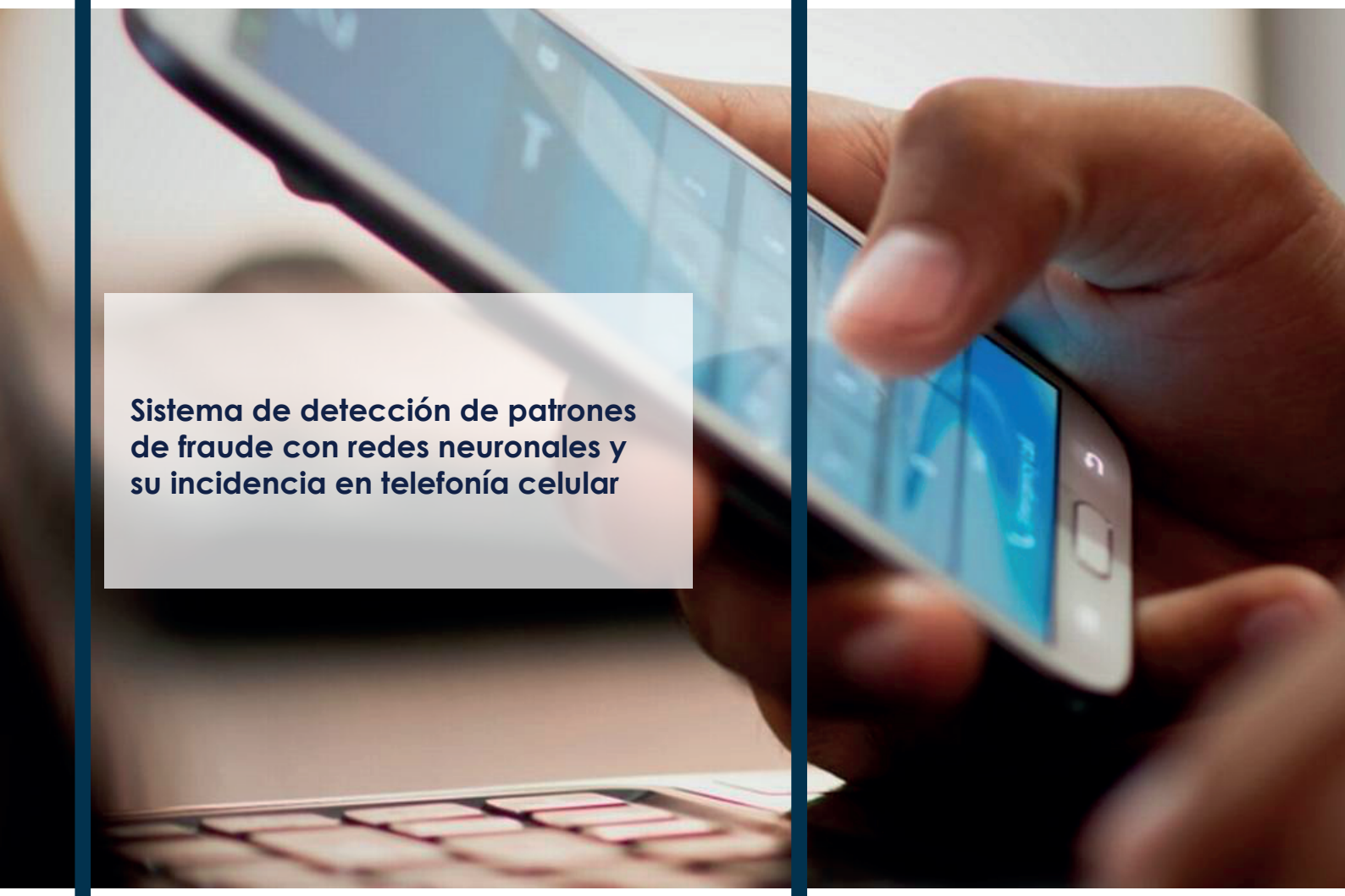
compAS
Grupo de capacitación e investigación pedagogica



Sistema de detección de patrones de fraude con redes neuronales y su incidencia en telefonía celular

Byron Oviedo Bayas
Roberto Pacheco Paliz

Sistema de detección de patrones de fraude con redes neuronales y su incidencia en telefonía celular

Byron Oviedo Bayas
Roberto Pacheco Paliz

Sistema de detección de patrones de fraude con redes neuronales y su incidencia en telefonía celular

compAs
Grupo de capacitación e investigación pedagógica

Título original:
Sistema de detección de patrones
de fraude con redes neuronales y
su incidencia en telefonía celular
Primera edición: enero 2020

Publicación derivada del 5to Congreso Multidisciplinario
de Investigación Científica.
Guayaquil-Ecuador

Editado en Guayaquil - Ecuador

ISBN: 978-9942-33-170-0

compÁs
Grupo de capacitación e investigación pedagógica

Cita.
B. Oviedo, R. Pacheco (2020) Sistema de detección de patrones de fraude con redes neuronales y su incidencia en telefonía celular , Editorial Grupo Compás, Guayaquil Ecuador, 91 pag

Índice

Capítulo 1
Sistemas de telecomunicaciones

Uno de los sistemas de telecomunicaciones que se ha desarrollado con mayor rapidez debido, al consumismo y al éxito que tiene el servicio de telefonía celular ha roto todos los pronósticos en los escenarios más optimistas realizados hace 22 años atrás, cuando las operadoras internacionales estudiaban al Ecuador como mercado potencial para introducir este negocio exitoso a nivel mundial, que genera riqueza y bienestar para la comunidad empresarial, sus empleados y distribuidores.

Tabla # 1.1 Resumen de Abonados en Telefonía Celular desde el (2009 – 2015).

RESUMEN			
FECHA	**TOTAL ABONADOS**	**TOTAL PREPAGO**	**TOTAL POSPAGO**
Mar-15	**16,174,686**	**12,095,818**	**4,078,868**

Fuente: Agencia de Regulación y Control de las Telecomunicaciones.

Asimismo, ha sido el motor de generación de micro emprendimientos. Las reparaciones de celulares, ventas de tarjetas, pines electrónicos, accesorios, teléfonos, etc., se han convertido en toda una actividad paralela, que complementa el negocio de telefonía celular, generando empleo pleno, subempleo e impuestos para el Gobierno, además de la satisfacción a sus usuarios, que encontraron en la telefonía celular probablemente como una necesidad primordial y necesario de su diario vivir, solo después de los alimentos. (Telecomunicaciones, 2015)

Las operadoras de telefonía celular en Ecuador son 3: CNT con el 6.35%, Conecel 64.75% y Otecel 28.91%, más de 16 millones están abonados a la telefonía móvil en el Ecuador con una penetración del 113.3% a julio del 2014 lo cual se entiende como que desde un recién nacido hasta el ciudadano de mayor edad en el país son clientes, superando incluso la población que es de 15.74 millones habitantes en el país, hasta marzo del 2015, se registraron 16174686, debemos descremar el mercado para establecer el mercado potencial real, restando a los niños desde los 8 años y los adultos mayores de 70 años, lo que nos da una población real estimada de 10.150 clientes, que representa una penetración telefónica del 80%. (Meza, 2011)

A todas luces, las estadísticas no son reales, es evidente que hay clientes con varios números asignados, teléfonos no operativos. La mezcla de teléfonos personales y corporativos evidencia la necesidad de regular el mejor registro estadístico por abonados a nivel corporativo y personas naturales, generar la guía telefónica celular virtual pública y privada en donde todos los clientes deben reportarse, y la excepción la constituyen los clientes protegidos a los que no se les publican sus datos. Lamentablemente, muchos de los delitos en telefonía celular transitan por vía celular, de allí el incremento de controles más rígidos y monitoreo de llamadas a números sospechosos. (Sellan, 2015)

La investigación se desarrollará en la Provincia de los Ríos, localizada en la Región Costa del país. Su capital es la ciudad de Babahoyo y su localidad más poblada es la ciudad

de Quevedo, se caracteriza por estar fuertemente comunicada por vías terrestres que la conectan a varias provincias y es un punto de conexión entre la sierra y la costa ecuatoriana. (Wikipedia.org, 2015)

Su territorio está ubicado en la parte central del litoral del país y limita con las provincias de Guayas, Santo Domingo de los Tsáchilas, Manabí, Cotopaxi y Bolívar. Abarca 6.254 km². Según el Instituto Nacional de Estadísticas y Censos, en 2010 su población era de 778.115 habitantes, siendo ligeramente superior el número de hombres, la densidad de población es de 124,42 hab./km². Cuenta con 13 cantones. (Wikipedia.org, 2015).

Los principales cantones de la provincia en cuanto a población son: Quevedo, con 173.575 habitantes, Babahoyo (153.776 hab.), Vinces (71.736 hab.), Ventanas (71.093 hab.), Buena Fe (63.148 hab.) y Valencia (42.556 hab.). (Wikipedia.org, 2015)

Dentro de los cantones citados, las localidades más pobladas son:

- Quevedo (150.827)
- Babahoyo (90.191 hab.)
- San Jacinto de Buena Fe (38.263 hab.)
- Ventanas (38.168 hab.)
- Vinces (30.241 hab.)
- Valencia (16.983 hab.)

Los Ríos es parte importante del conjunto de las seis provincias del litoral ecuatoriano que genera el 42,30% de las divisas no petroleras (alrededor de 75 818 millones de dólares cada año).

La Provincia genera más de 2000 millones de dólares como producción bruta al año, aproximadamente el 2,63% del total nacional[1]

Según estadísticas del INEC hasta el año 2013 la penetración de telefonía celular en la provincia de los ríos es del 48.1% de líneas activas.[2]

Los fraudes en el área de telecomunicación sobre todo en la telefonía celular se han presentados en el ecuador desde 1995 con la aparición del denominado Call Back o llamada revertida siendo este el primer tipo de fraude identificado en nuestro país. En 1996 se descubre el Bypass el mismo que se encuentre conformado por un enlace internacional, hasta este entonces únicamente se lo hacía con líneas fijas, haciendo llamadas internacionales y haciéndolas pasar como llamadas locales. Hasta el 2000 se empezó a combatir el Bypass este se realizaba mediante la identificación de grupos de líneas telefónicas fijas, instaladas en un mismo lugar, las cuales generaban considerables volúmenes de tráfico local, en el 2003 se descubre por primera vez que el bypass se estaba utilizando en líneas celulares, las probabilidades de encontrar las instalaciones clandestinas que forman parte de esos sistemas ilegales, son bastantes limitadas, a diferencia de las fijas la celular no se cuentan con medios físicos pares de cobre cuyo seguimiento permite descubrir la ubicación de dicha instalación ilegal y su posterior incautación.

[1] http://aplicaciones.los-rios.gob.ec/phocadownloadpap/PlandeOrdenamientoTerritorial/Sistema%20Economico.pdf
[2] http://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_Sociales/TIC/Resultados_principales_140515.Tic.pdf

En el 2011 se encontraron fraudes telefónicos a travez de las PBX o PABX son las siglas en inglés de Private Branch Exchange y Private Automatic Branch Exchange, en ese orden, cuya traducción sería Central Secundaria Privada Automática. Se refiere a cualquier central telefónica conectada directamente a la red pública de telefonía.

Este dispositivo, generalmente, pertenece a la empresa que lo tiene instalado y no a la compañía telefónica. Ésta será la que enrute la llamada hasta su destino final mediante enlaces unificados de transporte de voz, llamados líneas troncales, para gestionar, además de las llamadas internas, las entrantes y las salientes, con autonomía sobre cualquier otra central telefónico.

La SUPERTEL ha realizado investigaciones en varios casos relacionados con fraudes a PBX; en los cuales, se vuelve un elemento común la falta de atención a medidas básicas de seguridad. De la experiencia, se ha podido identificar algunos métodos utilizados por los defraudadores para atacar las PBX y realizar llamadas gratuitas. Técnicas como los ataques de fuerza bruta o conexiones remotas anónimas son las más utilizadas, Este tipo de fraudes se han vuelto más comunes en los últimos meses y lamentablemente, identificar a los responsables no garantiza que se eviten las pérdidas generadas. Estos representan un gran impacto para pequeñas y medianas empresas, para las cuales asumir el costo de una planilla telefónica mensual superior a 10 mil dólares puede resultar devastador. Sin embargo, tomando las medidas adecuadas, se puede evitar ser una víctima potencial,

entre las principales formas de ataques que han sido detectados en las centrales PABX, podemos encontrar los siguientes:

- Llamadas mediante acceso remoto,
- servicio de atención automática,
- mantenimiento remoto,
- transferencia de llamadas a operadoras internacionales,
- re direccionamiento de llamadas móviles, LDI y LDN

Fraude en telefonía celular

Fraude en redes de telecomunicaciones según GCI (2004) es el uso de servicios de una red de telecomunicaciones con la intención de no pagar por dichos servicios o modificar el pago de los mismos. Las organizaciones calculan las pérdidas monetarias por el dinero que dejan de percibir por los servicios que no cobraron debido a los fraudes. El dinero que han dejado de devengar las empresas ha ido en aumento año tras año.

Los fraudes son un reto de hoy para los operadores de telecomunicaciones del mundo, convirtiéndose en un problema en expansión que se va desarrollando paralelamente a la tecnología con una velocidad de adaptación elevada. Sus impactos se hacen sentir como innumerables pérdidas financieras y en un malestar de los clientes. Las comunicaciones móviles son las más afectadas por este flagelo y han dado abrigo a lucrativos negocios de organizaciones criminales que han hecho del fraude un próspero modo de vida.

Según CFCA se estima que las pérdidas de los operadores de telefonía móvil ascienden de unos 12 000 millones de dólares en 1999 a unos 35 en el 2003, 40 000 millones en el 2011. Estas pérdidas afectan tanto a las compañías como a los consumidores al incrementar los costos de operación de las compañías, lo que se revierte en la política de precios.

Los operadores conservan en silencio las medidas a tomar para resarcir los daños que los perpetradores les ocasionan como políticas de seguridad para no alertar a los criminales sobre sus sistemas de defensa. Una discusión abierta puede ser vista como un desafío a los criminales y el riesgo de compartir información confidencial. A la vez mantienen en secreto sus pérdidas por cuestiones de prestigio y por miedo a más ataques al revelarse como una red insegura.

El fraude en la telefonía celular tiene sus características propias, que lo distinguen del efectuado en la telefonía fija pública. Debido a la movilidad implícita de la telefonía celular se hace difícil la localización de los atacantes, actuando estos en ocasiones durante mucho tiempo sin poder ser detenidos. He ahí la necesidad de que los intentos de realizar llamadas ilegales sean detectados con la mayor brevedad posible para minimizar las pérdidas monetarias.

La inminente integración de la telefonía con los datos en las nuevas generaciones de telefonía celular con un enfoque centrado en conexiones a velocidades elevadas con Internet, ha lanzado predicciones de un incremento en las acciones fraudulentas a llevarse a cabo con una integración lógica del

fraude en las redes de datos junto al de la telefonía celular. Al avanzar el tiempo, la tecnología se va complejizando, y se hace más propensa a tener vulnerabilidades en el sistema, lo que la hace más susceptible ante nuevos ataques.

El fraude tiene una naturaleza compleja, es dinámico y evoluciona junto con las tecnologías. Para acometerlo no siempre se necesita tecnología de punta y el mercado irónicamente pone en las manos de los infractores las herramientas de forma barata para que estos cometan los delitos. Por otro lado es difícil de detectar pues la diferencia entre una conducta fraudulenta y una normal puede ser sutil.

En el fraude se deben tener en cuenta los escenarios en los que este se comete, la manera en que estos son perpetrados, cuáles partes de la red son vulnerables o usadas para cometer los fraudes. Se deben conocer los indicadores que indican que acciones ilegales se están cometiendo y que se clasifican según (Moreau, 1997) por su tipo y su uso en:

Por su tipo:

- Indicadores de uso, relacionados con la forma en que el móvil es usado.
- Indicadores de movilidad, relacionados con la movilidad del teléfono.
- Indicadores deductivos, que surgen producto de comportamientos fraudulentos (colisiones de llamadas, etc).

Por su uso:

- Indicadores primarios, que pueden ser usados aislados para detectar fraudes.
- Indicadores secundarios, que dan información valiosa por sí solos, pero que no son suficientes por ellos mismos.
- Indicadores terciarios que dan información de apoyo cuando se usan combinados con otros indicadores.

Varias son las formas de perpetrar los ataques a las redes de telefonía celular, por aire o por cables, con equipos más sofisticados o simplemente robando un terminal. Algunos de los fraudes más usuales son los siguientes, de los cuales analizaremos sus características principales:

- Hacking
- Clonación
- Tumbling
- Refiling
- Robo de identidad
- Hijacking
- Fraude interno Call back
- Abuso del roaming
- Tarjetas prepagadas falsas
- Robo de celulares
- Estafa de transferencia de llamadas

Hacking

Las estaciones base para comunicarse con el centro de conmutación o para que los centros de conmutación se

comuniquen entre sí o con la red pública conmutada usan cables, fibra u otro medio físico y tienen computadoras que son las que controlan todo estos procesos y las bases de datos de los clientes. Las estaciones de conmutación celular necesitan la máxima seguridad posible contra ataques pues estas computadoras no sólo controlan las conexiones celulares sino que también mantienen los registros de los números de serie electrónicos y el número de identificación del móvil junto con la información de tarificación.

Estas máquinas son accesibles por la red telefónica pública conmutada, la cual a su vez es accesible a través de Internet. Los ordenadores están físicamente conectados a módems, varias computadoras, LANs y WANs directa o indirectamente. Cualquier fallo de seguridad en los módems, computadoras o enlaces puede hacer una o más estaciones de conmutación vulnerables. Un ataque efectivo que le dé acceso a las bases de datos de usuarios a un atacante puede convertirse en pérdidas económicas enormes para el proveedor.

