INVESTIGACIÓN

Lógica Difusa para la detección de comportamientos anómalos en los sistemas de telefonía

Fuzzy logic for the detection of anomalous behavior in IP telephony systems

Ing. Daynet Álvarez Eng[1*], Dr.Sc. Godofredo Garay Álvarez[2], Msc. Yoan Martínez López[3].

Recibido: 11/2017 | Aceptado: 03/2018

PALABRAS CLAVE

Lógica Difusa

Telefonía IP

Minería de Datos.

RESUMEN

En esta investigación se exponen los principales fraudes y ataques detectados según la revisión bibliográfica, en las infraestructuras de telefonía IP. Se explican las técnicas de Minería de Datos (Data Mining), más utilizadas en los últimos años, y que son empleadas para detectar comportamientos anómalos en las redes de datos, como son las técnicas de Lógica Difusa. Además, se realizaron varios estudios y experimentos con las herramientas de minería de datos Open Source que se utilizan a nivel mundial, como el WEKA —*Waikato Enviroment for Knowledge Analysis*— y el KEEL —*Knowledge Extraction based on Evolutionary Learning*—, combinando varios algoritmos de lógica difusa con otros pertenecientes a los Árboles de Decisión y Reglas de Decisión, que mostraron experimentalmente los algoritmos a implementar para la detección de comportamientos anómalos.

KEYWORDS

Fuzzy logic

IP telephony,

Data Mining.

ABSTRACT

This research exposes the main frauds and attacks detected according to the literature review, in the IP telephony infrastructures. The data mining techniques (Data Mining), most used in recent years, are explained and used to detect anomalous behavior in data networks, such as Fuzzy Logic techniques. In addition, several studies and experiments were carried out with the Open Source data mining tools that are used worldwide, such as the WEKA -Waikato Environment for Knowledge Analysis- and the KEEL-Knowledge Extraction based on Evolutionary Learning-, combining several algorithms of fuzzy logic with others belonging to Decision Trees and Decision Rules, which experimentally showed the algorithms to be implemented for the detection of anomalous behaviors.for the realization of the experiments, traffic samples, of two types, according to the raised by the literature.

Introducción

Con la convergencia de las redes de telecomunicaciones, la voz analógica está migrando a la tecnología IP (Internet Protocol), surgiendo así la voz sobre IP (VoIP) y las empresas comienzan a emplear múltiples servicios de comunicaciones telefónicas basadas en IP sobre la infraestructura de las redes de datos ya establecidas. Esto posibilita que la adopción de sistemas telefónicos

1* Dirección de Operaciones de Seguridad, ETECSA, Cuba, daynet.alvarez@etecsa.cu

2 Profesor de la Universidad de Camagüey en la Facultad de Informática, Cuba

3 Profesor de la Universidad de Camagüey en la Facultad de Informática, Cuba

empleando IP sea especial para amortizar el costo del alquiler o arrendamiento de los canales de datos, realizar un mejor aprovechamiento y utilización de los mismos, mayor capacidad en el control de los gastos de las llamadas telefónicas, tener una plataforma que integre las funciones de tarifación, gestión de llamadas, gestión de líneas, inventario del parque de líneas telefónicas, la realización de llamadas telefónicas ordinarias entre otras características. (Amparo, 2011)

Entre los protocolos más utilizados para la implementación de infraestructuras de telefonía IP, están los protocolos SIP –*Session Initiation Protocol*– y la familia de protocolos H.323, regulados y aprobados por la UIT. Además, existen diversas configuraciones de otros protocolos y elementos de red que pueden ser utilizados igualmente para soluciones de telefonía IP, pero, por no estar debidamente regulados, generan malas configuraciones. Normalmente, la VoIP utiliza las IP de las redes y está sujeto a vulnerabilidades propias del medio, donde cada componente de red, dígase enrutadores (routers), conmutadores (switches), pasarelas (gateways) y cortafuegos (firewalls), tienen sus propios problemas de seguridad. La adición de tráfico de voz a estos componentes aumenta su lista de vulnerabilidades (Jared Ring, 2006).

Una revisión general de la literatura explica que un comportamiento anómalo es una manera o forma de actuar que no es usual, y si estas formas no usuales son utilizadas para violentar lo establecido, entonces es considerado fraude, ataques, amenazas, violaciones de seguridad, entre otros. A continuación se menciona y explica brevemente los principales ataques que se realizan a los sistemas de telefonía IP, a través de las redes de datos. (Ram Dantu May, 2009)

Todos estos tipos de ataques son realizados, de forma general, con la finalidad de:

- Evitar el completamiento de un gran número de llamadas, robando así minutos de llamadas a las compañías de telecomunicaciones, lo que afecta la imagen corporativa y sus ingresos.
- Escuchar a escondidas estos canales (Sniffer) y así obtener datos personales y/o de empresas, con el objetivo de venderlo a terceras personas.
- Según datos ofrecidos por CFCA –Communications Fraud Control Association– y FIINA –Forum for the International Irregular Network Access–, en el 2014 los Robocalling (robo de llamadas) y otros tipos de ataques maliciosos, constituyeron 20 millones de horas de trabajo SMBs –Sever Message Block, protocolo de red que pertenece a la capa de aplicación del modelo OSI–.

Además, durante el año 2015, los métodos de fraudes que ocasionaron pérdidas billonarias a las Empresas de Telecomunicaciones fueron: IP PBX Hacking y PBX Hacking (ataques a Pizarras con tecnología IP), *Abuse of Network - device or configuration weakness* (Abusos en la red, aprovechando debilidades en la configuración de los dispositivos) y el Phishing / Pharming (robo de datos bancarios por medio de Internet). (CFCA, 2015, 2013)

La literatura científica consultada define tres técnicas básicas para la detección de todos estos tipos de ataques, cada una dividida en sub-técnicas que sirven para detectar y clasificar comportamientos anómalos, estos son:

- Técnicas basadas en la Estadística, donde se define un comportamiento estocástico (comportamientos que son sometidos al azar y serán objetos de un análisis estadístico).
- Técnicas basadas en el conocimiento, donde se define una disponibilidad de los conocimientos previos a los datos. Brindando robustez, flexibilidad y escalabilidad.
- Técnicas basadas en el Aprendizaje Automático (machine learning), permite una categorización de patrones, brinda flexibilidad y adaptabilidad en la captura de interdependencias.

