisis es conocido como análisis absoluto y es útil para la detección de parámetros generales de la actividad fraudulenta en una red.

Análisis diferencial

En el análisis diferencial, el sistema realiza el seguimiento de los patrones de comportamiento de la red, compara las últimas actividades telefónicas con su comportamiento histórico y genera una alarma en el sistema cuando se presentan cambios significativos durante un corto período.

Técnicas de inteligencia artificial

Los métodos convencionales de detección de fraude, basados en simples umbrales, son usados para combatir la mayoría de los fraudes. Sin embargo, algunos problemas permanecen sin solución ante ellos, al desplazarse de su comportamiento habitual, escapando así de la rigidez de sus filtros. Las técnicas de inteligencia artificial (IA), pueden ser exitosamente usadas en la lucha contra estos problemas. El término inteligencia cubre muchas habilidades conocidas, incluyendo la capacidad de solucionar problemas, de aprender y de entender métodos de solución.

La creación de un sistema experto (SE), con reglas de inferencia diseñadas por expertos, que tengan en cuenta la incertidumbre de los datos procesados, parece ser una de las mejores variantes para el análisis de la detección del fraude.

Investigación en SE

Las categorías básicas de la investigación en sistemas basados en el conocimiento incluyen representación del conocimiento, uso del mismo —o solución de problemas— y su adquisición —el aprendizaje—.

Una aplicación importante de la investigación de los SE implica los métodos para razonar con datos y conocimientos inciertos; uno de los más adoptados se llama “lógica difusa” o “razonamiento borroso”.

Lógica difusa

La lógica difusa es una nueva forma de solucionar algunas de las debilidades de los SE, especialmente cuando se está ante problemas donde reina la incertidumbre o los términos medios. Mientras que los programas que se usan en las computadoras devuelven respuestas precisas como sí o no, los programas que utilizan la lógica difusa pueden usar valores entre 0 y 1 y, de esta manera, se asemejan más a la lógica humana. Es precisamente esta habilidad de la lógica difusa lo que la hace una herramienta útil para desarrollar aplicaciones para la toma de decisiones en casos donde se cuenta con datos imprecisos o donde los problemas tienen más de una solución. En la detección del fraude los datos son imprecisos y la interpretación de la información debe variar en dependencia de la hora del día o la época del año. En ocasiones pueden encontrarse patrones de fraude sin que estén presentes —falsos positivos—, por lo tanto, las medidas en la toma de decisiones no

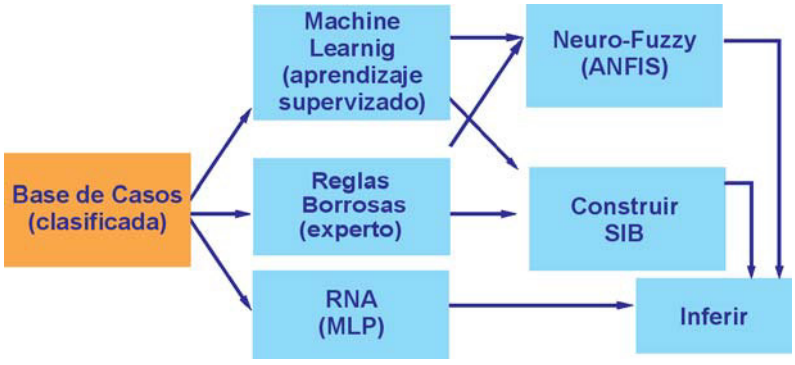

Figura 4 Técnicas para la clasificación

pueden ser rígidas porque pueden provocar afectaciones en la calidad de los servicios.

Técnica a emplear

Debe ser usada una técnica de IA o una combinación de estas técnicas; pero ¿qué procedimiento debe emplearse? Para la clasificación los métodos de mejores resultados son:

1.Sistema de aprendizaje automatizado —*Machine Learning*— que, a través de la base de casos (BC), crea reglas definidas por los casos analizados. Luego pueden pasarse estas reglas por el sistema ANFIS —*Adaptive-Network-based Fuzzy Inference System*— o partir directamente a la construcción de un sistema de inferencia borroso (SIB) y, por último, a inferir.

2.Reglas borrosas creadas por expertos, construir un SIB y luego inferir, o mejorar los términos lingüísticos del SIB a través del sistema ANFIS y entonces inferir.

3.Red neuronal artificial, específicamente una MLP —*Multi-Layer Perceptron*— a través de los datos contenidos en la BC clasificada, la red es capaz de aprender a detectar irregularidades y correlaciones en los mismos, para posteriormente inferir las respuestas en dependencia de los nuevos datos de entrada.

De todas las técnicas, la de mayor número de seguidores y mejores resultados, en la clasificación de patrones, ha sido la tercera variante (RNA-MLP); pero esta al igual que la primera variante, necesita una buena base de casos muy bien clasificada. En caso de no contarse con esta BC y decidir comprarla a un país que la posea, el resultado no sería óptimo pues a pesar de ser universales los tipos de fraude, las modalidades son propias de cada región. Por lo tanto, en tal caso la variante más adecuada es la segunda: a través de reglas borrosas creadas por expertos construir un SIB.