En 1993, los laboratorios GTE fueron seleccionados por la CTIA (Cellular Telecommunications Industry Association) para formar un laboratorio técnico de la industria para la detección de fraudes y en 1994 los laboratorios realizaron un ataque a una estación celular. Solamente se usaron técnicas de hackeo ordinarias, como la búsqueda de acceso a puertos abiertos y crackear contraseñas débiles.

El equipo atacante ganó rápidamente el privilegio de administración remotamente, alteraron el fichero de

contraseñas, obtuvieron información confidencial acerca de los ESN/MINs y la tarificación de los clientes.

El equipo dejó huellas deliberadamente con la esperanza de ser capturado, pero no fueron detectados. El equipo uso también ingeniería social para ganar acceso físico a las oficinas y al cuarto de las computadoras, quebrando el mecanismo de seguridad y puso un troyano en una computadora de una oficina.

A partir de ese momento se hizo un estudio sobre la vulnerabilidad de la industria celular y se propusieron recomendaciones sobre políticas de seguridad y estándares para la industria celular en cuanto a la seguridad.

Clonación

El fraude de mayor envergadura en la telefonía móvil es sin lugar a dudas la clonación Hai- Ping (1997). Los teléfonos celulares contienen dos números esenciales: el número de serie electrónico (ESN) y el número de identificación del móvil (MIN). El número de serie electrónico viene en la tarjeta física del teléfono y es un número único para cada terminal y el número de identificación del móvil se le ofrece al cliente por el proveedor. La clonación consiste en la escucha en el espacio radioeléctrico de estos números con el objetivo de grabarlos en la tarjeta de otro terminal telefónico y así valerse de la cuenta de un infortunado cliente para realizar llamadas gratis.

Los equipos de telefonía analógica son más golpeados por

este tipo de fraude pues no tienen un mecanismo de cifrado para transmitir los ESN/MIN y estos son captados con gran facilidad. En los teléfonos de segunda generación se han puesto en práctica algoritmos de cifrado para lograr una mayor confiabilidad.

GSM, líder de la telefonía celular de la segunda generación, confía en códigos criptográficos especiales para autenticar a los usuarios y cobrarles apropiadamente. Esto se efectúa gracias a una tarjeta inteligente llamada SIM (Subscriber Identification Module) registrada con el operador de red que guarda una clave secreta (un identificador único) que puede usarse para autentificar al consumidor y permitir que las llamadas sean cobradas a ese número.
Como cada teléfono tiene un único identificador es posible lograr que la tarjeta SIM se use solo con ese terminal telefónico y no sirva para ningún otro. Esto no se implementa en la realidad.

Cuando se hace una llamada telefónica el SIM prueba el conocimiento de la llave secreta al proveedor de servicios. El mayor problema de seguridad es, sin embargo que el conocimiento de la llave secreta y como programarlo a una SIM en blanco es suficiente como para clonar un teléfono. Los números de la SIM son generados de una forma especial y contienen dígitos de chequeo de forma que clonar no es solamente entrar un número aleatorio a una SIM en blanco. Interactuando con el SIM usando técnicas de craqueo disponibles por Internet se puede conocer la clave de la tarjeta SIM, Dempsey (1999).

Una vez que la SIM esta comprometida es posible realizar llamadas fraudulentas que le son cobradas a la víctima. Crackear la tarjeta SIM no rompe el algoritmo de la red entera pero una vez que la información y los parámetros de construcción del algoritmo son descubiertos, hace el crackeo de los demás más fácil.

Para lograr una mayor confiabilidad se añade un tercer número, llamado número de identificación personal (PIN), no obstante existen equipos que logran rastrear MIN/ESN/PIN en tiempo real.

Los equipos para la escucha radioeléctrica se encuentran a precios bajos en el mercado así como los softwares para el crackeo de las claves y la introducción del código en las tarjetas de los celulares. Por si esto fuera poco algunos proveedores han lanzado al mercado equipos clonadores de tarjetas profesionales como un servicio más a prestar a aquellos usuarios que tienen más de un número telefónico para integrarlos en un solo terminal GSM(2003). Casos de múltiples clonaciones a un mismo teléfono se han reportado, aumentando de esta forma las pérdidas. Llegado el momento el verdadero dueño del celular protesta y se le otorga un nuevo número de identidad, luego de cancelar la anteriormente existente para eliminar las infracciones del perpetrador. Este tipo de fraude manifiesta un cambio en el patrón de consumo del cliente.

Tumbling

Los criminales tienen una variante de clonación llamada tumbling, donde un teléfono es programado con las identidades de varios teléfonos (99 en un caso).Cada vez que se efectúa una llamada el número cambia su identidad de forma que sea más difícil detectarlo y asegurando de esta forma el funcionamiento del terminal fraudulento por un tiempo más prolongado. Este fraude es mucho más común en los teléfonos celulares analógicos. En este tipo de fraude el cambio en los patrones de consumo del cliente son menores a medida que aumente la cantidad de identidades que implemente el equipo fraudulento.

Refilling

Otro fraude común es el refilling, que consiste en que un tercer país enrute las llamadas hacia otro sacándole ventajas como intermediario a que las políticas de tarificación del tráfico a partir de este son más bajas. El país afectado ve disminuir sus ingresos de esta forma, pues percibe una menor cantidad de dinero que antes por el mismo tráfico, mientras que los otros se benefician ganado un tráfico adicional y tarifas más bajas.

Este es un tipo de fraude difícil de probar y como compromete empresas de otros países debe ser manejado cuidadosamente. El tráfico evidentemente manifiesta aquí una migración masiva de un país a otro en cuanto a la cantidad y frecuencia de llamadas, factor por el que se detecta.

Robo de identidad

Un fraude que ha estado ganando terreno es el robo de identidad, que es aquel en que una determinada persona adquiere datos de la víctima a estafar y en su nombre solicita un determinado servicio que posteriormente no paga.

Los ladrones de identidad han incrementado sus actividades en los últimos años y especialistas piensan incluso que sea el fraude del nuevo milenio Daniels (2001). Esto se debe a que las medidas que se han ido tomando contra la clonación de teléfonos han provocado una migración: los criminales siempre aprovechan las debilidades de los sistemas para hacer dinero fácil.

Estos defractores son capturados por la huella que dejan a su paso en forma de patrones y métodos de operación. Los hombres son personas de hábitos. Así, muchas veces se producen llamadas a los mismos números de call back con el reusó de las mismas cuentas o se mantiene el robo de servicios en las mismas áreas. Existe también un comportamiento bastante usual: el delincuente realiza llamadas a números comunes como el tiempo y la hora para verificar si el número está activado sin que relacionen al perpetrador con el teléfono (al hacer llamadas personales), luego se observa una demora de días desde la llamada de prueba hasta que el teléfono es usado por el usuario al que le venden el servicio.

Los delincuentes se llaman unos a los otros para negociar, pues suelen hacer negocios por vía telefónica. Muchas veces suelen tener listas codificadas de sus usuarios e incluso llegan a brindar

facilidades monetarias como premio a la fidelidad a sus subscriptores.

Las organizaciones criminales pueden tener numerosos usuarios los cuales realizan muchas llamadas por lo que capturar una banda reviste una gran importancia por el monto de las pérdidas que pueden generar. Muchas veces reusan una misma identidad para crear numerosas cuentas y pedir crédito. Este tipo de fraude también arroja cambios en los patrones de consumo históricos del usuario.

Hijacking

Otro posible ataque a través del aire es el hijacking. Una vez que el canal de voz es establecido entre una estación base y un móvil, un violador puede aumentar la potencia de transmisión de su terminal telefónico por encima de la del teléfono legítimo, pudiendo de esta forma hacer una llamada ilegal al robarse el canal de comunicaciones Hai-Ping (1997).

Fraude interno

Se entiende por fraude interno al cometido por los empleados de la empresa. Se valen para ello de su puesto o de la información que administran para usarla a su favor o a favor de terceros. Dentro de este caso se puede incluir la manipulación de los datos, los equipos y los sistemas. Personal despedido o descontento en la compañía pueden realizar este tipo de fraude para vengarse de la empresa.

Call back

El call back consiste en que empresas domiciliadas en el exterior, generalmente en los Estados Unidos, contactan a empresas nacionales que generan gran cantidad de tráfico internacional o a particulares. Una vez afiliados los usuarios llaman a un número determinado, generándose de esta manera una llamada que no es contestada. No obstante, un ordenador que está del otro lado de la línea identifica y guarda el número llamante para, una vez que este cuelgue, devolverle la llamada con tono de los EUA y tramitarla como si fuera local.

Las pérdidas monetarias son grandes con este tipo de fraude porque de esta forma se pierde mucho tráfico internacional. Países como los Estados Unidos brindan el call-back como un servicio telefónico más. Este ha sido motivo de discusiones con operadores de otros países que han logrado que numerosas compañías dejen de brindar este servicio. El call back tiene un comportamiento típico fácilmente detectable, de una llamada no establecida que tiene una respuesta en un período de tiempo breve y muchas llamadas son efectuadas a un mismo número. Existen variaciones más complejas de este fraude.

Abuso del roaming

El abuso del roaming es otro fraude de la telefonía móvil. La red móvil tiene que permitir el roaming entre redes y a través de otros países. Los estafadores pueden aprovechar el tiempo

que demora la actualización de los datos de tarificación y verificación del estado de las cuentas entre los operadores, de forma tal que pueda efectuar llamadas costosas durante el tiempo de viaje o incluso seguir usando un terminal cuya cuenta había sido cerrada en su lugar de origen.

Tarjetas prepagadas falsas

Las subscripciones con tarjetas prepagadas han sido diseñadas para permitir a los subscriptores pagar por una determinada cantidad de servicios de forma adelantada. Esta es una proposición atractiva en la que el cliente tiene pocas oportunidades de evitar el pago. Sin embargo, la seguridad en la producción de estos pequeños dispositivos debe ser buena, en un caso una impresora preparó un duplicado de tarjetas y se reportaron ventas por encima de un millón de dólares.

Cuando los verdaderos dueños de las mismas trataron de usar el monto de los servicios reservados, los sistemas de prepago habían grabado como usadas completamente las tarjetas y como que estas se habían vaciado. Se han reportado casos de fraudes recargando las tarjetas.

Robo de celulares

El robo de los celulares es otro fraude cometido con gran frecuencia. Más de 15 000 teléfonos móviles son robados cada mes en gran Bretaña de acuerdo a la compañía productora de teléfonos Ericsson. En el período de tiempo que sigue del hurto a su notificación para la desactivación del mismo, se pueden realizar una buena cantidad de

llamadas dejándose de devengar cantidades considerables de dinero.

Estafa de transferencia de llamadas

Otro tipo de estafa tiene que ver con la transferencia de llamadas. El estafador hace una llamada o deja una nota diciéndole a la víctima que ha ganado un premio en metálico y que para cobrarlo debe introducir un número de dos dígitos seguido por # o * y luego un número telefónico. El número telefónico es típicamente un número de larga distancia y el número de dos dígitos es el número de activación del servicio de transferencia de llamadas. Así, el estafador hace una llamada a la víctima y esta es enrutada hacia otro número telefónico y cargada a la cuenta del estafado.

Combatiendo el fraude

La protección de una red de telefonía celular conlleva consigo la creación de una política de seguridad bien trazada, entrenamiento del personal, realización de auditorías y la adquisición de productos para la detección de fraudes.

Para la adquisición de un producto de seguridad se deben tener en cuenta varios aspectos. El tráfico generado diariamente en las centrales telefónicas es enorme y esto presupone el tener que lidiar con cantidades realmente exuberantes de información, lo que conlleva consigo el uso de bases de datos igualmente gigantescas, en las que entonces juega un papel fundamental la velocidad de acceso a los datos para el procesamiento de la información.

Por esta misma razón la velocidad de procesamiento del sistema de detección de fraudes es crucial.

Igualmente debido a la naturaleza inherente a las redes celulares, estos sistemas deben estar distribuidos y ser capaces de realizar intercambios de información entre ellos. Otro parámetro importante al elegir un producto es la fiabilidad o número de alertas falsas que emite el sistema; si este número es elevado el sistema no es confiable.

Aunque la cantidad de fraudes sea alta, en comparación con la totalidad de los clientes representa un pequeño por ciento y el cliente no debe ser molestado innecesariamente para no crear malestar y opiniones negativas que puedan dañar el prestigio de la compañía en cuanto a los servicios que esta presta.

La efectividad del sistema, o sea, la cantidad de fraudes que detecta es un factor importantísimo en la adquisición del producto. Por último, la detección debe hacerse en el menor tiempo posible para minimizar las pérdidas, preferiblemente en tiempo real. Algunos sistemas se centran en la búsqueda de patrones de fraude para compararlos con los patrones de los usuarios y así determinar si hay algún comportamiento sospechoso al haber una similitud en los comportamientos. La idea opuesta es hallar cambios significativos en el comportamiento normal del sistema para así lanzar alarmas, pues los fraudes generalmente conllevan cambios en la conducta normal de los clientes. Numerosas compañías se han volcado a la producción de software y equipos para

la detección de fraudes y se estima que este es un mercado que genera millones de dólares. Los productos existentes en el mercado se distinguen por la cantidad de fraudes que logran identificar, la técnica empleada para la detección, capacidad para manejar la información de tráfico, prestaciones adicionales (como visualización de estadísticas de tráfico, etc), interfaz, precio, y la capacidad de estos de hacer la detección en tiempo real, cuestión esta indispensable para minimizar las pérdidas que ocasionan los fraudes.

Prevención vs. Detección

Empezamos distinguiendo entre prevención del fraude y detección del fraude. La prevención del fraude describe medidas para evitar que el fraude tenga lugar. En contraste, la detección del fraude implica detectar el fraude lo más rápido posible cuando este ocurre. La detección del fraude empieza a actuar cuando la prevención ha fallado. En la práctica, la detección del fraude se usa continuamente porque se desconoce el momento exacto en que la prevención falla.

La detección del fraude es una disciplina que se encuentra en una evolución constante. Una vez que se implementa un método de detección, los criminales adaptan sus estrategias a este y tratan de violarlo. Nuevos criminales entran constantemente a este campo. Muchos de ellos no están al tanto de los métodos de detección que han sido empleados con eficacia en el pasado, por lo que adoptarán estrategias que los llevaran a cometer fraudes identificables. Esto significa que las herramientas de detección más viejas deben usarse en combinación con las más modernas.

El desarrollo de nuevos métodos de detección de fraudes se vuelve más difícil por el hecho de que el intercambio de ideas y datos en el terreno de la detección de fraudes es limitado.

Muros a derribar

Dificultades a nivel de mercado, histórico, cultural e incluso de gobierno pueden conspirar en contra de la seguridad de los sistemas. La pasividad en cuanto a legislaciones que castiguen estos delitos ha llevado a que muchos de estos no sean reconocidos como tal, por lo que numerosos operadores hacen presiones y crean organizaciones para exigir legislaciones más fuertes y precisas para castigar este dañino fenómeno.

Dificultades a nivel de gobierno

Los gobiernos de diferentes países pueden influir en el debilitamiento de los estándares de seguridad. El ejemplo a continuación ilustra perfectamente esta afirmación.

Una de las características más atractivas de GSM es que es una red bastante segura. Las comunicaciones, tanto de voz como de datos son encriptados para prevenir que sean escuchadas por personas indeseadas. De hecho, en sus etapas más tempranas se descubrió que el algoritmo de encriptación que utilizaba era demasiado poderoso para ciertas tecnologías de regulación de exportaciones. Esto pudo haber traído serias limitaciones para la diseminación global de GSM, limitando el número de países a los cuales se hubiera

expandido. Afortunadamente el MoU (memorandum of understanding), terminó con esta amenaza.

Algoritmos alternativos fueron desarrollados para permitir la libre diseminación de la tecnología por el mundo entero. Esto significa que la protección en la telefonía celular fue deliberadamente debilitada como resultado de una intervención directa del gobierno de los Estados Unidos, particularmente en respuesta a la oposición del FBI y la NSA (National Security Agency) a los algoritmos de encriptación fuertes. La oposición derivaba del problema que supondría escuchar conversaciones por estas organizaciones en los teléfonos celulares. Como una consecuencia directa de esto, los algoritmos que protegen las comunicaciones GSM son menos fuertes de lo que pudieran ser.

En la transición hacia la tercera generación de la telefonía celular, concebida como un paso gradual, se percibe nuevamente el problema de que el algoritmo de encriptación no refleja el estado del arte (se usan curvas de encriptación elípticas en lugar de RSA) debido a las restricciones de importación de los EUA en el área. De esta forma el desarrollo de los estándares se revierte en una menor seguridad. Esto le permite a los perpetradores ganar ventaja sobre las escuchas de las agencias de inteligencia usando teléfonos clonados, mientras los usuarios legítimos permanecen con el riesgo de ser víctimas de fraudes.

Políticas de accesibilidad

Las consideraciones de la política de accesibilidad global no

permiten adoptar las mejores defensas tecnológicas como estándares para los teléfonos celulares. La implicación de este hecho es que otras tecnologías deben ser adoptadas para minimizar la incidencia del fraude. Un principio simple es que la prevención del fraude y los mecanismos de detección no deben evitar que algunas clases de usuarios usen la tecnología. Los dos elementos principales relacionados con la accesibilidad son la facilidad de uso y la privacidad.

En relación con la facilidad de uso, contraseñas, números de PIN y restricciones similares son frecuentemente más abusadas que usadas correctamente. Los números de PIN son típicamente basados en cumpleaños, aniversarios o secuencias fáciles de descifrar como por ejemplo 2468. Las contraseñas son frecuentemente nombres de mascotas o parientes cercanos.