En nuestro país, existen varios estudios sobre la implementación de infraestructuras de telefonía IP, pero según la literatura consultada, no se tiene conocimiento de que se hayan realizado estudios para la detección de comportamientos anómalos en los sistemas de VoIP, con la "Actualización de lineamientos de la Política económica y social del Partido y la Revolución para el periodo 2016 – 2021 aprobados en el VII Congreso del PCC en abril de 2016 y aprobados por la Asamblea Nacional del Poder Popular en Julio de 2016", los lineamientos números 89 y 108, están orientados para garantizar las telecomunicaciones (7mo. Congreso del PCC 2016)

Los especialistas han acumulado experiencia, en el reconocimiento de patrones de fraude y comportamientos anómalos, en las redes de Voz (telefonía), con la utilización de la señalización SS7. Mientras que, en el caso de las redes de datos, se están dando los primeros pasos, lográndose la especialización en el monitoreo y análisis de tráfico malicioso (malwares), la detección de Botnet (son un grupo de PC infectados y controlados por un

atacante de forma remota), ataques de DoS –*Denegación de Servicio y sus variantes*–, detección de páginas web falsas –*AP Phishing*–entre otras. Sin embargo, el espectro de ataques identificados a este tipo de red es mucho más amplio y se sigue incrementando constantemente. (G. Lorenz NY, 5–6 June 2001)

Ante esta situación, a los especialistas hoy en día, se les dificulta la detección de comportamientos anómalos, pues carecen del conocimiento necesario para ello, lo que trae consigo deficiencias en la clasificación de los diferentes tipos de métodos de fraude y por consiguiente no se toman las medidas necesarias para minimizar el impacto de los mismos.

Resultados y discusión

El servicio de telefonía IP, establece un modelo de solicitud/respuesta parecido al protocolo HTTP –HiperText Transfer Protocol–; en ambos modelos la transacción consta en una solicitud de un cliente que invoca a un método en particular o función específica del servidor y este responde al cliente de acuerdo a la respuesta dada por el método o función ejecutada (Figura 2). Por lo que su funcionamiento es fluido en cualquier red, basada en protocolo IP y a la vez es sensible a las vulnerabilidades de los protocolos de VoIP, así como a las del protocolo HTTP. (Cárdenas Rojas, 2015).

Las vulnerabilidades en los sistemas de Telefonía IP, son aprovechadas cuando se rompe la tríada que define su Seguridad: Confidencialidad, Integridad y Disponibilidad, al fallar algunos de estos elementos, causan las amenazas más comunes que inciden en garantizar la Seguridad, Fiabilidad y Calidad del Servicio (Qos) de las infraestructuras de Telefonía IP. Estas se clasifican en (VOIPSA, 2005).

- Amenazas sociales: dirigidas fundamentalmente hacia los usuarios. Se debe a configuraciones erróneas, error o malas interacciones en los protocolos de los Sistemas de Telefonía IP, que permitirían a personas malintencionadas, asumir identidades maliciosas de los usuarios de la red. Ejemplos de estos tipos de ataques tenemos los Phishing - Vishing, Spam, entre otras.
- Amenazas de escucha, espionaje, intercepción y modificación conocida como *eavesdropping*: permiten a personas malintencionadas sin autorización, realizar la escucha de las señalizaciones (configuración de las llamadas) o el contenido de las sesiones de VoIP, posibilitando incluso modificar aspectos de estas sesiones, evitando su detección. Ejemplo de estos tipos de ataques puede ser el re-enrutamiento de llamadas, interceptación de sesiones RTP no cifradas, entre otras.
- Amenazas de Denegación de Servicio (Dos): son concebidos para evitar, denegar los accesos de los usuarios a los servicios de VoIP. Afectar las capacidades de comunicación de un usuario u organización, es decir, cuando todas las comunicaciones VoIP y de datos se multiplexan sobre la misma red, explotando los fallos y/o defectos en la configuración de las llamadas o en la implementación de servicios, realizar ataques genéricos a los flujos de tráfico. También pueden implicar ataques con componentes físicos (por ejemplo, desconectar físicamente o cortar un cable) o a través de ataques informáticos u otras infraestructuras (deshabilitar el servidor DNS o apagar la alimentación).
- Amenazas de abuso del servicio: consisten en un uso indebido de los servicios de telefonía IP, especialmente cuando las infraestructuras de VoIP, tienen un uso comercial, dando lugar a distintos métodos de fraude para evitar la facturación de llamadas, como por ejemplo el Toll Fraud.
- Amenazas de acceso físico: se refiere al acceso físico no autorizado o inapropiado a los equipos de VoIP, o a nivel de red (siguiendo lo que establece el Modelo ISO de 7 capas).
- Amenazas a la interrupción de los servicios: se refiere a algunos problemas que se crean de manera no intencional, pero que inciden en que los servicios de VoIP, se vuelvan inutilizables e inaccesibles. Ejemplo de esto puede ser la caída de las redes eléctricas debido a inclemencias del tiempo, el agotamiento de los recursos de los sistemas por un incremento de tráfico y/o subscriptores no previsto, problemas en el rendimiento que degradan la calidad de las llamadas.

Estas amenazas diversifican las diferentes técnicas y métodos de ataques más usuales que utilizan los defraudadores para atacar los sistemas de telefonía IP.