Bases de casos para el aprendizaje

Las BC que deben emplearse para el estudio y la caracterización del fraude, deben cumplir dos condiciones fundamentales: una, ser lo suficientemente amplias para permitir la acertada caracterización del fenómeno y atenuar la influencia de cualquier anomalía temporal; y la otra, estar actualizadas para garantizar que la caracterización evolucione al ritmo del fenómeno.

Sistema de inferencia borroso

Aunque un SE consiste fundamentalmente en una base de conocimiento y un motor de inferencia, otras dos características deben mencionarse: la explicación de la línea del razonamiento y el razonamiento con incertidumbre.

Cuando la respuesta a un problema es cuestionable, debe conocerse el análisis razonado. Si este parece probable, habrá que creer la respuesta. Las explicaciones pueden ser generadas rastreando la línea del razonamiento usada por el motor.

Para tratar el conocimiento incierto, una regla puede tener asociado a ella un factor de confianza o un peso. El conjunto de métodos para usar el conocimiento incierto conjuntamente con datos inciertos en el proceso del razonamiento se denomina razonamiento con incertidumbre —para razonar con incertidumbre uno de los métodos más importantes se llama lógica difusa y los sistemas que los utilizan se conocen como sistemas difusos o sistemas borrosos—.

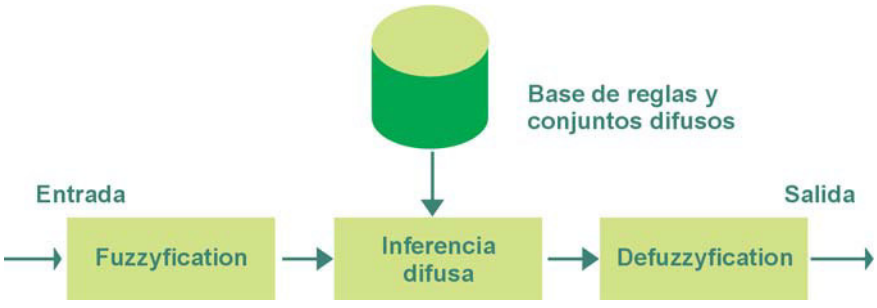

Figura 5 Sistema de inferencia borroso

Reglas

Las reglas relacionan las variables de entrada con las de salida, a través del empleo de sus etiquetas. Por lo tanto, el buen desempeño del uso de estas reglas depende de la cantidad de etiquetas empleadas para el control de la variable y lo bien definidas que estén las funciones de pertenencia de cada conjunto borroso.

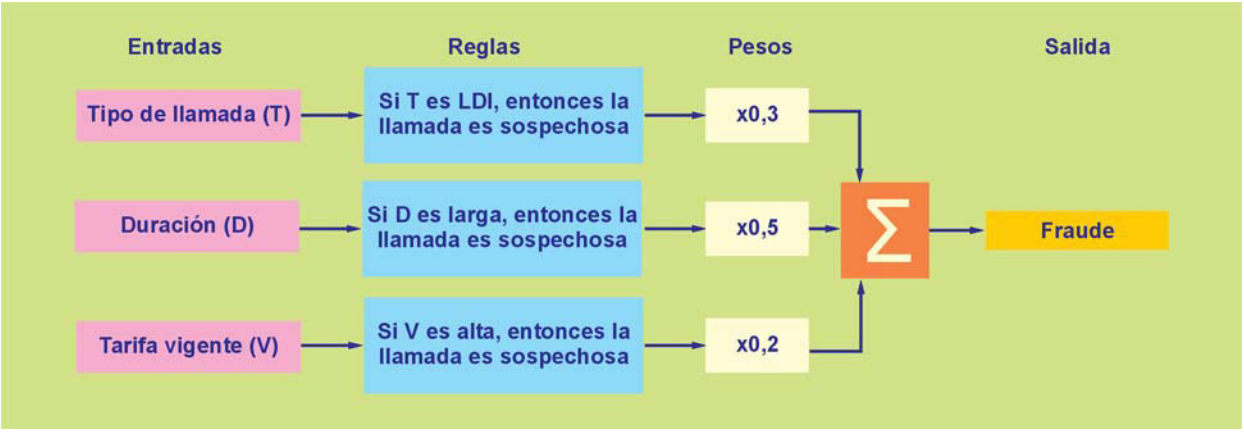

Figura 6 Análisis con reglas de inferencia

La correcta selección de las variables representativas para cada análisis y la adecuada ponderación de cada una de las reglas, en dependencia de su peso en el análisis, son otros factores imprescindibles para el uso de las reglas, en cada análisis.

El correcto tratamiento de estos dos factores evita las demoras innecesarias en el procesamiento, evade la aparición de algunos falsos positivos y disminuye el número de llamadas sospechosas que escapan al sistema.