Una razón típica para este comportamiento es que las personas no tienen el tiempo o la inclinación para recordar una serie de números inútiles que requieran el uso de su memoria.

En relación con las facilidades de uso, los dispositivos de identificación pueden resultar alienantes y no deben resultar incómodos al cliente.

Factores socio-culturales e históricos

Los SIMs tienen la posibilidad de contener datos biométricos que podrían ser usados para proveer un nivel extra de

seguridad. La dificultad con la adopción de este tipo de mecanismo de seguridad no es tanto técnica como histórica o cultural. Los datos biométricos son datos muy sensibles y muchas personas se perturban por el pensamiento de su uso y potencial abuso.

Técnicas de detección de fraudes

Numerosas técnicas se usan para la detección de fraudes, cada una con sus características que la hacen más efectiva contra uno u otro fraude. En la búsqueda de lograr una mayor cobertura contra todos los fraudes se unen unas con otras para fortalecer los sistemas o se usan en conjunto.

Algunas de las técnicas más usadas en la detección de fraudes son: autentificación, huella de radiofrecuencia digital, minería de datos, verificación de voz, asignación de PINs dinámicos, análisis de tráfico, agentes inteligentes y redes neuronales.

Análisis de tráfico

Hasta la fecha, el método más usado para detectar el fraude en esta rama es el análisis de tráfico. Varianzas notables en los patrones de llamada pueden implicar fraudes. El tiempo y la localización de las llamadas previas pueden posibilitar también por imposibilidad física la presencia de un clon. Al menos esto puede proveer evidencia para el individuo queesta siendo incorrectamente tarificado. Fraudes como el call back tienen un comportamiento típico fácilmente detectable de una llamada no establecida que tiene una respuesta en un periodo de tiempo breve. Actuar ante la aparición de datos

sospechosos puede ser perjudicial si no se maneja correctamente la situación (las personas cambian sus patrones de llamadas y una investigación basada solamente en variaciones de tráfico puede causar invasión en la privacidad del cliente).

Minería de datos

La minería de datos es un método popular para combatir el fraude por su efectividad. Detecta los fraudes por aproximaciones estadísticas. La minería de datos es un procedimiento bien definido que toma datos como entrada y produce salidas en forma de modelos o patrones. La tarea de la minería de datos es analizar una cantidad masiva de datos para extraer información útil que pueda ser usada posteriormente.

Al analizar la información debemos definir la meta de la minería de datos y hallar la estructura correcta de los posibles modelos o patrones que encajan con los datos. Una vez que tenemos el modelo correcto para los datos podemos usar el modelo para predecir eventos futuros al clasificar la información. En términos de detección la minería de datos puede verse como una clasificación de datos. Los datos de entrada son analizados con el modelo adecuado y se determina por comparación si hay sospechas de actividades fraudulentas.

Un modelo de clasificación bien definido se desarrolla reconociendo los patrones de comportamiento fraudulentos. Entonces el modelo puede usarse para predecir cualquier

actividad sospechosa implicada en los datos de entrada. Una limitación al uso de esta técnica es su problema de eficiencia. La construcción del modelo requiere una cantidad enorme de tiempo, lo que implica que no se pueda reconocer el fraude en tiempo real. Además si los criminales cambian sus patrones no los detecta, por lo que hay que construir nuevos modelos.

Agentes inteligentes

Con el rápido crecimiento de las tecnologías de la información, muchos métodos que explotan las capacidades de la inteligencia artificial han sido creados. Uno de estos métodos es el uso de agentes inteligentes, los que incorporan las tecnologías de computación junto a las técnicas de minería de datos.

Los agentes inteligentes son programas de computación que pueden actuar automáticamente para realizar diversas funciones y no se adhieren a un modelo o regla. Pueden construir nuevos modelos y reglas con sus capacidades de aprendizaje, necesitando poca intervención humana. En un entorno donde se empleen múltiples agentes, estos pueden actuar en paralelo y cooperar unos con otros. Esto no sólo acelera el proceso sino que incrementa la efectividad de la detección. Los agentes inteligentes permiten la detección de fraudes en tiempo real.

Al distribuirse los agentes se distribuye también la información, con lo que aumenta la velocidad, puesto que cada uno tiene

que consultar menor cantidad de datos. Así también se puede distribuir la información en varias PC de forma tal que no sean necesarias estaciones de trabajo con tantas exigencias en cuanto a memoria. Por último, se les pueden enseñar muchos modelos o reglas para optimizar la detección y ellos son capaces de derivar otros nuevos a partir de las entradas Wu (2001).

Compañías como Searchspace brindan soluciones en el mercado basado en agentes inteligentes para combatir el fraude Wrolstad (2001). Este sistema en particular crea perfiles de usuarios mediante módulos de software inteligentes llamados centinelas que analizan los tipos de llamadas hechas, los números llamados, la longitud de las llamadas y cuándo son hechas estas. Si alguna actividad sospechosa es detectada, el usuario es encuestado a introducir un PIN para verificar su cuenta que al no recibir la confirmación correcta puede terminar el servicio. Estos perfiles de usuarios son creados a partir de los tickets de las llamadas y se adaptan continuamente en los cambios en los patrones de uso verificando estos cambios para detectar el fraude.

Otro ejemplo de estos agentes es el sistema de aprendizaje de clasificación multiagente. Este sistema fue propuesto por Abu-Hakima (1997). Consiste en tres agentes: el agente de comunicación personal (PCAs), el agente de movilidad de red (MNAs), y el agente detector de fraudes (FBAs).

El PCA (agente de comunicación personal) confecciona un perfil de usuario. Puede monitorear todas las llamadas

salientes, los tiempos de llamada, la duración de estas y la información del receptor. Una vez que obtienen la información hacen una base de datos. Esta información es comparada con los datos históricos de los clientes. Si el agente hallauna llamada atípica trata de notificar al usuario mediante un correo electrónico o un teléfono regular. Si el usuario no puede ser notificado, la llamada es cortada por el agente. Los agentes de movilidad de red residen en los centros de conmutación móviles e interactúan con los PCA para crear un mejor perfil de usuario brindando información de pago acerca de un usuario a los PCA. Si un MNA detecta una llamada sospechosa alerta al PCA. Entonces el PCA alerta al usuario sobre la llamada sospechosa o guarda evidencias para probar que la llamada es falsa.

Los agentes detectores de fraudes (FBA) están equipados con varios algoritmos de clasificación y se especializan en detectar patrones de llamadas fraudulentos. Esos patrones incluyen llamadas internacionales largas, llamadas simultáneas originadas por el mismo teléfono celular, llamadas a centros criminales o regiones sospechosas. Estos agentes también residen en el centro de conmutación y basados en la información de los FBA, los MNA alertan a los PCA a chequear el perfil de usuario para comprobar las características de la llamada sospechosa.

Este es un sistema propuesto que no se ha implementado y que provee un nivel de seguridad adicional con cada agente, interactuando unos con otros para minimizar las falsas alarmas.

Redes neuronales artificiales

Las redes neuronales son mecanismos de procesamiento estadístico de la información compuesto por numerosos nodos o unidades de procesamiento distribuidos que realizan cálculos simultáneos y se comunican utilizando conexiones adaptables llamadas pesos.

Las redes neuronales están compuestas por un número variable de capas de neuronas, derivándose de esta forma diversas arquitecturas que definen las distintas redes. Los procesos de las redes neuronales comprenden tres etapas: entrenamiento, prueba y funcionamiento. En la etapa de entrenamiento los pesos son adaptados a las diferentes neuronas usando algoritmos de entrenamiento que tienen en cuenta los datos pasados. Luego, usando datos de prueba se evalúa si la red actúa de forma adecuada. Este proceso de entrenamiento y prueba se repite hasta que la red trabaje correctamente. Luego viene la etapa de explotación de la misma.

Las redes neuronales pueden ser la mejor solución en ocasiones en que la selección de reglas es difícil en términos de velocidad y complejidad.Las redes neuronales son preferidas por dos razones principales. En primer lugar, las características aritméticas de la red la hacen adecuadas para manejar grandes cantidades de información. Las redes ponen más atención a la identificación de patrones que al análisis de los datos. En segundo lugar, las redes pueden mantenerse alternando sus pesos entre los enlaces con los datos

acumulados durante el entrenamiento, de forma que se adaptan fácil y rápidamente a los cambios en las entradas. Esta es la razón principal por la que son consideradas una aplicación ideal para la detección de fraudes que enfrenta una gran cantidad de datos. Productos como Cerebrus liderean el mercado pero hay otras ofertas en el mercado como Minotaur de Neural Technologies.

Huella de radiofrecuencia digital

Este sistema electrónico evoluciono de la tecnología militar, donde era usada para determinar según las características de radio de los aviones si eran hostiles o amigables Patrick (1997).

Cada teléfono celular tiene características específicas de radiación que dependen de su edad, las tolerancias de sus componentes internas, el modelo, cuan usado esta su teclado, etc, Vadman (1997) Las tolerancias de construcción varían dándole al transmisor características únicas como son la velocidad y uniformidad con que la portadora sube desde cero hasta la potencia deseada. También se mide la operación del sintetizador de frecuencias hasta que la frecuencia se vuelve estable, así como el ruido de fase, los armónicos y la desviación pico de la señal.

Este equipo almacena el patrón de radiación de cada aparato telefónico celular en una base de datos y lo usa para verificar si el aparato coincide con la huella de radiofrecuencia. Además comprueba los números de identificación que se chequean habitualmente al realizarse una llamada. Una computadora compara las características de amplitud y

frecuencia medidas con las de las bases de datos para cada teléfono con el mismo número de serie electrónico. Si hay una distancia significante esto indica que ha ocurrido una clonación y la llamada no llega a efectuarse pues es desactivada en fracciones de segundo. El perpetrador oye el tono de marcar pero no logra establecer comunicación, eliminando de esta forma el fraude sin pérdidas económicas implicadas. Este sistema es muy caro, pues es difícil analizar una huella de radiofrecuencia y decidir si la llamada es correcta o no. Muchas variables atentan contra la correcta detección de los parámetros como son la atenuación multitrayecto, la velocidad relativa del móvil, la temperatura variable, las condiciones del terreno, etc., por lo que se debe acudir a niveles de tolerancia aceptables para las mediciones. Además, el sistema no ofrece una cobertura total contra todos los tipos de fraude y es caro Struthers (1997). Un equipo de esta índole costó 3,3 millones de dólares al adquirirlo en Costa Rica. No obstante, para operadores con altos niveles de fraudes como la ciudad de Nueva York estas inversiones se recuperan en corto tiempo, cuestión que no se cumple con las pequeñas empresas. El producto líder en el mercado se llama PhonePrint y es de la firma Corsair, una ex empresa militar.

Verificación de voz

Los adelantos en las técnicas de procesamiento de voz han posibilitado la implementación de esta técnica a la detección de fraudes con niveles de aceptación adecuados. La tecnología de reconocimiento de voz es biométrica. La identificación se produce al pronunciar una frase convenida, que es comparada con la información del

usuario que se encuentra en las bases de datos. Provee un mayor grado de seguridad y no hay necesidad de memorizar algún tipo de código. Muchos proveedores de servicio para redes inalámbricas ofrecen facilidades de marcado de voz y facilidades de traducción voz-texto. La tecnología de huella de voz fue desarrollada en la universidad Rutgers para identificar pilotos volando aviones de combate por el Departamento de defensa de los Estados Unidos. Como la tecnología fue creada para trabajar en ambientes ruidosos, es ideal para las comunicaciones móviles que envuelven estática, efecto doppler por la velocidad de los vehículos, etc, Meyers (1997). Esta tecnología reconoce los patrones de voz incluso si el usuario está enfermo, pues se basa en la huella de voz, datos biométricos y una contraseña grabada. El sistema analiza patrones del sistema vocal completo así que es muy robusto y difícil de violar.

El instituto SAS, desarrollador de sistemas de detección de fraudes para bancos y compañías telefónicas lanzó un producto que procesa patrones de voz como el tono que son indicativos de cuando la persona miente Harvey (2002). Compañías como T-Netix y Authentix ofrecen tecnologías de verificación de voz como Voice Verifi-Air.

Sistemas de control de acceso a las compañías

Estos sistemas de control de acceso pueden basarse en numerosas técnicas y se unen al personal de vigilancia para crear un mecanismo único de seguridad física. Se usan técnicas como el reconocimiento de voz para los accesos, el uso de

contraseñas, tarjetas magnéticas, identificación personal, etc., para regular el acceso a los diferentes departamentos, acceso que por demás debe ser departamentalizado, creando niveles de manipulación de la información de seguridad.

Auditorías

Muchas compañías se dedican a ofrecer auditorías para ayudar a mejorar los servicios que brinda la empresa así como su seguridad. Observando las operaciones que se realizan en la central se identifican riesgos de fraude, problemas de seguridad, se cuantifican los riesgos de ataques, se hacen mediciones, analizan estrategias y opciones y se recomiendan soluciones. Compañías como AMS brindan este servicio, revisando además los recursos críticos y los componentes de la empresa incluyendo la organización, los sistemas, productos y procesos.

Autentificación

Uno de los métodos más usados consiste en la identificación del móvil mediante un sistema de claves que se intercambian entre la base y el cliente. Estas claves no tienen que transmitirse necesariamente en el principio mismo de iniciarse la comunicación sino que puede que se transmita información antes de comprobarse la veracidad de la clave secreta. Estas claves se guardan en bases de datos en servidores. No se transmiten las claves en sí sino datos codificados con estas y pueden transmitirse dentro del canal de voz, lo que hace más difícil descifrarla. El cliente no tiene que memorizar ningún tipo de código pues este esta implícito en el teléfono. La empresa

Synacom Technology entre otras ofrece CloneSafe, un sistema de autentificación de clave secreta basado en las claves A.

PINs (Personal identification numbers) dinámicos

Otro esfuerzo realizado para mejorar la seguridad de las comunicaciones móviles es la asignación de números de identificación personales que se verifican con los guardados en las bases de datos en las centrales. Al efectuarse la llamada se convida al cliente a marcar su código de identificación y si este no coincide con el almacenado en la central la llamada no se llega a efectuar.

REDES NEURONALES

Según Kohonen "las redes neuronales artificiales son redes interconectadas masivamente en paralelo de elementos simples (usualmente adaptativos) y con organización jerárquica, las cuales intentan interactuar con los objetos del mundo real del mismo modo que lo hace el sistema nervioso biológico". Las Redes Neuronales son un conjunto de algoritmos matemáticos que encuentran las relaciones no lineales entre conjuntos de datos. Estas se usan en la práctica en muchas aplicaciones donde hay envueltas cantidades masivas de información, en la creación de patrones como clasificadoras de conjuntos de datos, en la optimización, como herramientas para la predicción de tendencias, etc. Se denominan neuronales porque están basadas en el funcionamiento de una neurona biológica cuando procesa información.

Caracterización de las redes neuronales artificiales

Una red neuronal es un modelo computacional que pretende simular el funcionamiento del cerebro a partir del desarrollo de una arquitectura que toma rasgos del funcionamiento de este órgano sin llegar a desarrollar una réplica del mismo. El cerebro puede ser visto como un equipo integrado por aproximadamente 10 billones de elementos de procesamiento (neuronas) cuya velocidad de cálculo es lenta, pero que trabajan en paralelo y con este paralelismo logran alcanzar una alta potencia de procesamiento.

A partir de esta visión del cerebro el modelo computacional desarrollado consiste de un conjunto de elementos computacionales simples (llamados también unidades o celdas), las cuales constituyen neuronas artificiales y que tienen la forma que se muestra en la figura2.1. Siendo:

$H_i(t)$ el potencial sináptico de la neurona i en el momento t.

X_j la entrada de datos procedentes de la fuente de información j. W_j el peso sináptico asociado a la entrada X_j.

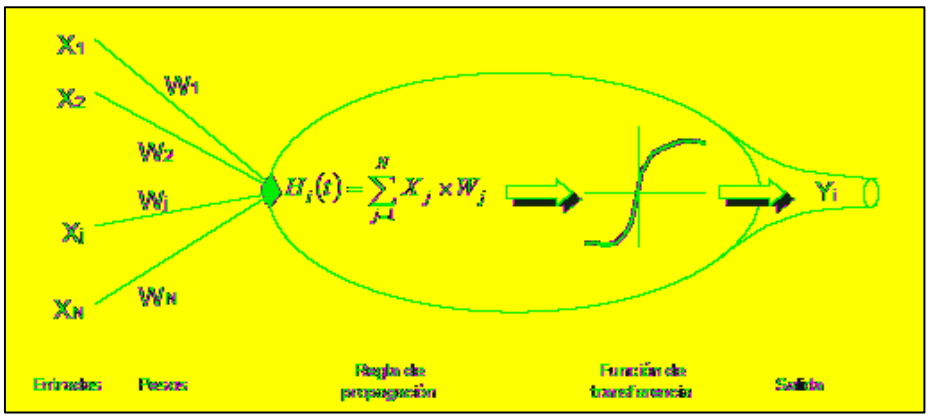


Figura 2.1 Esquema de una neurona artificial.

Si Xi no está conectada a Hj entonces Wij=0. Por convenio existe una unidad X0 con activación siempre igual a 1 que está conectada al resto de los elementos de procesamiento y el peso W0j es una constante que representa una predisposición de la unidad (bias), es decir, un valor inicial añadido al calcularse el nivel de actividad del elemento. (Bertona, 2005)

Los pesos positivos (Wij > 0) indican que el nivel de actividad de la unidad Xi refuerza el nivel correspondiente del elemento Hj, mientras que un peso negativo (Wij < 0) representa una inhibición. Los pesos determinan el comportamiento de la red, desde el punto de vista de la inteligencia artificial el conjunto de pesos W encierra el conocimiento del dominio de aplicación que la red neuronal modela.