Las Técnicas de Lógica Difusa –***Techniques fuzzy logic***–: Emplea la teoría de los conjuntos difusos bajo la cual el razonamiento es aproximado; al contrario de la precisión que plantea la lógica clásica. Estas técnicas se utilizan en el campo de la detección de anomalías, principalmente porque las características

ISSN: 1813-5056, RNPS: 0514, Vol. 14, No. 2, julio-diciembre, 2018, pp. 73 - 84 fono

a considerar pueden considerarse como variables difusas. Este tipo de esquema de procesamiento considera una observación como normal si se encuentra dentro de un intervalo dado. (Idris Mala, 2013)

Dentro de las técnicas antes mencionadas, unas de las más utilizadas a nivel mundial, para la detección de comportamientos anómalos en los sistemas de telefonía IP, son las técnicas Fuzzing y/o de Lógica Difusa (fuzzy Logic), que pueden ser combinadas con otras técnicas para una mayor efectividad, por ejemplo la combinación con Redes Neuronales, con los algoritmos genéticos, entre otras (Narendra Shekokar, 2011).

Gracias a la aplicación de las Técnicas de Lógica Difusa, se ha llegado a encontrar gran cantidad de ataques de Denegación de Servicio (DoS), Denegación de Servicios Distribuidos (DDoS) y ataques por Desbordamientos de Búfer –*buffer overflows*–, en los productos que implementan los protocolos SIP, H.323 y sus variantes según lo aprobado por la UIT (Eric Y. Chen 2008). Además se utilizan para censar la seguridad, determinar vulnerabilidades no previstas en configuraciones o son empleadas por personas maliciosas para apropiarse de diferentes datos sobres estas infraestructuras de telefonía IP (H. Abdelnur, 2006). Para ello se profundiza en las Técnicas de Lógica Difusa, por su versatilidad en la detección de varios ataques ya conocidos

Minería de datos, conceptos y técnicas más utilizadas

La minería de datos, como el proceso de extracción de la información dentro de los grandes volúmenes de datos (Big Data), de origen matemático, a través de algunos teoremas de eficacia de algoritmos y ecuaciones lineales, cuadráticas entre otras, sumado a un tratamiento estadístico y un aprendizaje automático o semi-automático de los datos, permite demostrar la utilidad de las cosas en las diferentes áreas del conocimiento y del saber, extraer patrones interesantes hasta ahora desconocidos, como los grupos de registros de datos (análisis clúster), registros poco usuales (la detección de anomalías) y dependencias (minería por reglas de asociación), antes de la puesta de su funcionamiento. (Palomo Miñambres, 2005)

Las técnicas de la minería de datos provienen de la inteligencia artificial y de la estadística, las mismas, no son más que algoritmos, más o menos sofisticados que se aplican sobre un conjunto de datos para obtener otros.

Las técnicas más representativas y utilizadas según la literatura consultada son:

Redes neuronales. - Son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida. Algunos ejemplos de red neuronal son (Narendra Shekokar, 2011):

El perceptrón

El perceptrón multicapa

Los mapas auto organizados, también conocidos como redes de Kohonen

Regresión lineal. - Es la más utilizada para formar relaciones entre datos. Rápida y eficaz pero insuficiente en espacios multidimensionales donde puedan relacionarse más de 2 variables.

Redes bayesianas. - Llamadas redes de creencias o probabilísticas. Permite invertir las dependencias de los atributos; con una distribución previa, proporciona una forma de incorporar información externa al proceso de análisis de los datos, dando una distribución posterior, la cual vuelve a ser actualizada con la distribución previa.

BayesNaive. - Es un método heurístico basado en los teoremas de Bayes, sin las dependencias que puedan existir, en donde se emplea la teoría de la probabilidad, para encontrar lo más probable. Es utilizado para la clasificación de texto. También como clasificador para descartar redundancia e irrelevancias en los predictores (Muhammad Ali Akbar, 2012).

Árboles de decisión. - Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial y el análisis predictivo, dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema. Ejemplos:

Algoritmo ID3

Algoritmo C4.5

Lógica e Inferencia Difusa. - Se utiliza en escenarios donde no existe un modelo de solución simple o matemático preciso. Trabaja con variables lingüísticas o con datos imprecisos, con reglas de tipo IF-THEN, definidas a partir de la opinión de expertos o de un sistema de aprendizaje (red neuronal). La lógica difusa es el puente entre la alta precisión y la alta complejidad de la difusividad.

Modelos estadísticos. - Es una expresión simbólica en forma de igualdad o ecuación que se emplea en

todos los diseños experimentales y en la regresión para indicar los diferentes factores que modifican la variable de respuesta.

Agrupamiento o *Clustering*. - Es un procedimiento de agrupación de una serie de vectores según criterios habitualmente de distancia; se tratará de disponer los vectores de entrada de forma que estén más cercanos aquellos que tengan características comunes. Ejemplos: Algoritmo K-means, Algoritmo K-medoids.

Reglas de asociación. - Se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos.

Algoritmos genéticos. - Representan la modelación matemática de la biología en los cromosomas como un marco evolucionista que alcanza la estructura y composición más óptima en aras de la supervivencia. Comprendiendo la evolución como un proceso de búsqueda y optimización de la adaptación de las especies que se plasma en mutaciones y cambios en los genes o cromosomas. Se combinan posteriormente con otras técnicas como las redes neuronales, los árboles de decisión, los sistemas difusos, entre otros, después de ser utilizadas para la selección de las variables a predecir.

Se han estudiado los diferentes ataques a los que pueden ser sometidas las infraestructuras de telefonía IP, esto se refleja en determinados comportamientos anómalos en el tráfico VoIP generado, para la detección de estos comportamientos, se han realizado diferentes estudios sobre técnicas que permitan una pronta y eficaz detección.

La aplicación de las técnicas de Fuzzing en este campo ha permitido realizar diferentes estudios de muestras de tráfico de VoIP. Se pretende realizar experimentos para obtener resultados mediante un sistema Difuso (FLS) y los algoritmos de árboles de decisión, para ello se utilizarán las herramientas de minería de datos Open Source: WEKA y KEEL, combinando las librerías de lógica difusa, los algoritmos ID3 y C.4.5, de los árboles de decisión y combinándolos con las reglas de decisión, para posteriormente realizar comparaciones de los resultados obtenidos.

Lógica difusa para la detección de comportamientos anómalos en los sistemas de telefonía IP en la Empresa de Telecomunicaciones de Cuba S. A.