Como puede apreciarse aparece una nueva variable —tarifa vigente— que permitirá al sistema adaptarse rápidamente a posibles cambios en el comportamiento de los clientes. Si se producen rebajas en las tarifas, el número de llamadas comenzará a aumentar y estas serán ahora de mayor duración, y no implicaría fraude. Con el uso de esta variable, asignada en la interfaz del sistema, pueden controlarse situaciones como la descrita.

Otra solución para valorar sería el tratamiento diferenciado en días festivos, días conmemorativos y ciertos períodos del año, pues en esos días el comportamiento de los clientes evidentemente varía, lo que podría provocar la generación continua de alertas por el sistema y la continua aparición de llamadas evaluadas como fraudulentas, cuando en realidad no lo son. Además, a través de filtros de grupo se creará una BC y, de comprobarse que un comportamiento antes atípico ahora pasa a común, entonces debe generarse una alerta especial para que se evalúen las modificaciones necesarias de las ponderaciones y las reglas. Este tratamiento es vital para el trabajo con los PU pues la conducta de los usuarios puede cambiar y su influencia en su historial ser poco significativa.

Luego de demostrarse la eficacia del sistema pueden ser empleadas las técnicas de redes neuronales y entrenar un MLP —*Multi-Layer Perceptron*— a través de una BC clasificada por el SIB. Una vez organizada la BC a emplear para la caracterización del fraude debe ser seleccionado un 75% para el entrenamiento del MLP y reservado para las pruebas un 25%. Esta solución híbrida puede ofrecer mejores resultados que los anteriormente obtenidos.

Conclusiones

Este artículo presenta un sistema para la detección del fraude en las redes de telecomunicaciones, fundamentalmente en telefonía, utiliza como principal fuente de información la red de señalización No.7. Luego de mostrar los tipos de fraudes más comunes y relevantes se realiza un análisis de las características más importantes a considerar en el diseño de un sistema eficiente para su rápida detección, y abarca el filtrado, la potencia de cálculo, la arquitectura a emplear y algunos de los atributos a seleccionar para ser procesados, entre otros. El principal aporte está en valerse de técnicas de IA para procesar la información y obtener los resultados. En este punto se demuestran las ventajas de estas técnicas sobre el análisis absoluto y el análisis diferencial. Se sugiere la utilización de un sistema experto que aplique lógica difusa, para fortalecer el tratamiento con términos medios o con incertidumbre. Luego de recorrer las características de las bases de casos que pueden ser utilizadas y del sistema de inferencia borroso, el trabajo culmina con la muestra de una regla y los factores que deben tenerse en cuenta para la elaboración de las mismas.

Notas

[1]Kvarnström, H. *Combining Fraud and Intrusion Detection* (2000). Disponible en: http://www.ce.chalmers.se/edu/course/EDA262/04 report_combining_fraud_and_intrusion_ detection.pdf (Consultado: febrero, 2004).

[2]Paniagua, G. *El fraude en las telecomunicaciones* (2000). Disponible en: http://www.grupoice.com/esp/cencon/docs/art_noticias/articulos/articulos3.htm (Consultado: febrero, 2004).

[3]Nortel Networks. Opiniones de Nortel Networks Fraud Solutions, 2001.

Bibliografía

Botía, J. A. "Sistemas Difusos". Departamento de Ingeniería de la Información y las Comunicaciones, Universidad de Murcia. Disponible en: http://www.angelfire.com/ia3/angel82/otros/Modelado_Difuso_4.pdf (Consultado: enero, 2004).

Cerebrus Solutions Limited. "Fraud Primer" (2002). Disponible en: http://www.cerebrussolutions.com/pdf/Fraud_Primer-Nov02.pdf (Consultado: enero, 2004).

Cerebrus Solutions Limited. "Neural Network Primer" (2002). Disponible en: http://www.cerebrussolutions.com/pdf/Neural_Network_Primer_Nov02.pdf (Consultado: mayo, 2004).

Dempsey, G. "The Changing Face of Fraud: Phone Fraud" (1999). Disponible en: http://www.aic.gov.au/conferences/outlook99/dempsey.pdf (Consultado: abril, 2004).

Iglesias, C. A. "Aplicaciones de Sistemas Inteligentes a las Telecomunicaciones" (2000). Disponible en: http://www.gsi.dit.upm.es/~cif/cursos/ssii/aplic4.pdf (Consultado: marzo, 2004).

Marín, R. "¿Qué es el Razonamiento Temporal Borroso?". Disponible en: http://intelec.dif.um.es/~aike/roque/pub/RMARINTH.pdf (Consultado: febrero, 2004).

Pignani, J. M. "Sistemas Expertos". Disponible en: http://www.modeladoeningenieria.edu.ar/utnfrro/orientacionl/monografias/pignani-sistemasexpertos.pdf (Consultado: febrero, 2004)

Montejo, Samuel. "Sistema para la Detección de Patrones de Fraude en las Redes de Telecomunicaciones". (Trabajo de tesis). Las Villas: Universidad Central "Marta Abreu", 2003.