Cuando el resultado de la regla de propagación supera un cierto número, denominado umbral, entonces la neurona se activa y el número resultante de la regla de propagación se evalúa en una función denominada función de transferencia. Esquemáticamente se podría representar de la siguiente manera:

$$H_i(t) > \theta \text{ entonces } f[H_i(t)]$$

La función f se conoce como función de activación y puede ser una función dura como la función signo, o una función no lineal diferenciable creciente monótonamente con forma de S (función sigmoidal) tal como la tangente hiperbólica.
La forma en que se calcule Hj determina que la red se considere un clasificador lineal o no lineal. Los clasificadores lineales

están limitados en su capacidad, y por supuesto están limitados a formas linealmente separables de discriminación de patrones. Los clasificadores más sofisticados con alta capacidad son no lineales.

Se puede escoger diferentes funciones para la función de transferencia Cuatro funciones de transferencia típicas que determinan los distintos tipos de neuronas son según (Bertona, 2005):

1. Función escalón
2. Función lineal
3. Función Sigmoidea
4. Función Tangente

Sigmoidea Función escalón:

La función escalón se utiliza cuando la neurona tiene salidas binarias cero o uno. La neurona se activa cuando el valor del potencial pos sináptico es mayor o igual a cierto valor umbral. Por ejemplo:

Sea $f(x) = 1$ cuando $Hi(t) \geq 0$ Y
$f(x) = 0$ cuando $Hi(t) < 0$

En este caso el umbral es cero. Cuando la función de propagación supera el valor cero, la función tomará valor uno. En caso contrario la función tomará el valor cero.

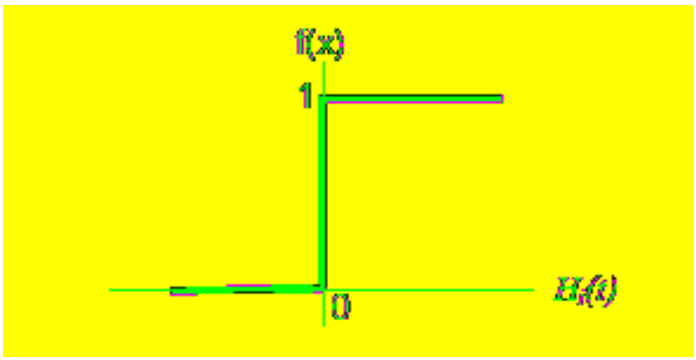


Figura 2.2 Función escalón.

Función lineal

La función lineal responde a la expresión f(x) = Hi(t). Una variación de la función lineal sería la función lineal a tramos donde la salida de la neurona sería la función identidad siempre y cuando el valor del potencial pos sináptico estuviese dentro de un rango de valores. Al estar fuera del rango la función se torna constante. Un ejemplo sería:

Se	f(x) = 1	cuand	Hi(t) > 1
Y	f(x) =	cuand	-1 ≤ Hi(t) ≤
Y	f(x) = -1	cuand	Hi(t) < -1

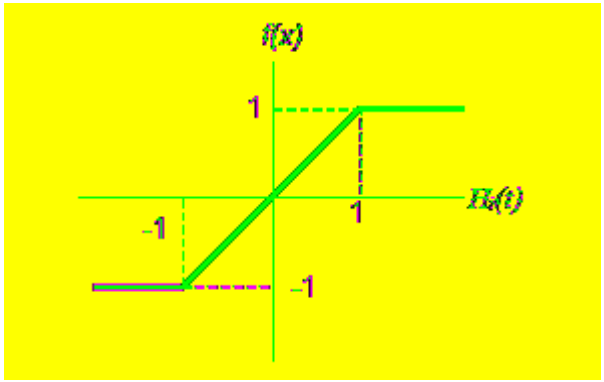


Figura 2.3 Función lineal a tramos.

En la figura 2.3 se aprecia como la función pasa a ser la función identidad a partir de que la función de propagación tome el valor –1 hasta el valor + 1. Fuera de este rango la función se torna constante con un valor de –1 desde - ∞ hasta –1 y con un valor de +1 desde +1 hasta + ∞.

Función Logarítmica Sigmoidea:

La salida de esta función siempre será continua en el rango entre cero y uno. Con esta familia de funciones se pueden utilizar datos continuos o digitales proporcionando salidas exclusivamente continuas. Su ecuación es la siguiente:

$$f(x) = \frac{1}{1+e^{-x}}$$

Siendo x=Hi(t)

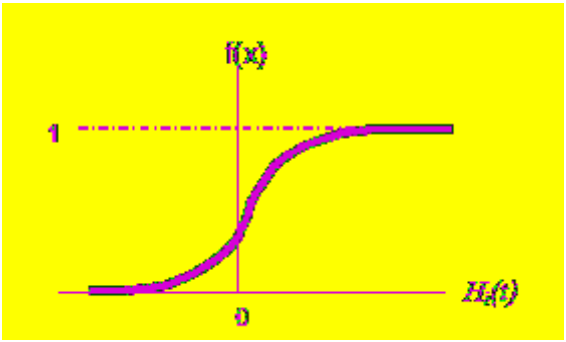


Figura 2.4 Función Logarítmica Sigmoidea.

En la función representada en la figura 2.4 se observa como la función adopta valores muy próximos a cero cuando X es pequeño, pero que según aumenta el valor en el eje de las abscisas la función pasa a ser creciente. Al principio la pendiente de la función crece hasta llegar a un punto de

inflexión, momento en el cual la pendiente comienza a descender hasta llegar a cero, a partir del cual la función vuelve a dar un valor constante e igual a uno.

Función Tangente Sigmoidea:

Esta es una de las funciones más utilizadas en las redes neuronales por su flexibilidad y el amplio rango de resultados que ofrece. Las ventajas de utilizar una tangente sigmoidea frente a una sigmoidea reside en que la segunda sólo ofrece resultados en el rango positivo entre cero y uno, en cambio la tangente sigmoidea da resultados entre –1 y 1, por lo que se amplía a los números negativos los posibles resultados. La función tiene la siguiente ecuación:

$$f(x) = \frac{2}{1+e^{-2x}} - 1$$

Siendo x=Hi(t)

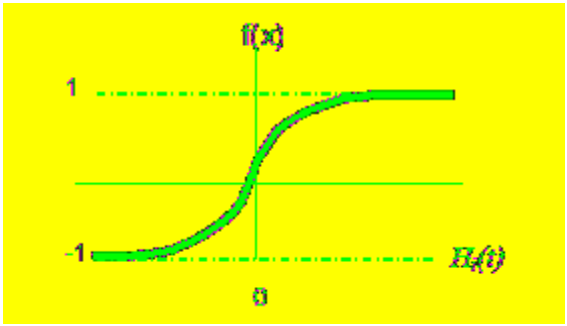


Figura 2.5 Función Tangente Sigmoidea.

El resultado que nos ofrece cada una de estas funciones será el dato de salida de la neurona que se dirigirá bien a otra

neurona, bien al resultado final.

Topología de red

Grupos de neuronas artificiales pueden ser interconectadas en una variedad de maneras para formar redes neuronales artificiales. Se define como topología de una red neuronal a la organización o arquitectura del conjunto de neuronas que la forman, a la distribución espacial de las mismas y los enlaces entre ellas.

Normalmente los elementos de proceso se organizan como una secuencia de capas con un determinado patrón de interconexión entre los diferentes elementos de proceso que las forman, y con un patrón de conexión entre los elementos de proceso de las distintas capas. Uno de los rasgos que puede ayudar a definir una capa es el hecho de que todos los elementos de proceso que la forman usan la misma función de transferencia. En muchas de las arquitecturas de redes neuronales se puede hacer la siguiente distinción entre las capas:

Capa de entrada: Es la capa que recibe los estímulos del entorno. No suele tener asociado un mecanismo de aprendizaje, es decir, sus pesos se mantienen constantes, y su misión simplemente es la de distribuir dicha entrada al resto de los elementos de proceso que constituyen la red.

Capa de salida: Es la capa sobre la que se forman las salidas de la red. Capas ocultas: Son las demás capas que no son ni de

entrada ni de salida. A continuación se presentan algunas topologías básicas, Bello (1993).

Neurona simple

La expresión más simplificada de una red es aquella en la cual se tiene solamente una neurona. Esta funciona como una unidad de procesamiento que recibe entradas y calcula un nivel de activación que definirá su respuesta.

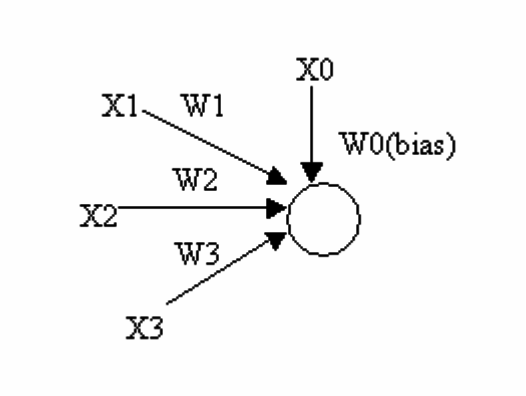


Figura 2.6. Una neurona simple.

Red simple

En esta topología se organiza un conjunto de N neuronas de la forma siguiente: se utilizan M unidades censoras, las cuales captan la información de entrada. Estas M unidades se conectan a las N neuronas mediante caminos pesados. No existe conexión entre las neuronas. Cada neurona calcula una respuesta, y la salida de la red será un vector con N componentes, una por cada neurona.

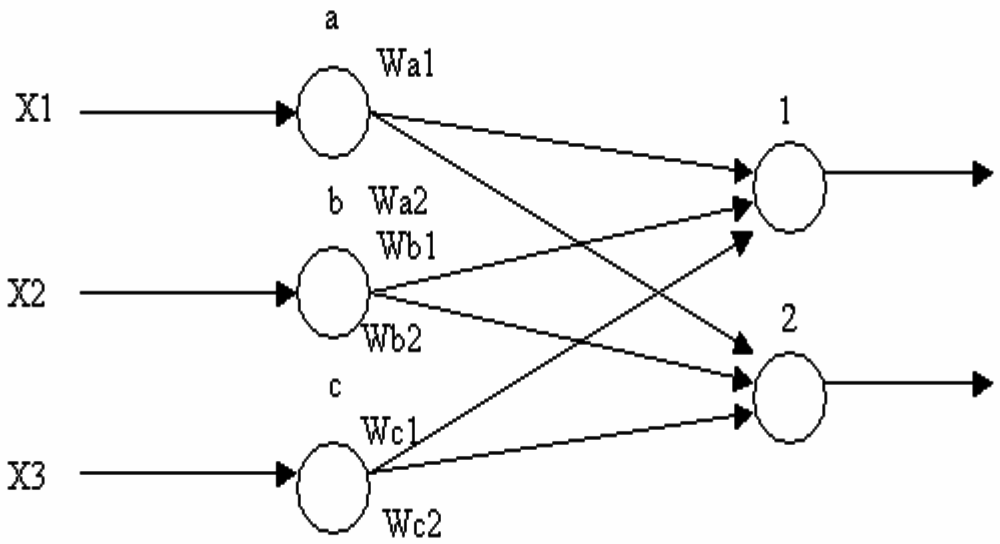


Figura 2.7 Topología de una red simple.

Red con elementos de asociación

La arquitectura anterior se puede desarrollar introduciendo un conjunto de unidades de asociación en las cuales se combinan las entradas (pueden utilizarse para esto operadores lógicos) y las combinaciones producidas sirven de entrada a las neuronas, las cuales calculan las salida de la red.

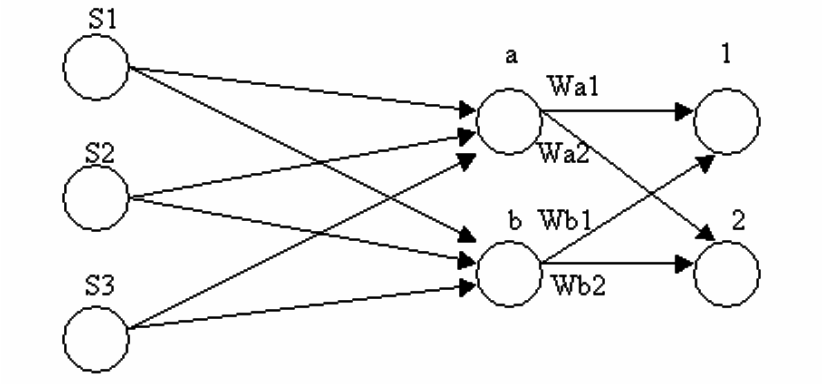


Figura 2.8 Topología de una red con unidades de asociación

Redes multicapas.

Una de las topologías más poderosas y difundidas ordena el conjunto de elementos de procesamiento en niveles, de modo que los enlaces se establecen desde unidades en el nivel i a unidades en el nivel j (i<j), por lo que la información fluye unidireccionalmente desde las unidades de entrada a las unidades de salida. Esta arquitectura se conoce como dirigida adelante. Típicamente existe una capa de unidades censoras, una o más capas ocultas de neuronas, y una capa de neuronas que producen la salida.

En esta topología resulta de interés analizar lo referente a cuántos niveles o capas, así como la cantidad de neuronas por capas, son necesarias. En el nivel inicial existe una unidad sensora por cada rasgo de entrada a la red, en la capa de salida se colocan tantas neuronas como sean necesarias. Se ha probado que dos capas ocultas son suficientes para resolver cualquier problema. No existe un criterio riguroso para determinar la cantidad de neuronas en las capas ocultas.

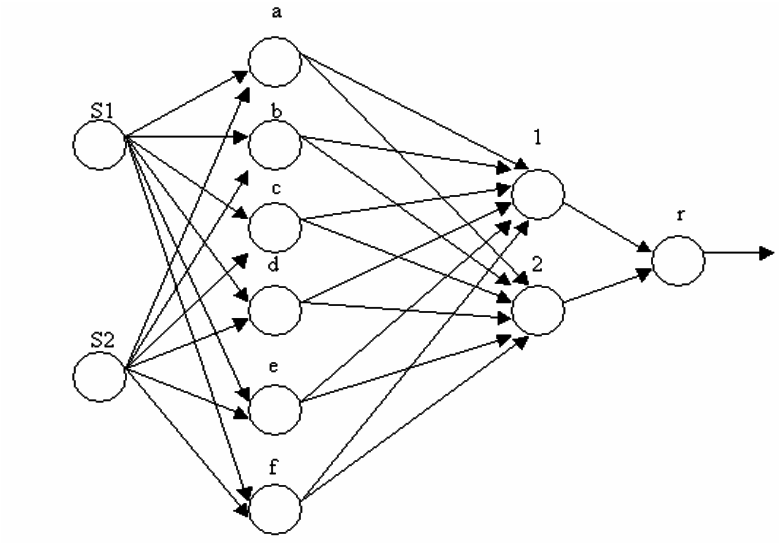


Figura 2.9 Topología de una red con capas ocultas.

Topología del modelo interactivo

Se tiene un conjunto de N neuronas las cuales se conectan completa y mutuamente, es decir, todas las unidades sirven como unidades de entrada y como neuronas para calcular la salida; cada neurona se conecta a las N-1 restantes mediante caminos pesados. El procedimiento de cálculo de las salidas se realiza mediante un procedimiento iterativo en el cual las neuronas se inicializan con los valores de entrada y recalculan estos valores hasta caer en un estado estable.

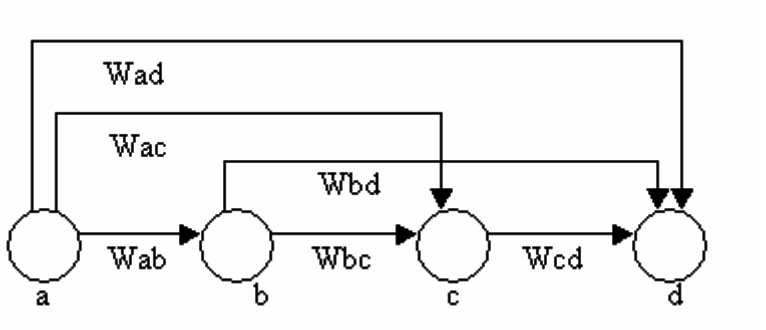


Figura 2.10 Topología del modelo interactivo.

Métodos de entrenamiento

Una de las propiedades fundamentales de las redes neuronales es la capacidad de adaptarse al entorno, aprendiendo a proporcionar la respuesta adecuada ante los estímulos que reciba de este entorno. Este aprendizaje se plasma en la modificación de los pesos de las conexiones entre los distintos elementos que forman la red, de tal forma que memoriza los ejemplos de entrenamiento que se le presentan.

Un conjunto de reglas bien definidas que describen el

método de adaptación o modificación de los pesos de acuerdo con el entorno en el que se encuentra sumergida la red recibe el nombre de regla de aprendizaje, y su transcripción en forma de procedimiento se denomina algoritmo de aprendizaje.

Existe una relación muy fuerte entre la arquitectura de una red neuronal artificial y el o los algoritmos de aprendizaje que puede usar, de tal modo que diferentes arquitecturas de redes neuronales requieren diferentes algoritmos de aprendizaje.

Existen muchos tipos de aprendizaje dependiendo del modo en que es realizado el ajuste de los pesos. En un principio, los pesos pueden ser considerados parámetros libres, aunque es posible, si se conoce información acerca de la naturaleza del problema que se va a tratar, fijar restricciones a los valores iniciales del los pesos, o a los valores que puedan tomar a lo largo del proceso de aprendizaje.