El monitoreo de aplicaciones distribuidas como VoIP es un desafío, porque en muchos casos el seguimiento de una sola fuente de datos no es suficiente para revelar la imagen completa de algunas actividades maliciosas. Por lo tanto, en las infraestructuras de VoIP, se deben configurar la obtención de tres tipos de fuentes de datos: el tráfico de red, los registros de detalle de llamadas y los registros del servidor. Cada uno de estos tipos de CDR –*Call Details Records*– y Logs, son importantes para un grupo de enfoques de detección. Por otra parte, la correlación de los patrones que abarcan múltiples fuentes de datos puede revelar más pruebas sobre ciertas anomalías que permitirá el monitoreo y análisis de la Confidencialidad, Integridad y Disponibilidad. A continuación, se destaca la importancia de cada tipo de datos (Ruiz-Agundez, 2014):

El tráfico de red: Especialmente los protocolos que son esenciales para el funcionamiento normal del servicio VoIP (por ejemplo, SIP, DNS), constituyen las fuentes de datos típicos para la detección de anomalías basadas en la red. Se analiza fundamentalmente para detectar una serie de inundaciones y ataques SPIT.

Los registros de detalle de llamadas (CDR) son particularmente importantes para la detección de fraudes y la mitigación de SPIT. El historial y análisis de las llamadas ayuda a crear perfiles de usuarios, ya sea desde el punto de vista del usuario y/o grupos de usuarios, para detectar llamadas fraudulentas.

El registro y las estadísticas de los servidores VoIP, son el origen de los datos típicos para la detección de intrusos basada en host. Por ejemplo, los logs que proporcionan los equipamientos que componen las infraestructuras de Telefonía IP: núcleo, memoria, estadísticas sin estado, estadísticas de transacciones, ubicación del usuario y registro. Estos parámetros pueden ser monitoreados en línea con el fin de revelar un procesamiento anormal y las cargas de memoria.

Para el estudio y la aplicación de la lógica difusa en la detección de comportamientos anómalos, se tomarán varias muestras de tráfico de VoIP, generada en las Centrales IP, de los tipos "Tráfico de Red", que nombraremos para los experimentos "tipo 1" y "Registros de Llamadas (CDR)", que nombraremos para los experimentos "tipo 2". Las muestras se componen aproximadamente por 65 000 registros, divididos en:

- La muestra del tráfico tipo 1, está compuesta por tráfico normal de datos y la otra muestra con tráfico unido combinado con normal y anómalos de datos, originados por la utilización del SIPP[1],

1 https://www.sipp.sourceforge.net, herramienta open source, para el testeo de infraestructuras VoIP.

ISSN: 1813-5056, RNPS: 0514, Vol. 14, No. 2, julio-diciembre, 2018, pp. 73 - 84 tono

herramienta de testeo utilizada en infraestructuras de telefonía IP.

- La muestra del tráfico tipo 2 está compuesta por tráfico normal de CDR y tráfico con CDR considerados totalmente anómalos, originados por la utilización del Star Trinity, en su versión no comercial, herramienta de testeo utilizada en infraestructuras de telefonía IP.

Las herramientas de testeo para el servicio de VoIP, son empleadas para detectar malas configuraciones, vulnerabilidades, hacer estudios de calidad (Qos) del servicio que comprende: calidad de la señal, calidad de los canales de voz, ancho de banda, entre otras. Para ellos existen diversas herramientas Open Source y bajo licencia que son instaladas y configuradas, en los sistemas de VoIP.

En la literatura consultada, una herramienta de Testeo de infraestructuras de VoIP, debe abarcar los siguientes puntos, para que sea efectiva en la protección y mantenimiento del servicio de telefonía IP:

Test de los protocolos que integran el servicio de VoIP.

Generación de tráfico de VoIP de prueba.

Contenga herramientas para la evaluación de las diferentes Redes de Datos.

La herramienta utilizada en el experimento, para la generación del tráfico anómalo del tipo 1, fue el SiPp, es una herramienta Open Source. Con ella se pueden configurar y generar diferentes escenarios de introducción de tráfico anómalo, en la figura 5 se muestra la estructura a nivel de red implementada.

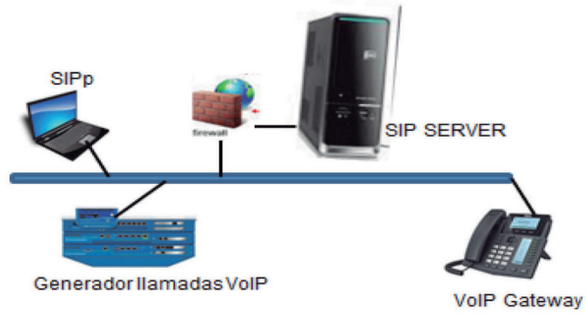

Figura 5. Estructura adoptada para la generación del tráfico anómalo utilizando el SIPp.

Para la evaluación del tráfico de tipo 1, se definieron varios parámetros a evaluar y que, en su análisis, dieron lugar a la variable de decisión definida por:

Llamada normal (NORMAL).

Llamada con parámetros anómalos (WARNING).

Llamada de fraude (FRAUDE).

Las variables principales a analizar son:

Ip Origen de la llamada.

Describir cómo valorar la IP origen cuando es 0.0.0.0.

Describir cómo valorar cuando la IP origen no cumple con la norma IPv4.

Ip destino de la llamada.

Describir cómo valorar la IP origen cuando es 0.0.0.0.

Describir cómo valorar cuando la IP origen no cumple con la norma IPv4.

Date: fecha y hora de arribo de la llamada.

d – duración de la llamada.

Cuando es 0 segundos de duración.

Cuando es 0.1 segundos hasta 3600 segundos ($0.1 \leq d \geq 3600$).

Cuando es mayor de 3600 ($d > 3600$), se considera llamada anómala.

n – estatus de la llamada:

Llamada completada, valor 1.

Llamada no completada, valor 0.

Llamada con valores distintos de 1 y 0 ($n \neq 1, n \neq 0$)

Estas variables fueron configuradas en un Sistema de Lógica Difusa (Hosseinpour Mahsa, 2016), implementado en C# (Anexo VIII), para posteriormente aplicarle las herramientas de minería de Datos WEKA y KEEL.