Para lograr una buena generalización el conjunto de entrenamiento debe contener una cantidad de ejemplos al menos varias veces mayor que la capacidad de la red, o sea: Np > Nw/Ny, donde Np denota la cantidad de ejemplos de entrenamiento, Ny la cantidad de neuronas de salida y Nw cantidad de pesos. Esto se puede entender intuitivamente notando que si la cantidad de grados de libertad de una red (Nw) es mayor que la cantidad de restricciones asociadas con la función de respuesta deseada (Np*Ny), el procedimiento de entrenamiento sería incapaz para restringir completamente los

pesos en la red.

El paradigma de aprendizaje indica la forma en que el entorno influye en el proceso de aprendizaje. Así, el paradigma de aprendizaje puede ser:

Prescriptivo: El método de programación prescriptiva de la red se basa en fijar los pesos asociados a cada enlace, cada peso se calcula directamente, utilizando una expresión matemática que incluye entre sus términos la información almacenada en el conjunto de entrenamiento. Esto significa que al crear la red, junto con la definición de su topología se asignan pesos a los enlaces.

Supervisado: Se presentan los conocimientos en forma de pares de [entrada, salida deseada]. Al calcular la red una salida, esta es comparada con la salida deseada y de su diferencia se calcula un error con el que se corrigen los pesos hasta que la adaptación de estos de como resultado la salida deseada.

No supervisado: Durante este proceso de aprendizaje a la red no se le presenta la salida deseada y solamente se le introducen las entradas y los ejemplos para el entrenamiento, de forma tal que la red por sí sola halle los resultados correctos.

Por refuerzo: Es una mezcla entre los dos aprendizajes anteriores. El instructor o maestro exterior sólo indica cuando la salida es correcta o no, pero no indica en cuanto sediferencia de la salida buscada. Si se compara este paradigma con el supervisado, se observa que si bien el

supervisado proporciona una información relativa a la dirección en la que se deben realizar los cambios en el sistema (ajuste de los pesos), en el caso de un aprendizaje por refuerzo no se tiene información acerca de la "dirección" del cambio, lo cual hace que su ámbito de aplicación sea mucho más reducido comparado con el modo supervisado, aunque presenta interés en la comunidad científica dedicada al estudio de las máquinas capaces de aprender.

Aprendizaje híbrido: Se trata de una combinación del aprendizaje supervisado y del no supervisado. Parte de los pesos se ajustan por medio de un esquema de aprendizaje supervisado, y el resto se obtienen por medio de un aprendizaje no supervisado.

La teoría del aprendizaje mediante ejemplos conlleva tres aspectos muy importantes a tener en cuenta Vivaracho (2001): determinar la capacidad de aprendizaje, la complejidad de los ejemplos utilizados y la complejidad computacional del proceso en sí.

La capacidad es un concepto relacionado con la cantidad de patrones que pueden ser almacenados y qué funciones y contornos de decisión puede sintetizar una red neuronal artificial.

La complejidad de los ejemplos determina el número de los patrones de aprendizaje necesitados para entrenar la red de tal manera que quede garantizado un determinado grado de generalización.

Un escaso número de patrones de aprendizaje comparado con el número de pesos (parámetros libres), puede dar lugar a problemas de sobre entrenamiento.

La complejidad computacional se refiere al tiempo requerido para que el algoritmo de aprendizaje se aproxime a la solución usando los patrones de entrenamiento. Lo más corriente es que los algoritmos de aprendizaje sean computacionalmente muy complejos.

Implementación de las redes neuronales

En la búsqueda de sistemas inteligentes en general, se ha llegado a un importante desarrollo del software (en la actualidad ya existen lenguajes de procesamiento simbólico de la información).Sin embargo estos lenguajes se apoyan en arquitecturas convencionales de ordenadores.

Actualmente las direcciones de investigación consisten en la búsqueda de nuevas arquitecturas mas adecuadas para este tipo de tarea, Hilera (1995). Dentro de esta línea se encuentran algunos de los neurocomputadores más conocidos.

Un neurocomputador es básicamente un conjunto de procesadores conectados con cierta regularidad que operan concurrentemente. En la actualidad ya existen una serie de neurocomputadores comerciales destinados a la realización de redes neuronales; podemos citar el MARK III y IV, el ANZA, el ANZA PLUS o el Delta-Sigma. A pesar de los logros alcanzados

en este campo, está muy lejos el día en que estas estructuras alcancen el nivel de desarrollo y difusión de los ordenadores convencionales.

Por otro lado, otra forma completamente distinta de realizar redes neuronales consiste en la implementación de estas por medio de uno o varios circuitos integrados específicos, para así poder obtener una estructura que se comporte lo más similar posible a como lo haría una red neuronal. Aunque esta segunda aproximación se halla básicamente en manos de las universidades y centros de investigación, existen algunos productos comerciales como el N64 de Intel, chip que contiene 64 neuronas y 10 000 sinapsis y que puede procesar 2500 conexiones por segundo o el MB 4442 de Fujistsu, con una sola neurona y capaz de procesar 70 000 conexiones por segundo.

Como diferencias con respecto a las neurocomputadoras cabría resaltar el menor número de neuronas de los chips neuronales, consecuencia del hecho de que dentro del propio chip se incluyen todas las interconexiones y por tanto su velocidad resulta varios órdenes de magnitud superior que para los neurocomputadores. En la actualidad, estos chips se utilizan para aplicaciones en las que la red neuronal ya se ha diseñado y probado por otros métodos y lo que se requiere es una mayor velocidad para trabajar en tiempo real.

Aunque la tecnología microelectrónica parece que es actualmente la más adecuada para la realización de redes neuronales, existen varios problemas sin resolver, como es la

dificultad de obtener el alto grado de interconexión propio de estas redes o el problema de la entrada/salida masiva condicionada por el número de pines o por último, el poder conseguir sinapsis con pesos variables, necesarios si se quiere que la red tenga una verdadera capacidad de aprendizaje.

Otra tecnología que podría ser apropiada en la implementación de las redes neuronales es la tecnología electro-óptica, con la ventaja de utilizar la luz como medio de transporte de la información, permitiendo la transmisión masiva de datos. Los intentos de aplicación de esta tecnología en la realización de redes neuronales se encuentran todavía en estados de investigación básicos.

Realización de redes neuronales

1.- La realización de redes neuronales más simple e inmediata consiste en simular la red sobre un ordenador convencional mediante un software específico. Es un procedimiento rápido, poco costoso e insustituible por el momento para realizar el entrenamiento y evaluación de las redes, pero cuya mayor desventaja consiste en el hecho de que se intentan simular redes con un alto grado de paralelismo sobre máquinas que ejecutan secuencialmente las operaciones. Valores intrínsecos de las redes neuronales no se pueden obtener de esta forma.

2.- Realización de redes neuronales a través de arquitecturas orientadas a la ejecución de procesos con un alto grado de paralelismo, tales como redes de transputers, arquitecturas sistólicas, etc. El objetivo de estas redes es acelerar la simulación de la red neuronal permitiendo si es posible una

respuesta en tiempo real. Esta segunda línea puede verse como una optimización de la primera en cuanto al tiempo de proceso, pero subsiste el hecho de que el comportamiento real de la red sigue siendo simulado por una estructura ajena a la estructura intrínseca de una red neuronal.

3.- Una tercera aproximación radicalmente distinta es la realización de redes neuronales mediante su implementación por uno o varios circuitos integrados específicos. Se intenta de esta manera construir un elemento o conjunto de elementos que se comporten lo más similar a como lo haría una red neuronal. Son los llamados chips neuronales. Las neuronas y las conexiones se emulan con dispositivos específicos de forma que la estructura del circuito integrado refleja la arquitectura de la red. Se consigue de esta forma realizaciones que funcionan a alta velocidad, permitiendo en muchas ocasiones el proceso en tiempo real, pero a costa de una pérdida notable de flexibilidad.

Softwares

La comercialización de productos software ha sido y sigue siendo la forma más extendida de simular redes neuronales. La diferencia entre los distintos productos software radica en aspectos tales como el tipo y número de arquitecturas de red que soporta (SOM, Hopfield,etc), velocidad de procesamiento (número de conexiones por unidad de tiempo), interfaz gráfica, exportación del código para el desarrollo auténtico de aplicaciones, etc.

A la hora de decidirse por un software hay que tener muy claro el uso que se que se va a hacer del mismo. Es preferible elegir entre aquellos productos que se centran en pocas arquitecturas, pues en muchos casos cuando ofrecen una gran variedad de las mismas no todas las incluidas son plenamente funcionales. Otros aspectos importantes a tener en cuenta son la forma de leer los datos (si leen del disco en vez de la memoria, el proceso se hace largo y tedioso), la inclusión de módulos para el análisis de resultados o la posibilidad de generar código C incluyendo los pesos obtenidos.

En el mercado se encuentran productos tales como Neural Desk, de Neural Computer Sciences, para PC con sistema operativo Windows o Expert Net, de Image Soft, para PC con sistema operativo Windows.

Neurocomputadoras de propósito general

En las neurocomputadoras comerciales actuales se ha adaptado el concepto de memoria virtual, propio de los ordenadores convencionales a la idea de red neuronal virtual. Según esta idea, el usuario define la red neuronal que quiere simular y mapea dicha red sobre una serie de procesadores de manera que cada procesador simula en tiempos diferentes, distintas neuronas de la red. Esto obliga a que cada procesador debe tener acceso a cierta cantidad de memoria donde almacenará los estados de aquella neuronas que en algún momento van a contribuir con su estado a la activación de la neurona simulada por el procesador y los pesos a través de los cuales le llegarán estas contribuciones.

Un neurocomputador de propósito general estará formado por un conjunto de procesadores de forma que en cada paso, cada uno de ellos recupera de su memoria local los pesos Wij de los estados Xi de las neuronas conectadas a la neurona que en ese momento está simulando, realiza la suma de los productos y transmite el nuevo estado de al resto de los elementos a través del bus de comunicaciones. Paralelamente el procesador está recibiendo los estados actualizados del resto de las neuronas y los almacena en la memoria de nuevos estados. Cuando todas las neuronas hayan actualizado la memoria de nuevos estados se carga en la memoria de estados actuales, que servirá como base para el siguiente cálculo. De manera similar para poder trabajar con reglas de aprendizaje, cada procesador debe llevar asociado dos memorias de pesos, la de pesos actuales y la de pesos nuevos.

Las Neurocomputadoras de propósito general se pueden subdividir en dos grupos:

-Placas coprocesadores.
-Matrices de procesadores paralelos.

Estas dos categorías se diferencian básicamente en el número y complejidad de las unidades de proceso que utilizan, de forma que su frontera es a veces difícil de establecer. Son placas basadas en microprocesadores convencionales junto con cierta cantidad de memoria que se conectan a la PC y permiten acelerar significativamente la simulación de la

red neuronal.

Un ejemplo ilustrativo es Mark III, formado por un conjunto de procesadores Motorola 68020 asociados cada uno de ellos a un coprocesador de punto flotante 68881. Se distribuye con un paquete de software que se encarga de la distribución dinámica de los elementos de proceso de las neuronas virtuales. Soporta hasta 65000 neuronas y 1000000 de conexiones con velocidad de proceso de 45000 conexiones por segundo.

Se trata de matrices de unidades procesadoras conectadas con una topología más o menos regular. Son extensiones de las estructuras vistas hasta ahora.
Uno de estos productos es el CONE (computation network environment) de IBM, basado en el NEP (network emulation processor), un procesador expansible que actúa como coprocesador para PC. Se pueden interconectar hasta 256 NEP a través del bus de comunicaciones NEPBUS. El conjunto NEPCONE puede procesar 1000000 neuronas con 4000000 de conexiones. El NEP incluye un TMS320, RAM de datos de 64kx16, RAM para programa de 4kx16 y 3 interfaces. Cada NEP tiene 4000 neuronas y 16000 conexiones con velocidades de 30 a 50 actuaciones de la red por segundo.

Neurocomputadoras de propósito especial

Son aquellas que han sido diseñadas para implementar un modelo específico de red neuronal. Ejemplos de estas neurocomputadoras son el Sistema de reconocimiento de imágenes de Wisard o el coprocesador para redes

realimentadas del Centro Nacional de Microelectrónica y la Universidad Autónoma de Barcelona (CNM-UAB). CNM-UAB es un procesador conectable a PC, especialmente indicado para la emulación de redes neuronales realimentadas tipo pseudo-Hopfield con pesos discretos y 4 Mbits de RAM.

Tiene 4096 neuronas binarias, posibilidad de aprendizaje (pesos variables) y 10^9 sinapsis por segundo de velocidad de procesamiento.

Implementación microelectrónica VLSI

Debido a las dificultades para adaptar correctamente las grandes estructuras de procesamiento de la información a las necesidades de las redes neuronales, se han buscado soluciones basadas en la implementación VLSI de dichas redes. En general, si la red ha estado previamente simulada y su configuración de conexiones perfectamente determinada, se busca la implementación sobre un circuito de conexiones fijas.

El primer problema que supone la solución microelectrónica corresponde a la elección de los dos aspectos siguientes:

1-Implementación analógica o digital.
2-Arquitectura en matriz de neuronas o de sinapsis.

La respuesta a la primera cuestión tiene muchos matices y es necesario hacer consideraciones sobre los aspectos relacionados con la precisión, velocidad, consumo, inmunidad al ruido, memoria, capacidad de aprendizaje, etc. Sin embargo, ningún aspecto es definitivo para una u otra

alternativa. Existen algoritmos mejor adaptados a una solución u otra y problemas que se resuelven mejor con una alternativa que con otra.

En cuanto a la segunda interrogante, en la matriz de neuronas cada nodo o neurona esta localmente conectado a sus vecinos. El tipo de red y algoritmo deciden la necesidad de interconexión siendo este un parámetro muy fundamental, ya que consume una gran cantidad de área del chip .La densidad de conexionado y la conexión con el exterior (entrada/salida) limitan, en definitiva, la capacidad de integración sobre un dado de silicio. En una matriz de sinapsis se da una organización tipo PLA (programable logic array), con una serie de líneas horizontales correspondientes a los equipotenciales de entrada y una serie de líneas verticales correspondientes a las líneas de suma de las neuronas. Cada sinapsis es una multiplicación.

La presentación del panorama que actualmente existe en el ámbito de la implantación de las redes neuronales resulta complicada por la gran cantidad de aportaciones existentes tanto en soluciones analógicas como en digitales y con la alternativa de matrices de neuronas o de sinapsis.

Redes neuronales SOM

Las redes neuronales SOM (Self Organizing MAP o Mapas Autoorganizados) crean una correspondencia entre los datos de entrada y un espacio dimensional de salida, creando mapas topológicos de dos o incluso tres dimensiones, de tal

forma que ante datos de entrada con características semejantes se deben activar neuronas situadas en zonas próximas de la capa de salida.

Su arquitectura típica muestra una capa de entrada y una de salida. Las neuronas de la capa de entrada están conectadas a todas las de la salida y las neuronas de esta última capa están organizadas en forma de un espacio bidimensional en su representación habitual. Las neuronas adyacentes en la capa de salida ejercen una interacción entre ellas que está en función de la cercanía a la que se encuentren unas de otras, siendo esta interacción despreciable ante largas distancias. Esta zona de interacción puede ser circular, cuadrada, hexagonal u otro polígono regular centrado en dicha neurona.

Una de las características que hacen de esta red una de las más populares es el hecho de que use un método de entrenamiento no supervisado, por lo que no se le pasan salidas deseadas a la red y esta hace por sí sola a partir de los ejemplos del entrenamiento la clasificación de los datos en forma de mapa.

Esta es una red competitiva, donde al presentarle un vector de entrada a la red, ella evoluciona hacia una situación estable en la que se activa una neurona, la vencedora. En esta red las neuronas topológicamente próximas son sensibles a entradas físicamente similares. Por esta causa es especialmente útil para establecer relaciones desconocidas previamente entre conjuntos de datos.

Esta red necesita un entrenamiento prolongado y no aprende nuevos datos sin entrenamiento, observándose en ella un procedimiento de entrenamiento y prueba y luego el de aplicación propiamente dicho, no obstante, es una de las más populares y se utiliza ampliamente en el reconocimiento de patrones (de voz, texto, imágenes, señales, etc), codificación de datos, compresión de imágenes, resolución de problemas de optimización y en la robótica.

En el trabajo utilizaremos las redes neuronales SOM. Sus características la han hecho propicia para nuestra aplicación específica. En primer lugar está uno de sus mayores atractivos: usa un método de entrenamiento no supervisado, lo que nos evita el tener que encontrar criterios de clasificación, ya que la red los provee por sí misma.
Por otro lado, la red es capaz de adaptarse a cada patrón de usuario por separado, haciendo una clasificación de datos particular para cada cliente, lo que nos permite realizar un análisis personalizado de los consumos. Otra facilidad que provee es la forma de salida de sus datos en forma de mapa topológico, lo que facilita el posterior análisis de los resultados mediante la graficación de los datos de salida. Por las facilidades que brinda esta red es que la hemos elegido para la implementación del método.

Capítulo 2
Resultados y proceso metodológico

Entre las estrategias de investigación que utilizamos están la exploratoria y la explicativa, exploramos diferentes técnicas y tendencias en la creación y aprendizaje de patrones neuronales con vista a desarrollar un modelo computacional novedoso y que resuelva las deficiencias de otros reportados en la bibliografía.

Se enfoca en el entrenamiento de la red y la regeneración de los patrones para construir posteriormente los perfiles de usuario, en el análisis de las llamadas de los usuarios con alto consumo y el correspondiente análisis y detección de alarmas.
Actualización del perfil UPH con cada llamada (f=1 llamada) y bajo umbral Hellinger (H) para el lanzamiento de alarmas de cambio de comportamiento, Actualización del perfil UPH una vez por día (f= 1 día) y alto umbral Hellinger (H). (Bertona, 2005)

Utilizado para estudiar diferentes propuestas de sistemas de detección de patrones de fraudes y seleccionar la que más se ajusta a las condiciones actuales de la provincia. También nos ayudó a definir el tipo de comportamiento de los usuarios y definir patrones de acuerdo a nuestro comportamiento. De la misma forma se hizo uso de esta metodología para definir de manera precisa el problema, los objetivos y la justificación de la investigación

Utilizado para estudiar diferentes proyectos de patrones de fraudes en diferentes países y de esta manera formalizar una propuesta aplicable a la provincia de Los Ríos

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 los perfiles de las llamadas de los usuarios con alto consumo y el correspondiente análisis y detección de alarmas posteriormente los perfiles de usuario; la segunda prueba se enfocó en el análisis. 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). (Bertona, 2005)

Lo primero que se realizó es definir las variables independientes y dependientes que se consideraran en el cuasi experimento.

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 mismos 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 (Bertona, 2005)

El presente trabajo de investigación se centralizó en la Provincia de Los Ríos, es una de las 24 provincias de la República del Ecuador. Localizado en la región costa del país. Su capital es Babahoyo y su ciudad más poblada en Quevedo. Los principales cantones de la provincia en cuanto a población son: Quevedo, con 173.575 habitantes, Babahoyo (153.776 hab.), Vinces (71.736 hab.), Ventanas (71.093 hab.), Buena Fe (63.148 hab.) y Valencia (42.556 hab.).

En cuanto a los resultados obtenidos se realizó varios experimentos.

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

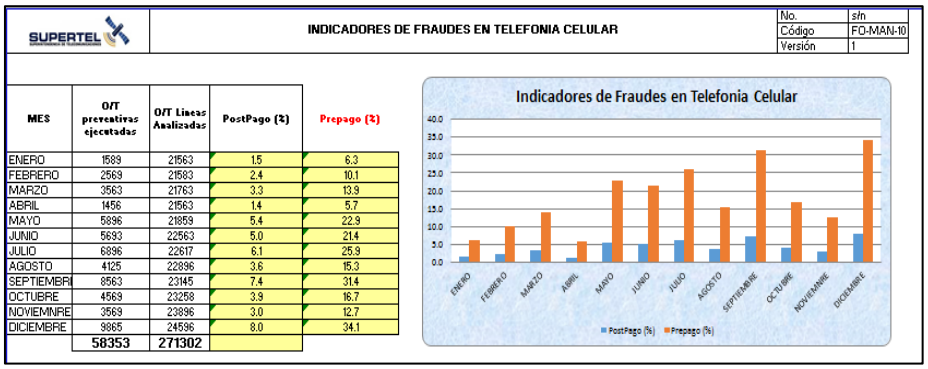
Dimensión de la red neuronal para clasificar llamadas locales (NLxML) = 12x12
Dimensión de la red neuronal para clasificar llamadas nacionales (NNxMN) = 8x8
Dimensión de la red neuronal para clasificar llamadas internacionales. (NIxMI) = 6x6
Tasa de aprendizaje estática (□) = 0,5
Distancia máxima de neurona (DVMAX) = 12
Los mismos valores definen la dimensión de los perfiles CUP y UPH:
Cantidad de patrones para clasificar las llamadas locales (PL) = 244
Cantidad de patrones para clasificar las llamadas DDN (PN) = 164

Cantidad de patrones para clasificar las llamadas internacionales (PI) = 136
Dimensión de los perfiles CUP y UPH (K) = 544

Una vez que se obtuvo resultados luego de los entrenamientos se tomó los valores de los perfiles y se los comparo e interpreto, 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, donde los resultados arrojados por el análisis histórico y presente de los CDR son utilizados por algoritmos fuzzy para mostrar valores que podrán ser analizados por los auditores de fraudes de las distintas operadores telefónicas. (Bertona, 2005)

Fraudes identificados en los servicios telefónicos

Para analizar los principales fraudes en telefonía celular y la incidencia y el impacto que tiene estos, se tomó datos del CAC de Babahoyo de CNT de la encuesta realizada para identificar los delitos en telecomunicaciones más comunes en el Ecuador. Entre los principales resultados de la encuesta se encuentra la clonación de teléfonos celulares, Call back y Fraude tercer país Según la SUPERTEL.



SUPERTEL

INDICADORES DE FRAUDES EN TELEFONIA CELULAR

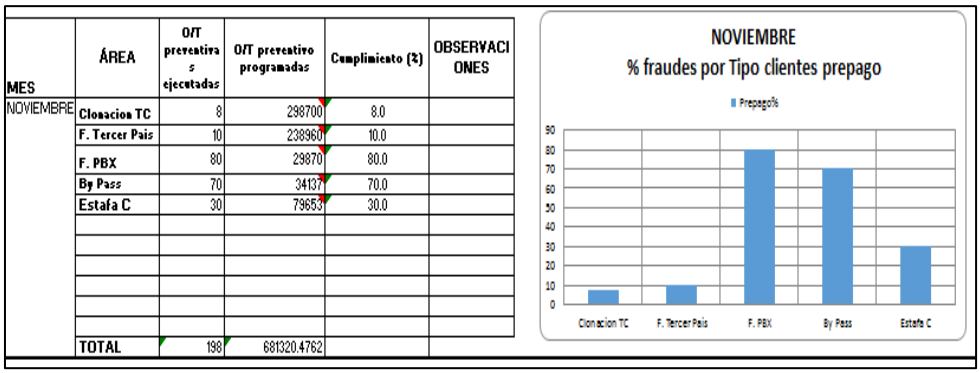
No.	s/n
Código	FO-MAN-10
Versión	1

MES	O/T preventivas ejecutadas	O/T Lineas Analizadas	PostPago (%)	Prepago (%)
ENERO	1589	21563	1.5	6.3
FEBRERO	2569	21583	2.4	10.1
MARZO	3563	21763	3.3	13.9
ABRIL	1456	21563	1.4	5.7
MAYO	5896	21859	5.4	22.9
JUNIO	5693	22563	5.0	21.4
JULIO	6896	22617	6.1	25.9
AGOSTO	4125	22896	3.6	15.3
SEPTIEMBRE	8563	23145	7.4	31.4
OCTUBRE	4569	23258	3.9	16.7
NOVIEMNRE	3569	23896	3.0	12.7
DICIEMBRE	9865	24596	8.0	34.1
	58353	271302		

Figura 4.1 Indicadores de fraudes en Telefonía Celular

Fuente: elaborado por el Autor

En los estudios realizados a partir de una encuesta aplicada a los usuarios de telefonía celular en la ciudad de Babahoyo se encontraron los siguientes tipos de fraudes.



MES	ÁREA	O/T preventivas ejecutadas	O/T preventivo programadas	Cumplimiento (%)	OBSERVACIONES
NOVIEMBRE	Clonacion TC	8	298700	8.0	
	F. Tercer Pais	10	238960	10.0	
	F. PBX	80	29870	80.0	
	By Pass	70	34137	70.0	
	Estafa C	30	79653	30.0	
	TOTAL	198	681320.4762		

Figura 4.1 Fraudes por tipo de clientes en telefonía móvil pre pagado

Fuente: elaborado por el Autor

Análisis y descripción del entrenamiento de patrones

Los sistemas de detección de fraude tienen que procesar cantidades considerables de información, por lo que los tiempos de acceso a las bases de datos deben ser mínimos. Es por esto que las bases de datos deben ser de acceso directo y se deben minimizar las lecturas al disco duro, pues estas tienen una mayor duración. Se deben hacer asignaciones de memoria dinámica de forma tal que se trabaje con la memoria RAM que es más rápida. Los algoritmos se deben optimizar para hacerlos más eficientes, así como también se necesita lograr un compromiso entre capacidad de almacenamiento y de procesamiento de la información, recursos que la PC dedicada a esta tarea debe tener en abundancia.

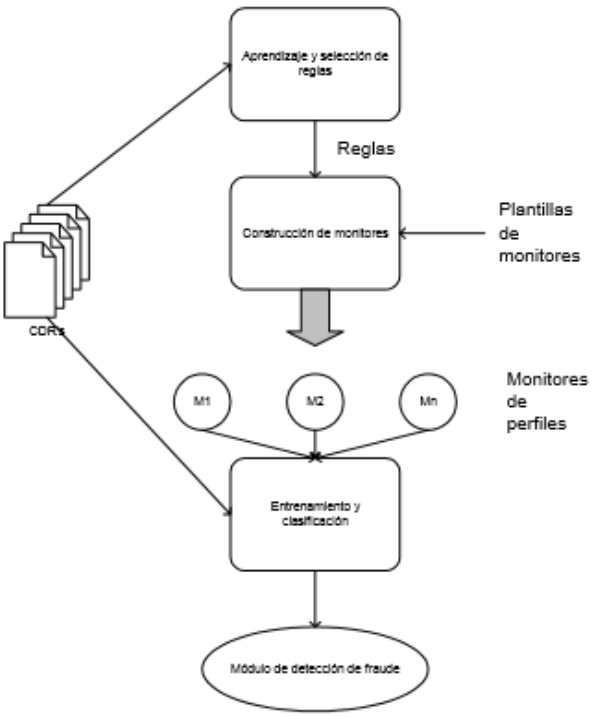


Figura 4.2 Esquema de las reglas

Construcción metodológica del objeto de Investigación.

Al generarse una llamada telefónica, la central genera registros de datos que contienen detalles referidos a una llamada celular que se está intentando realizar.

Esta información es transmitida hacia el operador de la red a través de las celdas o switches con los que el teléfono celular se estaba comunicando en un determinado momento y puede ser de diversos tipos, en dependencia del sistema utilizado (GSM, AMPS, etc) y del software analizador de tráfico. Estos datos son almacenados en formatos específicos de cada aplicación en formas no legibles a simple vista y deben ser procesados para extraer de ellos una información útil. Adicionalmente, muchos campos de estos datos son opcionales o se pueden representar en más de un formato de forma configurable a pesar de ser procesados por el mismo software. Estos datos son llamados CDR (call detail record o registro detallado de llamada) o tickets de llamadas y se utilizan con fines de facturación, análisis de tráfico y en diversas modalidades de detección de fraude. Los CDR contienen mucha información que puede servir para conocer el comportamiento del usuario y detectar usos fraudulentos del servicio. En dependencia de los campos que contengan los tickets se estudia cuáles de estos pueden ser indicadores de la comisión de un fraude. Para proteger la privacidad de los clientes, en varios sistemas se en cripta la información mientras es procesada por el detector.

Los sistemas existentes de detección de fraude intentan consultar secuencia de tickets 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 Tickets y son buenos para detectar casos extremos de la actividad fraudulenta. Los fraudes más sofisticados no se detectan mediante el análisis de un solo ticket, sino que se necesitan investigar una gran cantidad de tickets, por lo que analizar el uso absoluto de la red y los cambios en su comportamiento se hace necesario. (Bertona, 2005)

Para realizar un análisis diferencial, se monitorean patrones de comportamiento del teléfono celular comparando sus más recientes actividades con la historia de uso del mismo. (Bertona, 2005) La información sobre el comportamiento reciente del usuario es recogida en un perfil actual y la información de mayor tiempo se recoge en un perfil histórico. Estos perfiles contienen información condensada del comportamiento del usuario en vez de una secuencia de tickets. Cuando llega un nuevo ticket, los dos perfiles son actualizados de forma tal que la información en los tickets previos es adaptada. (Bertona, 2005)

Un cambio en el patrón de comportamiento es una característica sospechosa de ser un escenario fraudulento. (Bertona, 2005). Así se pueden establecer umbrales que se activan cuando los patrones de uso cambian significativamente en un corto período de tiempo. Hay muchas ventajas en realizar un análisis diferencial a través de perfiles de comportamiento de los usuarios. Ciertamente, ciertos

patrones de comportamiento que se pueden ver como sospechosos en uno y por lo tanto como indicadores de fraude pueden ser normales en otro. (Grosser, 2003)

Con un análisis diferencial se pueden desarrollar criterios flexibles que detecten cambios en el perfil de consumo basados en perfiles detallados del comportamiento de los usuarios. Esto lleva la detección de fraudes a niveles personales detectando así infracciones que sólo se pueden analizar a estos niveles donde el análisis absoluto no es factible. Como el usuario normal no es fraudulento, la mayoría de los criterios que lanzarían una alarma en el análisis absoluto se verían como un gran cambio en los patrones de consumo en el análisis diferencial. De esta forma el análisis diferencial puede verse como una aproximación del absoluto.

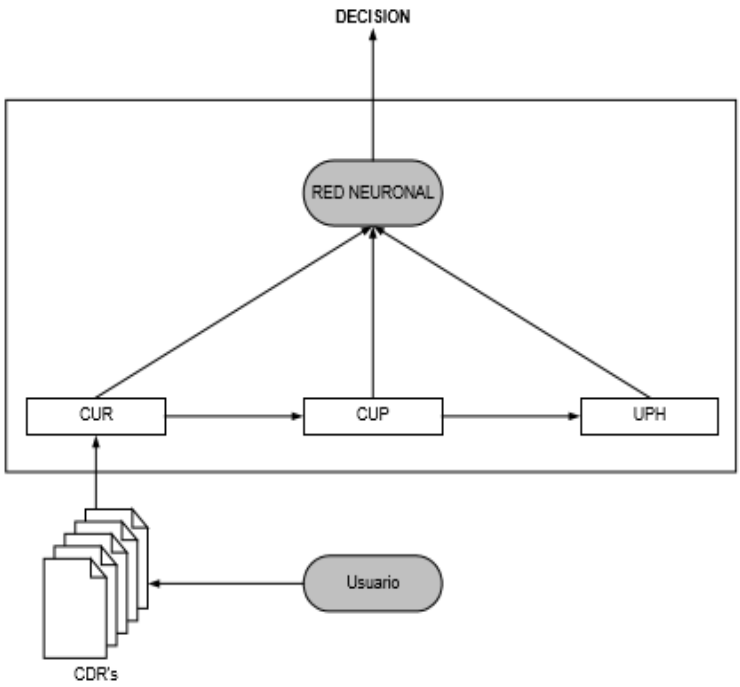


Figura 4.3 Enfoque basado en redes neuronales

Discusión de la información obtenida en relación a la hipótesis.

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". (Bertona, 2005) 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 (QL) 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 q u e 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

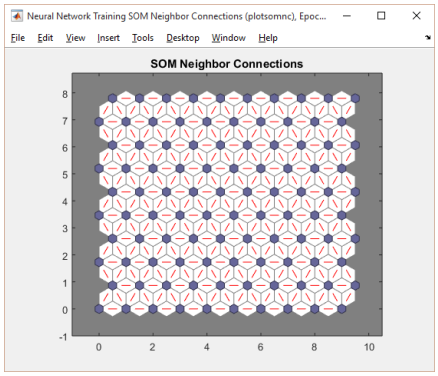


Figura 4.4 Conexión de neuronas
Fuente elaborada por el autor

Desarrollo del método

En el proceso de detección crearemos perfiles de usuario que nos mostrarán el comportamiento estadístico del individuo. Utilizaremos dos perfiles: el actual y el histórico del usuario. Para crear los perfiles entrenaremos una red neuronal de tipo SOM (Self Organizing Map o mapas autoorganizados), que creará los patrones de llamadas de los usuarios con los que conformaremos distribuciones de frecuencia. Estas distribuciones de frecuencia se hallan respecto al patrón establecido por la red como probabilidad de efectuar cada llamada mediante distancias ponderadas a cada uno de los puntos del patrón. Con estas distribuciones y distintos

coeficientes de adaptación llenaremos los perfiles. Luego, mediante distancias vectoriales compararemos el comportamiento actual del consumidor con su comportamiento histórico y estableceremos mediante umbrales optimizados la generación de alarmas.

La herramienta a utilizar para la implementación práctica del sistema es el Matlab matlab – Shortcut 2015R2

Los datos que extraeremos de los registros de llamadas serán el número telefónico en formato 000000, la fecha en formato AAAAMMDD, la hora en minutos 0000(desde 0 a 1440) y la duración de la llamada en segundos 00000(desde 0 hasta 86400).

La razón por la que se toma la hora en minutos es que el Matlab tiene como menor resolución $2.225073858507201 \times 10^{-308}$ y debemos sobreponernos a esta pequeña limitación de cálculo, pues de tomar la hora en segundos la distancia ponderada a calcular sobrepasaría estos límites. Por otra parte, si se toma la hora en el formato HHMMSS , al no ser este un valor decimal, las operaciones aritméticas como la resta no arrojarían un valor verídico y el patrón generado por la red neuronal tomaría distancias que realmente no existen entre hora y hora, produciéndose así una mala clasificación de los datos.

El que se obvien los segundos en la hora obedece al criterio de que los segundos en la hora de la llamada en cuanto a la descripción del comportamiento del consumidor no revisten

gran importancia, no así con la duración de la llamada, en la cual los segundos tienen gran significación.

Se confeccionaron varios programas en Matlab para llevar a cabo el proyecto. El primer programa, denominado baseorgs.m (Anexo 1), es el encargado de leer los registros de llamada que se encuentran en el fichero database.txt, procesarlos y devolverlos en el fichero datbaorg.txt, con los campos número telefónico, fecha, hora y duración en los formatos especificados anteriormente. Estos datos son almacenados de forma organizada, o sea, todas las llamadas telefónicas de un mismo número están una a continuación de la otra, cuestión esta que favorece el posterior análisis y procesamiento de las llamadas.