La herramienta utilizada en el experimento para la generación del tráfico anómalo del tipo 2, fue el Star Trinity, en su versión no comercial, ofrece un esquema de generación de llamadas que tributa a unos de los fraudes que actualmente más afecta a las compañías de telecomunicaciones, el FAS –*False Answer Supervision*–, es una herramienta comercial, con una versión libre. Con ella se pueden configurar y generar diferentes escenarios de introducción de tráfico anómalo, en la figura 6, se muestra la estructura a nivel de red implementada.

Para la evaluación del tráfico de las llamadas de VoIP, está compuesto por varios CDR (Call Details Records), campos y parámetros, de acuerdo a los componentes de la Infraestructura VoIP. En la evaluación de las llamadas se tendrán en cuenta seis parámetros que forman parte de la estructura de un CDR, que, en su análisis, darán lugar a la variable de decisión (y) definida por:

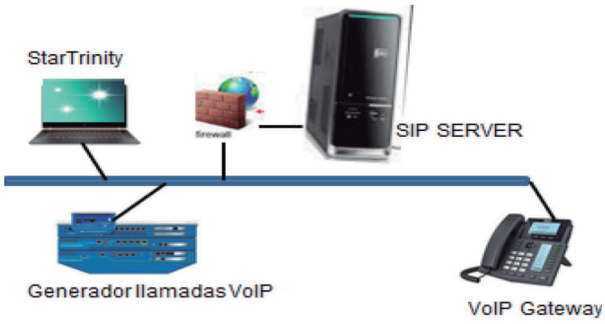

Figura 6. Estructura adoptada para la generación de tráfico anómalo utilizando el Star Trinity.

Llamada normal (NORMAL)

Llamada con parámetros anómalos (WARNING)

Llamada de fraude (FRAUDE)

Estos parámetros definidos inicialmente para el estudio de los comportamientos anómalos, se implementaron en el Sistema de Lógica Difusa –*Fuzzy Logic System* (FLS)– que fue programado en C#, son:

o – Número Llamador.

Describir cantidad de llamadas originadas de diferentes destinos.

e – Numero Llamado.

t- fecha y hora de arribo de la llamada.

d – duración de la llamada.

Cuando es 0 segundos de duración.

Cuando es 0.1 segundos hasta 3600 segundos ($0.1 \leq d \geq 3600$).

Cuando es mayor de 3600 ($d > 3600$), se considera llamada anómala.

n – estatus de la llamada:

Llamada completada, valor 1.

Llamada no completada, valor 0.

Llamada con valores distintos de 1 y 0 ($n \neq 1$, $n \neq 0$)

i – Duración del timbre (segundos).

Cuando la duración del timbre es 0 segundos ($i = 0$).

Cuando la duración del timbre es entre 1 segundo y 5 segundos.

Cuando la duración del timbre es mayor que 5 hasta 50 segundos.

Cuando es mayor que 50, se considera anómala.

Las muestras de tráfico (dataset) fueron analizadas en dos experimentos comparativos, tratando de que cumplieran con los estándares del tráfico KDDCUP'99 (Mahbod Tavallaee, 2009). A los tráficos de fraude del tipo 1 y tipo 2, después de haber pasado por el sistema difuso, se le aplicaron las herramientas de minería de Datos WEKA –*Waikato Enviroment for Knowledge Analysis*– y KEEL –*Knowledge Extraction based on Evolutionary Learning*–.

Herramienta WEKA

En el WEKA se utilizaron los algoritmos de clasificación J48 perteneciente a los algoritmos de clasificación de los árboles de decisión (Tree), el algoritmo J48, es una adaptación para la Herramienta WEKA del algoritmo C4.5, y el JRIP (JRIPPER), perteneciente a los algoritmos que se basan en la generación de reglas.

Resultados Experimentales para el Tráfico Tipo 1.

Se muestra el árbol generado con la variable ya fusificada en el tráfico tipo 1 DRT, figura 7:

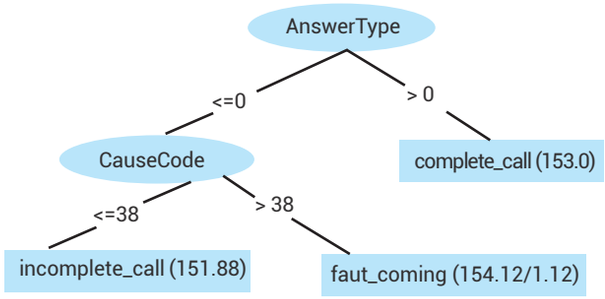

Figura 7. Árbol Decisión J48.

Y la generación de las ramas del árbol de decisión:

AnswerType <= 0

| CauseCode <= 38: incomplete_call (151.88)

| CauseCode > 38: fault_coming (154.12/1.12)

AnswerType > 0: complete_call (153.0)

Number of Leaves: 3

Size of the tree : 5

REGLAS:

IF (AnswerType <= 0) AND (CauseCode <= 38) THEN incomplete_call (151.88)

IF (AnswerType <= 0) AND (CauseCode > 38) THEN fault_coming (154.12/1.12)

IF (AnswerType > 0) THEN complete_call (153.0)

Matriz de confusión

a	b	c	<-- clasificada como
4586	0	34 |	a = incomplete_call
0	2980	0 |	b = complete_call
0	0	6 |	c = fault_coming

ISSN: 1813-5056, RNPS: 0514, Vol. 14, No. 2, julio-diciembre, 2018, pp. 73 - 84 tono

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.993	0.001	0.000	1.000	0.993	0.991	1.000	1.000	incomplete_call
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	complete_call
1.000	0.004	0.150	1.000	0.261	0.386	0.997	1.000	fault_coming

Tabla 2. Precisión detallada por clase.