Una vez que se han transcrito los datos de forma legible y ordenada se procede al entrenamiento de la red SOM (mapas autoorganizados) para la clasificación de los datos y la creación de los patrones de llamadas. Para esto se aprovechan las funciones de Matlab que crean, entrenan e implementan redes neuronales para crear y entrenar la red utilizada en específico.

El segundo programa, neurorgm.m (Anexo 2), hace la lectura de la base de datos organizada en datbaorg.txt, entrena redes neuronales tipo SOM bidimensionales de 16, 25 y 36 neuronas con los datos hora y duración y almacena las neuronas en el fichero neuronas.txt. Este programa lee los datos del fichero y los guarda en una matriz, acelerando el proceso de lectura pues el fichero es leído una sola vez y

se utiliza la memoria RAM que es mas rápida que las lecturas al disco para el procesamiento de los datos. Las neuronas se guardan de forma secuencial para que puedan ser visualizadas para su estudio. Estas se guardan en el siguiente formato: número telefónico y neuronas.

Luego un tercer programa pfinalm.m (Anexo 3) se encarga de leer los datos a clasificar de datbaorg.txt y las neuronas de neuronas.txt, halla la distancia geométrica entre los puntos (hora, duración) y cada una de las neuronas y las pondera de acuerdo con la ecuación:

$$V=e^{-distanciai} / \Sigma_j e^{-distanciaj}$$

creando de esta forma una distribución de frecuencia con la probabilidad de que la llamada se efectúe con ese patrón y de forma tal que al ponderarse las llamadas más lejanas tengan una probabilidad mucho menor de pertenencia al patrón.

Después se procede a la actualización de los perfiles. Estos son del mismo tamaño que la red neuronal que la originó. Se llenan a partir de la distribución de frecuencias creada anteriormente. Cada miembro del perfil actual se multiplica por un grado de adaptación y se llena con la distribución hallada, según la fórmula:

$$CUPi=axCUPi+(1-a)xVi.$$

Al actualizarse el perfil actual se compara con el perfil histórico

utilizando una distancia vectorial. En este caso utilizamos la distancia de Hellinger al cuadrado, usada para comparar vectores de distribuciones de frecuencia, que nos ofrece la ventaja de que su valor se encuentra siempre entre cero y dos, cero para distribuciones iguales y dos para distribuciones ortogonales, Costa (2002). La distancia mencionada anteriormente tiene la fórmula:

$$h=\Sigma(\sqrt{CUPi}-\sqrt{UPHi})^2.$$

El programa analiza los distintos umbrales que se pueden establecer contando las alarmas repetidas y la cantidad de usuarios que generan alarmas. Usa como criterio para la evaluación de las alarmas que los días de las llamadas analizadas sean distintos, aunque sepueden usar varios criterios como evaluar las alarmas a una hora determinada del día en que haya poco tráfico, etc.

Luego se procede a la actualización del perfil histórico mediante una adaptación del perfil actual según la fórmula:

$$UPHi=\beta xUPHi+(1-\beta)CUPi$$

El programa hace un corrido de las tasas de adaptación a, β y los umbrales y guarda los resultados en el fichero results.txt para que estos puedan ser analizados.

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 (QL) 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

Construcción de perfiles y detección de cambios de comportamiento

El análisis diferencial necesita información acerca de la historia reciente del usuario y la más antigua. Para esto crea perfiles de usuario.

Un perfil es un modelo de un objeto (una representación compacta que describe las características más importantes), creado en la memoria del ordenador, y que es utilizado como representante del objeto en las tareas computacionales. Estas tareas incluyen por ejemplo: comparaciones, almacenamiento, resúmenes y análisis. Los perfiles cumplen la tarea de describir las características más vitales del objeto en un formato compactado, para que los programas los puedan usar en sus tareas.

Hay tres métodos principales para crear perfiles: el método explícito o manual; el método colaborativo y el método implícito. En el método explícito o de creación manual, los datos son introducidos por el usuario escribiéndolos directamente en su perfil de usuario, respondiendo a formularios, etc. También se puede crear y modificar un perfil a partir de la interacción colaborativa con otros perfiles, con los que se relaciona, recurriendo a conocimiento específico del domino y heurísticas inteligentes.

En el método implícito, se extraen/crean y modifican/actualizan

automáticamente los perfiles, recurriendo normalmente a técnicas de Inteligencia Artificial para realizar estas tareas. Es importante señalar que estos métodos no son rígidos, y muchas veces se utilizan simultáneamente como métodos híbridos, para producir perfiles más precisos y comprensibles. En nuestro caso usaremos un método híbrido entre el implícito y el colaborativo.

Una forma de clasificar el perfil de usuario, es considerar su comportamiento con relación al cambio. Alguna información del perfil de usuario es estática, como la fecha de nacimiento, el nombre, el número telefónico, etc, y puede ser introducida manualmente. Otra sin embargo es dinámica, como por ejemplo el comportamiento de las llamadas del usuario, que cambia y por consiguiente es aconsejable que sea determinada automáticamente.

Esto significa que para obtener un perfil más actual y preciso, es necesario acompañar las acciones del usuario de la forma más cercana posible. Por eso se recoge, procesa y guarda información de las acciones del usuario, en nuestro caso según los campos útiles de los tickets de las llamadas telefónicas, que sirven para proceder a las depuraciones y actualizaciones que se tengan que realizar. Hay otro conjunto de aspectos que pueden condicionar la creación de los perfiles, como son los objetivos del profiling de usuario, el dominio de aplicación, etc. Esto refuerza la necesidad de emplear técnicas que automaticen de forma inteligente las tareas de creación y gestión de los perfiles de usuario. Aquí es donde se aplican un sinnúmero de técnicas de

inteligencia artificial, desde agentes inteligentes, sistemas expertos hasta redes neuronales. Estos brindan la facilidad de realizar estas clasificaciones de los datos, en nuestro caso incluso sin tener reglas fijas o predefinidas, basándose solo en el comportamiento habitual del cliente.

Una aproximación inicial para la creación de nuestros perfiles puede ser extraer información útil de los tickets y guardarla en un determinado formato en perfiles de usuario que no son más que un resumen de su actividad. La información reciente se guarda en un registro de actividad reciente mientras que la actividad pasada se guarda en un registro de actividad histórica. Ambos perfiles se pueden mantener como dos colas finitas en las que al llegar un nuevo ticket la entrada más vieja del perfil histórico es descartada y la más vieja del perfil actual pasa a ser la más nueva del perfil histórico, Moreau(1997). Claramente esto no es óptimo para ver el comportamiento de los perfiles por lo que más efectivo es hacer un registro acumulativo en el que no hay necesidad de guardar el ticket y en el que para cada nueva entrada habría que hacer una distribución de frecuencias y atenuar los perfiles según coeficientes de adaptación que representan la magnitud en que cambia el comportamiento recogido en el perfil ante las nuevas actividades. Esto nos permitiría representar el perfil como una distribución de frecuencias y nos daría la certeza de que no habría una atenuación de los perfiles, de forma que nunca se harían cero.

Hay otros dos requerimientos importantes para confeccionar los perfiles. Deben ser eficientes en cuanto a cómo guardar

la información de los usuarios, llegándose a un compromiso entre la efectividad del sistema para la representación del comportamiento del usuario y la capacidad de almacenamiento; no podemos olvidar que nos estamos enfrentando a cantidades realmente considerables de datos. Esto nos debe hacer pensar incluso en el formato en que guardaremos la información a nivel de número de bytes de las variables utilizadas y en la cantidad de elementos con que logramos satisfacer nuestros niveles de detección de forma óptima.

El otro requerimiento es en cuanto a la actualización. Esta debe tener los niveles adecuados de adaptación para lograr una real caracterización de la conducta del usuario. El objetivo del perfil es realizar una precisa descripción de la conducta del usuario para facilitar una detección de fraudes confiable. Toda la información que se necesita para la detección de los fraudes esta derivada de los tickets.

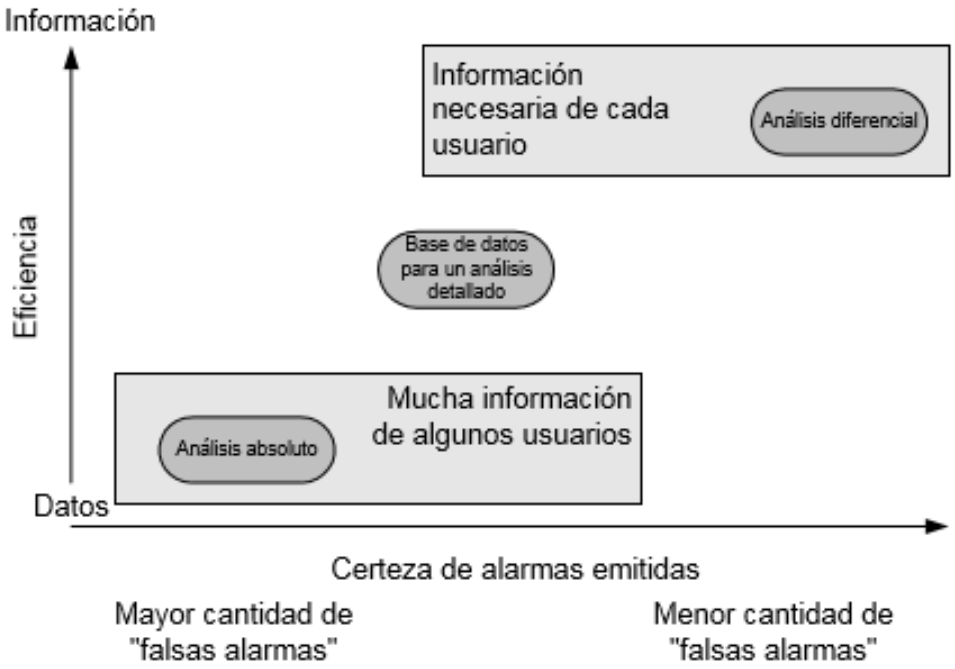


Figura 4.5 Análisis diferencial

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

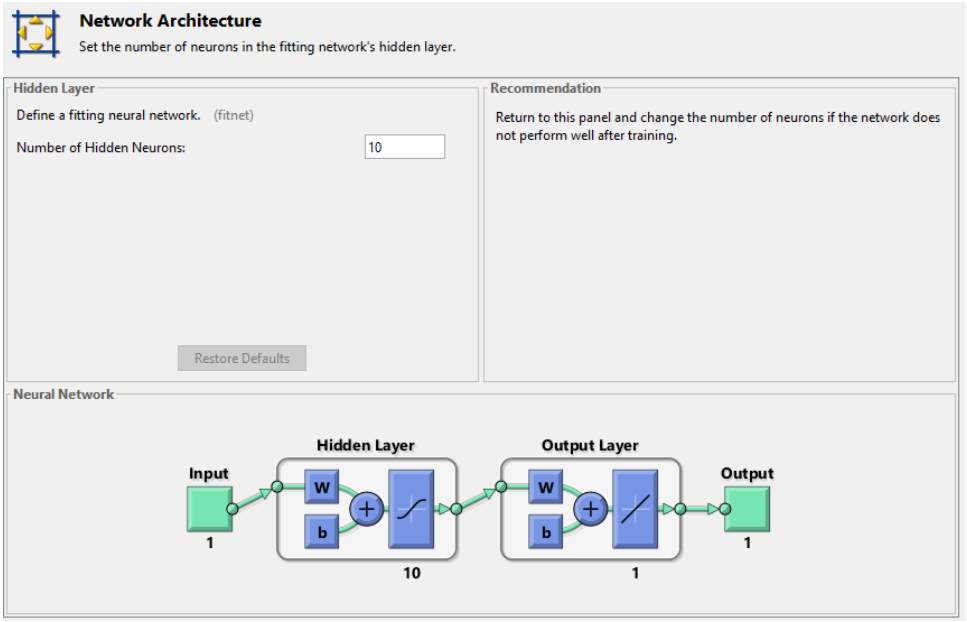


Figura 4.6 Contruccion de la red neuronal Matlab neural tool

Fuente elaborada por el autor

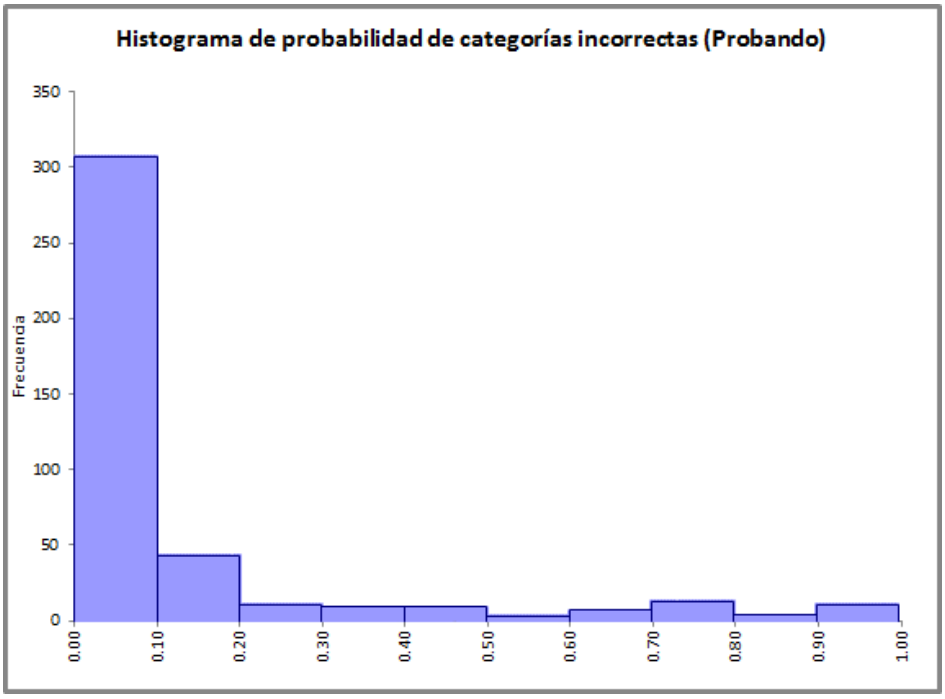


Figura 4.7 Histograma de entrenamiento neural tool

Fuente elaborada por el autor

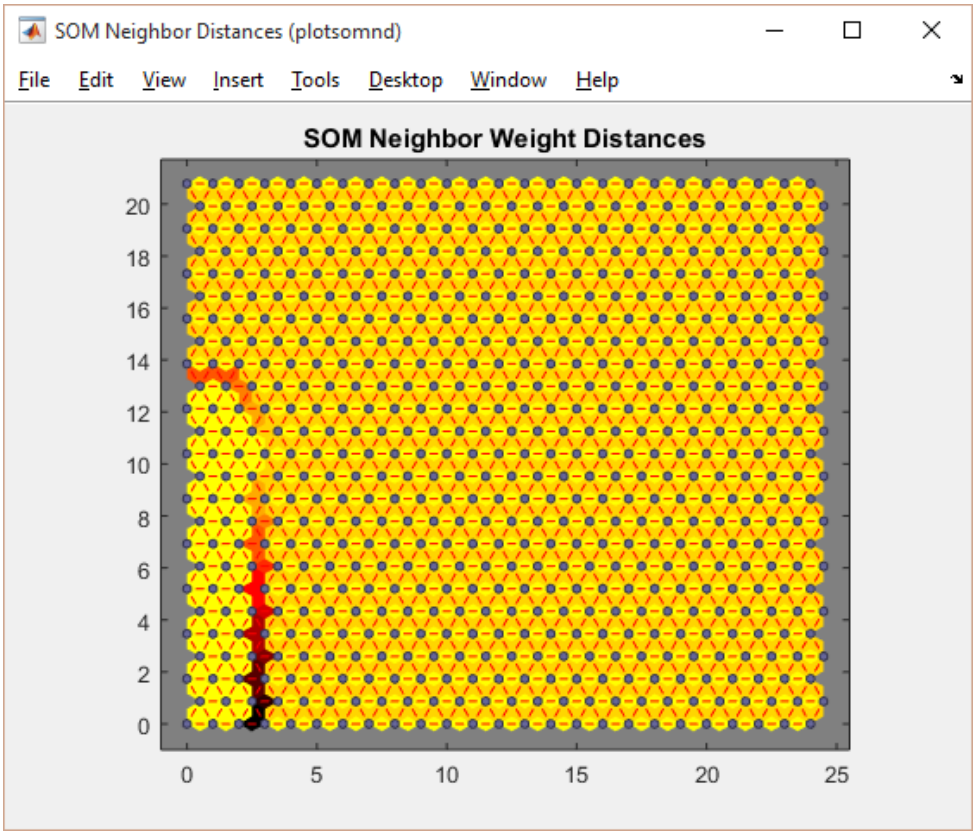


Figura 4.8 Distancia Matlab Neural tool

Fuente elaborada por el autor

Descripción de la Información Obtenida

Muchos sistemas de detección de fraude se basan en la creación de perfiles de usuario y la detección de anomalías en el patrón de consumo del usuario, las cuales indican posibles comportamientos fraudulentos que son valorados por expertos para analizar si efectivamente tuvo lugar un hecho delictivo.

El sistema específico elabora dos perfiles de usuario: el perfil actual (CUP o current user profile y el histórico (UPH o user profile history). Utilizando una red neuronal de tipo SOM se hace una clasificación de los datos durante su etapa de aprendizaje de forma tal que se hace una caracterización de un patrón de llamadas.