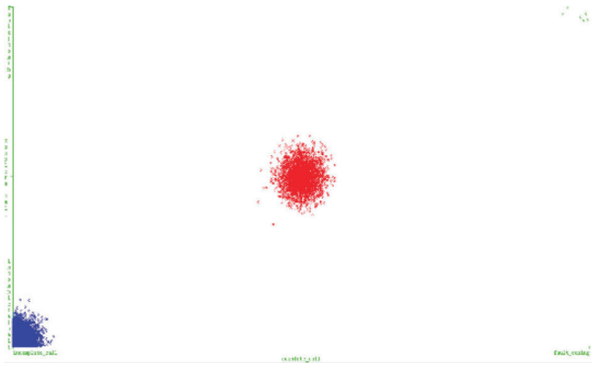

Figura 8. Gráfico generado por la implementación del JRIP.

Se muestra el gráfico generado por la implementación del algoritmo JRIP, (Figura 8):

Y la generación de las reglas del algoritmo, con la variable fusificada, para tráfico tipo 1.

(EOS >= 7351) => CallStatus=fault_coming (6.0/0.0)

(AnswerType >= 1) => CallStatus=complete_call (2980.0/0.0)

=> CallStatus=incomplete_call (4620.0/0.0)

Number of Rules : 3

REGLAS:

IF (EOS >= 7351) AND (CallStatus=fault_coming) THEN fault_coming (6.0/0.0)

IF (AnswerType >= 1) and (CallStatus=complete_call) THEN complete_call (2980.0/0.0)

IF (CallStatus=incomplete_call) THEN incomplete_call (4620.0/0.0)

Matriz de confusión

a	b	c	<-- clasificado como
4620	0	0	a = incomplete_call
0	2980	0	b = complete_call
0	0	6	c = fault_coming

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1	0	1	1	1	1	1	1	incomplete_call
1	0	1	1	1	1	1	1	complete_call
1	0	1	1	1	1	1	1	fault_coming

Tabla 3. Precisión detallada por clase.

Resultados Experimentales para el Trafico Tipo 2.

Se muestra el árbol generado por el algoritmo J48, con la variable ya fusificada en el tráfico tipo 2 DRT. (Figura 9):

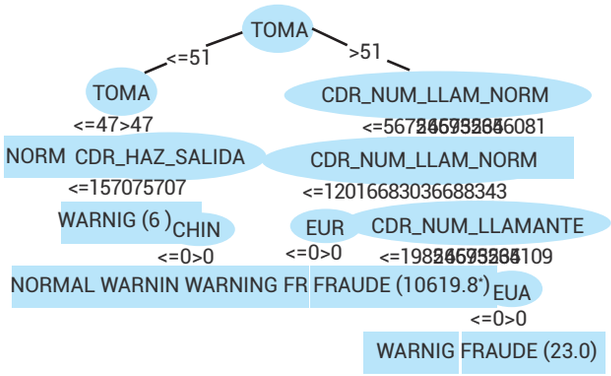

Figura 9. Árbol de decisión J48 para el tráfico tipo 2.

Y la generación de las ramas del árbol de decisión:

TOMA <= 51

| TOMA <= 47: NORMAL (5604.0/6.0)

| TOMA > 47

| | CDR_HAZ_SALIDA <= 15707: WARNING (6.0)

| | CDR_HAZ_SALIDA > 15707

| | | CHI <= 0: NORMAL (83.0/5.0)

| | | CHI > 0: WARNING (3.0/1.0)

TOMA > 51

| CDR_NUM_LLAM_NORM <= 5672236081

| | CDR_NUM_LLAMANTE <= 12016688343

| | | URA <= 0: WARNING (10.9/1.0)

| | | URA > 0: FRAUDE (18.0)

| | CDR_NUM_LLAMANTE > 12016688343

| | | CDR_NUM_LLAMANTE <= 19853174109: FRAUDE (10619.87/0.97)

| | | CDR_NUM_LLAMANTE > 19853174109

| | | | URA <= 0: WARNING (2.97)

| | | | URA > 0: FRAUDE (23.0)

| CDR_NUM_LLAM_NORM > 5672236081: WARNING

(284.26/9.1)

Number of Leaves: 10

Size of the tree : 19

REGLAS:

IF (TOMA <=51) AND (TOMA <=47) THEN NORMAL

IF (TOMA <=51) AND (TOMA > 47) AND (CDR_HAZ_SALIDA <= 15707) THEN WARNING

IF (TOMA <=51) AND (TOMA > 47) AND (CDR_HAZ_SALIDA > 15707) AND (CHI <= 0) THEN NORMAL

IF (TOMA <=51) AND (TOMA > 47) AND (CDR_HAZ_SALIDA > 15707) AND (CHI > 0) THEN WARNING

IF (TOMA > 51) AND (CDR_NUM_LLAM_NORM <= 5672236081) AND (CDR_NUM_LLAMANTE <= 12016688343) AND (URA <= 0) THEN WARNING

IF (TOMA > 51) AND (CDR_NUM_LLAM_NORM <= 5672236081) AND (CDR_NUM_LLAMANTE <= 12016688343) AND (URA > 0) THEN FRAUDE

IF (TOMA > 51) AND (CDR_NUM_LLAM_NORM <= 5672236081) AND (CDR_NUM_LLAMANTE <= 12016688343) AND (CDR_NUM_LLAMANTE > 12016688343) AND (CDR_NUM_LLAMANTE <= 19853174109) THEN FRAUDE

IF (TOMA > 51) AND (CDR_NUM_LLAM_NORM <= 5672236081) AND (CDR_NUM_LLAMANTE <= 12016688343) AND (CDR_NUM_LLAMANTE > 12016688343) AND (CDR_NUM_LLAMANTE > 19853174109) AND (URA <= 0) THEN WARNING

IF (TOMA > 51) AND (CDR_NUM_LLAM_NORM <= 5672236081) AND (CDR_NUM_LLAMANTE <= 12016688343) AND (CDR_NUM_LLAMANTE > 12016688343) AND (CDR_NUM_LLAMANTE > 19853174109) AND (URA > 0) THEN FRAUDE

IF (TOMA > 51) AND (CDR_NUM_LLAM_NORM > 5672236081) THEN WARNING

Matriz de confusión

a	b	c	<-- clasificado como
5676	8	0	a = NORMAL
11	296	1	b = WARNING
0	3	10660	c = FRAUDE