Luego se halla la distancia entre cada llamada y todos los puntos del patrón, se ponderan estas distancias exponencialmente y se genera una distribución de frecuencias de acuerdo a la probabilidad de ocurrencia de cada elemento del patrón. Con esta distribución de frecuencias se actualiza el perfil actual de usuario al multiplicarlo por un factor de adaptación y el perfil histórico se actualiza con otro factor de adaptación y el perfil actual. Estos dos perfiles se usan para compararlos y analizar si las distribuciones de frecuencia han tenido cambios apreciables, indicativo de que ha variado la conducta actual con respecto a la histórica. Esta comparación se hace mediante el uso de distancias vectoriales y luego de actualizar el perfil actual. Finalmente se establecen umbrales de generación de alarmas, los cuales deben ser óptimos al igual que los factores de adaptación para el funcionamiento eficiente del método.

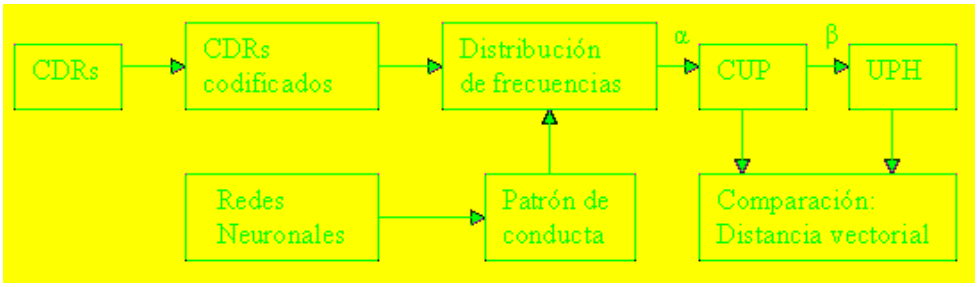


Figura 6.1 Esquema del algoritmo a utilizar

Fuente elaborada por el autor

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 (NLxML) = 12x12

Dimensión de la red neuronal para clasificar llamadas nacionales (NNxMN) = 8x8

Dimensión de la red neuronal para clasificar llamadas internac. (NIxMI) = 6x6

Tasa de aprendizaje estática (□) = 0,6

Distancia máxima de neurona "vecina" afectada (DVMAX) = 10

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

Cantidad de patrones para clasificar las llamadas locales (PL) = 144

Cantidad de patrones para clasificar las llamadas DDN (PN) = 64

Cantidad de patrones para clasificar las llamadas internacionales (PI) = 36

Dimensión de los perfiles CUP y UPH (K) = 244

NeuralTools: Entrenamiento, Auto-Prueba y Auto-Predicción de red neuronal
Ejecutado por: Palisade
Fecha: miércoles, 07 de septiembre del 2015 03:09:05 p.m.
Conjunto de datos: Datos de central
Red: Red entrenada en Datos

Resumen	
Información de red	
Nombre	Red entrenada en Datos Celulares
Configuración	Predicción de categoría PNN
Localización	Este libro de trabajo
Variable de categoría independiente	2 (Número de días desde que el cliente se suscribió a esta compañía, Plan internacional)
Variables numéricas independientes	11 (Duración de la cuenta, Plan de buzón de voz, Mensajes de correo de voz, Minutos por la mañana, Llamadas por la mañana, Minutos por la tarde, Llamadas por la tarde, Minutos por la noche, Llamadas por la noche, Minutos internacionales, Llamadas internacionales)
Variable dependiente	Var. de categoría (Llamadas al servicio al cliente)
Entrenando	
Número de casos	1667
Tiempo de Entrenamiento	00:01:21
Número de pruebas	127
Razón de la parada	Auto-Parada
% de predicciones incorrectas	1.4397%
Probabilidad incorrecta media	4.5565%
Desviación estándar de probabilidad incorrecta	10.3009%
Probando	
Número de casos	417
% de predicciones incorrectas	9.1127%
Probabilidad incorrecta media	12.5026%
Desviación estándar de probabilidad incorrecta	23.0434%
Predicción	
Número de casos	10
Predicción en Vivo activada	SÍ
Conjunto de datos	
Nombre	Datos de abandono
Número de filas	2094
Etiquetas manuales de caso	NO
Análisis de impacto de variable	
Mensajes de correo de voz	34.0486%
Llamadas internacionales	32.0212%
Llamadas por la mañana	11.9903%
Llamadas por la tarde	9.7807%
Minutos internacionales	4.1038%
Número de días desde que el cliente se suscribió a est	3.2734%
Plan internacional	2.0043%
Plan de buzón de voz	0.7120%
Llamadas por la noche	0.6951%
Duración de la cuenta	0.5744%
Minutos por la noche	0.3834%
Minutos por la tarde	0.3229%
Minutos por la mañana	0.0898%

Figura 4.9 Resultados del entrenamiento neural tolos

Fuente elaborada por el autor

Limitaciones

Este método se confeccionó para hacer la detección de patrones de fraude cuando existen cambios en los patrones de consumo del cliente, por lo que si el perpetrador mantiene un patrón estable de consumo y la red fue entrenada cuando ya estaba cometiendo el delito no será detectado. La red debe ser entrenada con consumo normal del usuario, de ser entrenada con información fraudulenta y mantener el criminal una conducta estable, este no será detectado.

Para sobreponerse a este inconveniente se puede hallar una correlación entre el consumo promedio de los usuarios y analizar los consumidores que se desvían de este promedio. Para esto se podría implementar una red que tomaría como entrada todos los parámetros de las demás redes para hallar el consumo promedio de los usuarios de la central. Mediante el uso de distancias vectoriales se compararía el patrón estándar con los demás para clasificar el tipo de consumidor y se analizarían los casos que se alejen del promedio. Esto sirve además para hacer estudios acerca del comportamiento de los clientes.

Por otro lado, si los patrones de consumo de los usuarios de la compañía tienen cambios apreciables de forma permanente, las redes deben ser reentrenadas. En días festivos en los que los usuarios tienen cambios en su conducta habitual se dispararán varias alarmas. Es por este motivo que se deben tener factores de adaptación especiales para los días festivos. Esto también

sería efectivo para disminuir la cantidad de alarmas puesto que una vez que esta es lanzada tiene un tiempo de latencia hasta que el perfil es adaptado correctamente a los cambios; este tiempo disminuiría al cambiar la adaptación de forma flexible.

Eficiencia

Hay una inmensa cantidad de datos con una porción muy pequeña de fraudes, lo que indica que estadísticamente estos datos están dispersos por lo que deben escrutarse con mucho cuidado para no molestar al usuario normal por equivocaciones del sistema.

Existen dos tipos de errores principales: falsas alarmas y casos sin descubrir. Los costos de los dos errores son diferentes. Frecuentemente las alarmas necesitan verificación por parte de expertos y son puestas en una cola, por lo que las falsas alarmas causan pérdidas de tiempo, mientras los casos sin descubrir causan pérdidas monetarias. Los llamados métodos sensibles al costo que respetan estos tipos diferentes de costo se deben usar. Las distribuciones dispersas y la necesidad de los sistemas sensibles al costo hacen que la evaluación del rendimiento de los métodos de detección de fraudes sea difícil, incluso en los métodos más sencillos.

Parámetros utilizados para la construcción de perfiles y detección de cambios de comportamiento, (Bertona, 2005) los valores utilizados para la construcción de perfiles y detección de alarmas fueron los siguientes:

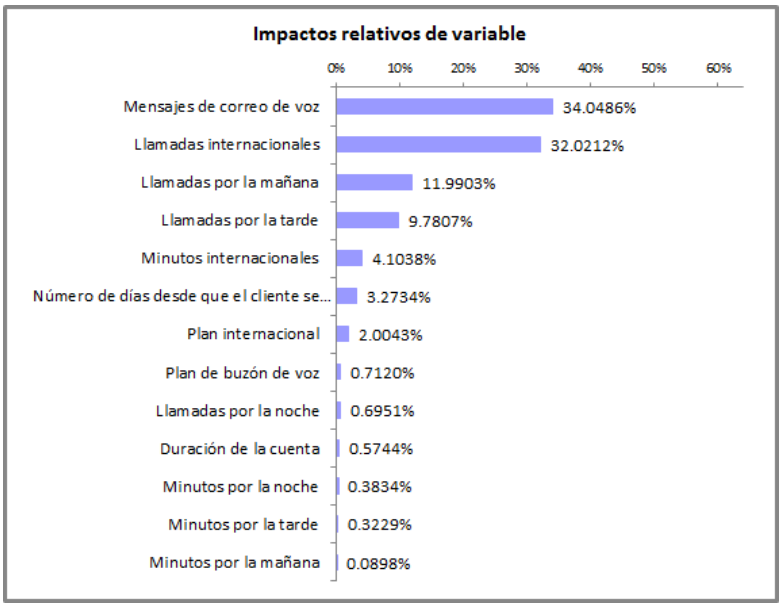


Figura 4.10 Porcentajes de resultados neural tolos

Fuente elaborada por el autor

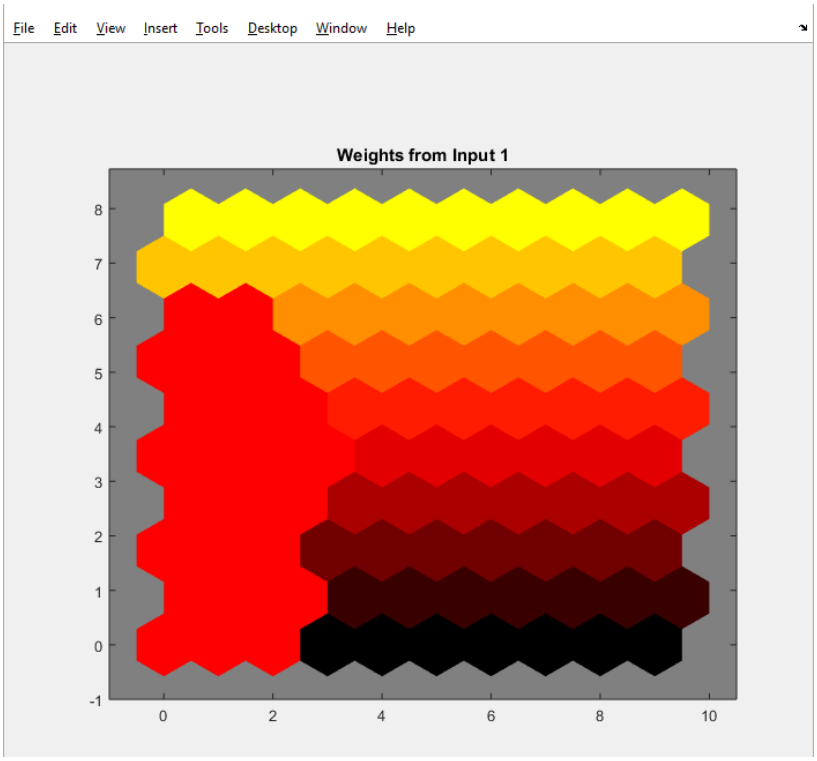


Figura 4.11 Resutados de llamadas realizadas MatLab Neural Tools

Fuente elaborada por el autor

Validación

Para la validación de los resultados alcanzados se procedió a formar una base de casos de llamadas sospechosas, implementada usando técnicas fuzzy. Los conjuntos fuzzy se usan para clasificar valores que pueden pertenecer a diversos grupos dependiendo de su valor con cierto grado de duda. Un ejemplo típico ilustrativo de esta situación es la altura, pues se puede ser bajo, mediano o alto con mayor o menor certeza en dependencia del valor de la variable, como se muestra en la figura 6.2.

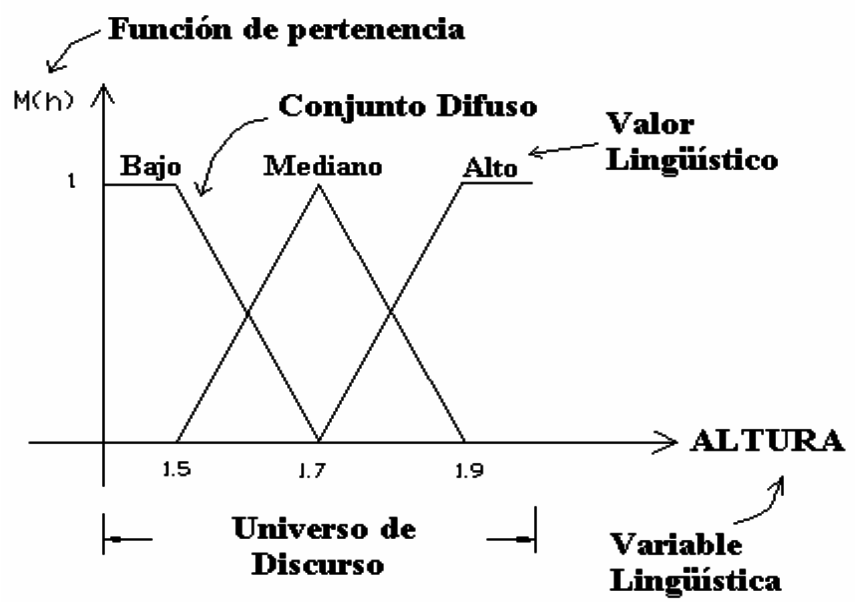


Figura 6.2 Conjunto fuzzy

Las técnicas fuzzy permiten que a una variable lingüística (se le dice así porque es descrita verbalmente), con un determinado universo de discurso o rango de valores permisibles, se le atribuyan valores lingüísticos que son "subclasificaciones" de la

misma variable modeladas según distintas funciones de pertenencia, Solsona (2003). La función de pertenencia asocia a cada elemento de un conjunto difuso el grado con que pertenece al valor lingüístico asociado. De esta forma se hace una clasificación de la misma variable en dependencia del grado de pertenencia que tenga con las "subclasificaciones" que de ella se hacen, como se ilustra en la figura 3.2.

Para formar esta base de casos antes se procedió a la conformación del patrón estándar de consumo de la central en sí, usando una red neuronal tipo SOM. Esta red tuvo como entradas los campos hora y duración al igual que las anteriores, pero los datos con los que se entrenó fueron los puntos de las redes individuales halladas anteriormente, con lo que se hizo una generalización del comportamiento de todos los usuarios (Anexo 4). Posteriormente a la creación de este patrón se inició el proceso de búsqueda de reglas para la implementación por medio de algoritmos fuzzy de la base de casos. En la creación de los algoritmos fuzzy se utilizaron las facilidades que nos brinda el Matlab, el cual contiene un toolbox para el trabajo con los conjuntos borrosos.

En la confección del programa (Anexo 5) se tomaron como variables de entrada la duración de las llamadas y la hora en que estas se efectuaron. Cada variable tiene tres funciones de membresía. En el caso de la duración esta tuvo como funciones de membresía cortas, normal y largas, definidas mediante las funciones zetamoidal, gaussiana y sigmoidal respectivamente. La hora tuvo como funciones de membresía temprano, común y tarde, definidas según las curvas

zetamoidal, gaussiana y sigmoidal respectivamente. En la salida se definió una sola variable: comportamiento, con funciones de membresía habitual, dudoso y alarma, las que a su vez estaban definidas por las funciones zetamoidal, gaussiana y sigmoidal.

En este trabajo se han cumplido los objetivos trazados al comienzo del mismo. Para ello se realizó un estudio del fraude en la telefonía celular y de las técnicas más usadas para combatirlo. Además se hizo una revisión acerca de las características fundamentales de las redes neuronales para escoger una red específica para la aplicación. Una vez escogida la red a utilizar, se implementó el algoritmo de detección y se obtuvieron los resultados que luego se validaron usando técnicas fuzzy para crear una base de casos con la que se compararon los resultados. El fraude en la telefonía móvil en ecuador no alcanza los niveles encontrados en otros países, fenómeno debido a la pequeña cantidad de suscriptores y a factores culturales y tecnológicos.

Las redes neuronales son una variante válida a considerar en la búsqueda de sistemas antifraude en el mundo moderno. La aplicación aquí presentada permite hacer estudios de consumo, estableciendo redes estándar que modelan el comportamiento general de los clientes. Los grupos de estudio se pueden conformar luego en dependencia de cuánto se alejan los usuarios de este patrón generalizado.

Por otra parte, se pueden hacer estudios a nivel personal del consumo de cada cliente mediante consultas a su perfil de

usuario, el cual presenta la ventaja de tener un formato de salida que permite su graficación con facilidad, de forma tal que la verificación de las alarmas se puede realizar analizando los perfiles individuales de forma gráfica sin la necesidad de hacer análisis de datos.

Los algoritmos tienen valores de adaptación variables que pueden ser manipulados a criterios propios, convirtiéndose en un instrumento flexible que puede ser adaptado a diversos criterios regionales, puesto que el fraude es un fenómeno que tiene diversas manifestaciones propias de la ubicación geográfica.

Las redes aquí empleadas para la conformación de los patrones utilizan criterios propios para hacer la clasificación de los datos de consumo del usuario a nivel personal, haciendo un análisis diferencial con el que se detectan cambios a nivel individual del comportamiento habitual del cliente.

Existen valores de adaptación y umbrales óptimos propios para cada red en dependencia del número de neuronas que esta posea y de los valores permisibles de las variables de entrada.

Descubre tu próxima lectura

Si quieres formar parte de nuestra comunidad,
regístrate en **https://www.grupocompas.org/suscribirse**
y recibirás recomendaciones y capacitación

compAs

Grupo de capacitación e investigación pedagógica

@grupocompas.ec
compasacademico@icloud.com

compAs
Grupo de capacitación e investigación pedagógica

@grupocompas.ec
compasacademico@icloud.com

ISBN: 978-9942-33-170-0

@grupocompas.ec
compasacademico@icloud.com

compAs
Grupo de capacitación e investigación pedagógica