Se muestra la gráfica generada por la implementación del algoritmo JRIP, (Figura 10):

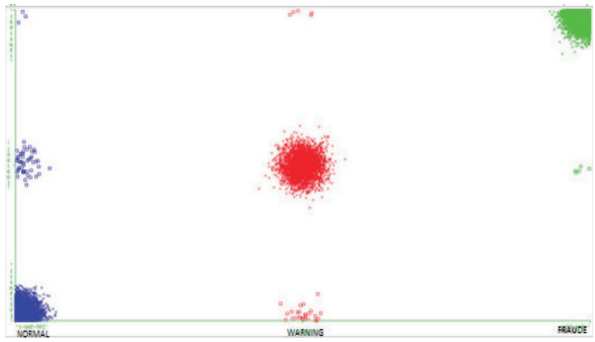

Figura 10. Gráfico generado por el algoritmo JRIP.

Y la generación de las reglas, con la variable fusificada.

(DRT = '(30-60]') and (CDR_NUM_LLAM_NORM >= 5989376065) and (TOMA >= 52) => class=WARNING (275.0/6.0)

(EUA <= 0) and (TOMA >= 51) => class=WARNING (24.0/4.0)

(CDR_HAZ_SALIDA <= 15705) and (DRT = '(30-60]') and (TOMA <= 51) and (TOMA >= 48) => class=WARNING (6.0/0.0)

(CDR_NUM_LLAM_NORM >= 13057670190) and (TOMA >= 50) and (I31 <= 0) => class=WARNING (6.0/2.0)

(VCLA >= 0.993655) and (SEGS_DUR >= 97) => class=WARNING (2.0/0.0)

(TOMA <= 51) => class=NORMAL (5678.0/6.0)

=> class=FRAUDE (10664.0/1.0)

Número de reglas: 7

REGLAS:

IF (DRT = '(30-60]') and (CDR_NUM_LLAM_NORM >= 5989376065) and (TOMA >= 52) then class=WARNING

ELSE

IF (URA <= 0) and (TOMA >= 51) then class=WARNING

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.999	0.001	0.998	0.999	0.998	1	NORMAL
0.961	0.001	0.964	0.961	0.963	0.996	WARNING
1	0	1	1	1	1	FRAUDE

Tabla 4. Precisión detallada por clase.

ISSN: 1813-5056, RNPS: 0514, Vol. 14, No. 2, julio-diciembre, 2018, pp. 73 - 84 tono

ELSE
IF (CDR_HAZ_SALIDA <= 15705) and (DRT = '(30-60]') and (TOMA <= 51) and (TOMA >= 48) then class=WARNING
ELSE
IF (CDR_NUM_LLAM_NORM >= 13057670190) and (TOMA >= 50) and (I31 <= 0) then class=WARNING
ELSE
IF (VCLA >= 0.993655) and (SEGS_DUR >= 97) then class=WARNING
ELSE
IF (TOMA <= 51) then class=NORMAL
ELSE
class=FRAUDE

Matriz de confusión

a	b	c	<-- classified as
1879	4	0	a = NORMAL
7	106	3	b = WARNING
0	0	3664	c = FRAUDE

Herramienta KEEL

En la experimentación con el KEEL, se utilizará el algoritmo FURIA –*Fuzzy Unordered Rule Induction Algoritm*–, es un algoritmo basado en los algoritmos de lógica difusa y el JRIP. (Hühn 2009)

Resultados experimentales obtenidos con el tráfico tipo 1

En la figura 11 se observarán los algoritmos utilizados para el análisis de la muestra (dataset) de tráfico del tipo 1, ya fusificada con diversas variables de entrada (crisp value) y una salida, en la importación del dataset se continuó con el formato ARFF, del WEKA, la combinación del FURIA con el algoritmo ID3-D, perteneciente a los árboles de decisión, arrojó los siguientes resultados factibles:

Reglas propuestas por el algoritmo FURIA:

(AbnormalId = call_normal) => tipoLlamada=normal (CF = 1.0)

(AbnormalId = abnormal_call) => tipoLlamada=-fraude (CF = 1.0)

Número de Reglas: 2

Describiendo el experimento con el KEEL la herramienta fue guiando la combinación de los algoritmos, para analizar la muestra con dos módulos del Algoritmo FURIA y de mediación el algoritmo ID3, al final se carga el módulo de visualización general. Según la literatura consultada esto nos ofrece una panorámica de lo realizado y obtenido en los experimentos.(Project, 2015)

Resultados Experimentales para el tráfico tipo 2:

En la figura 12 se observarán los algoritmos utilizados para el análisis de la muestra (dataset) de tráfico del tipo 2, no fusificada, en la importación del dataset se continuó con el formato ARFF, del WEKA y la combinación del FURIA con el algoritmo MinMax-Tr, perteneciente a los algoritmos de transformación. Es importante señalar que la muestra de tráfico (dataset) no fue fusificada por el FLS realizado y arrojó los siguientes resultados factibles:

Reglas propuestas por el algoritmo FURIA:

(DRT <= 46(-> 47)) and (SEGS_DUR <= 6(-> 7)) => class=NORMAL (CF = 1.0)

(DRT <= 44(-> 47)) and (DRT >= 1(-> 0)) => class=NORMAL (CF = 1.0)

(CDR_NUM_LLAM_NORM >= 5989376065(-> 5378832339)) and (DRT >= 47(-> 46)) => class=WARNING (CF = 0.99)

(CDR_NUM_LLAM_NORM <= 4549360436(-> 5352557091)) and (DRT >= 49(-> 46)) and (CDR_NUM_LLAMADO_NORM >= 5352557091(-> 5352552397)) => class=WARNING (CF = 0.8)

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.998	0.002	0.996	0.998	0.997	0.999	NORMAL
0.914	0.001	0.964	0.914	0.938	0.972	WARNING
1	0.002	0.999	1	1	0.999	FRAUDE

Tabla 5. Precisión detallada por clase.

Figura 11. Aplicación de los algoritmos con el KEEL.

Figura 12. Aplicación de los Algoritmos con el KEEL.

(CDR_AREA_DEST = VCLA) and (CDR_ZONA_ORIG = EUPA) => class=WARNING (CF = 0.51)

(DRT >= 55(-> 54)) and (CDR_NUM_LLAM_NORM <= 5378832339(-> 13057670190)) => class=-FRAUDE (CF = 1.0)

(TOMA >= 59(-> 58)) and (CDR_NUM_LLAM_NORM <= 5378832337(-> 12397388879)) => class=-FRAUDE (CF = 1.0)

Número de Reglas: 7

Al describir el experimento realizado con la herramienta KEEL, se mostró la combinación de los algoritmos, para analizar la muestra no fusificada con un módulo del Algoritmo FURIA y un módulo del algoritmo MinMax-Tr.

Conclusiones

La experimentación y la implementación son acciones que van cogidas de la mano. Facilitarían una rapidez y eficacia en la detección de comportamientos anómalos ya conocidos, que generalmente derivan en fraude, pérdidas de dinero, mala imagen de los operadores de Telecomunicaciones antes sus clientes.

Los resultados obtenidos nos muestran la pertinencia de la combinación de la Lógica Difusa con otros algoritmos de clasificación, en este caso, los algoritmos ID3 y J48, pertenecientes a los algoritmos de los Arboles de Decisión, el JRIP y MD5, y estos a su vez a los algoritmos de Reglas de Decisión y los algoritmos que implementan sistemas difusos como el FURIA.

Los experimentos realizados abren nuevos caminos para detectar nuevas formas de analizar lo que aún no es visible bajo el Iceberg. Las herramientas de minería de datos como el WEKA y el KEEL, brindan la posibilidad de encontrar nuevos caminos para implementar el "Data Mining".

Referencias

7mo. Congreso del PCC , A. N. d. P. P. A. (2016). "Actualización de los Lineamientos para el periodo 2016-2021."

Amparo, P. (2011). "Planificación de Seguridad en Voip."

Auza, J. (25 de noviembre 2010). "5 of the Best Free and Open Source Data Mining Software."

Cárdenas Rojas, M. (junio 2015). "Análisis y Diagnóstico al Sistema de Gestión de Seguridad de la Información de la Red Voip."

Cristóbal Romero, S. V., Carlos de Castro, Enrique GarcíaCristóbal Romero, Sebastián Ventura, Carlos de Castro, Enrique García (2005). "Algoritmos Evolutivos para Descubrimiento de Reglas de Predicción en la Mejora de Sistemas Educativos Adaptativos basados en Web " Revista Iberoamericana de Informática Educativa Revista Iberoamericana de Informática Educativa- IE Comunicaciones.

D' Negri Carlos Eduardo, D. V. E. L. (2006). "Introduccion al razonamiento aproximado: Logica Difusa." Revista Argentina de Medicina Respiratoria.

Eric Y. Chen, M. I. (2008). "Scalable Detection of SIP Fuzzing Attacks." IEEE.

G. Lorenz, T. M., G. Manes, J. Hale, S. Shenoi (NY, 5–6 June 2001). "Securing SS7 Telecommunications Networks." **2001 IEEE**.

García González , Francisco J. (2013). "Aplicación de técnicas de Minería de Datos a datos obtenidos por el Centro Andaluz de Medio Ambiente (CEAMA)." Universidad de Granada.

Garcıía-Teodoro, P., et al. (2008). "Anomaly-based network intrusion detection: Techniques, systems and challenges." Computers & security **28**(2009): 18 - 28.

Gruber Markus , C. S., Florian Fankhauser, Martin Moutran, Thomas Grechenig (2013). "Architecture for Trapping Toll Fraud Attacks Using a VoIP Honeynet Approach." Springer-Verlag Berlin Heidelberg.

Guillermo, M.-L. (17 de febrero de 2002). "Introducción a la lógica difusa." Centro de Investigación y Estudios Avanzados del IPN.

H. Abdelnur, V. C., R. State and O. Festor (2006). "VoIP Security Assessment: Methods and Tools." IEEE: 6.

Hosseinpour Mahsa , M. H. Y. M., Seyed Amin Hosseini Seno, Hossein Khosravi Roshkhari (october, 2016). "Modeling SIP Normal Traffic to Detect and Prevent SIP-VoIP Flooding Attacks Using Fuzzy Logic." IEEE.

Housam Al-Allouni, A. E. R., Mohammed Hashem, Ali El-moghazy, Abd El-Aziz Ahmed (March17-19,2009). "VoIP Denial of Service Attacks Classification and Implementation." 26 th NATIONAL RADIO SCIENCE CONFERENCE (NRSC2009).

Hühn, J., Eyke Hüllermeier (2009). "FURIA: an algorithm for unordered fuzzy rule induction." Springer Briefs in Computer Science: 27.

Humberto Abdelnur, O. F., Radu State (July, 2007). " KiF: A stateful SIP Fuzzer." IPTCOMM '07, New York USA.

Idris Mala, P. A., Tariq Javid Ali, Syed Saood Zia (2013). "Fuzzy Rule Based Classification for Heart Dataset using Fuzzy Decision Tree Algorithm based on Fuzzy RDBMS." World Applied Sciences Journal.

Jared Ring, K.-K. R. C., Ernest Foo, Mark Looi (march 20, 2006). "A New Authentication Mechanism and Key Agreement Protocol for SIP Using Identity-based Cryptography." **Information Security Institute Queensland University of Technology**.

Jason Ostrom, J. K. (2007). "VoIP Hopping: A Method of Testing VoIP security or Voice VLANs." SecurityFocus.com.

Jesús Alcala-Fdez, S. G., Francisco Jose Berlanga, Alberto Fernández, Luciano Sánchez, M.J. del Jesus and Francisco Herrera (march 2008). "KEEL: A data mining software tool integrating genetic fuzzy systems." IEEE.

Krasheninnikova, E. (2013). "SEGURIDAD EN VOIP : APLICACIÓN DE SEÑUELOS " Universidad Politécnica de Madrid.

Kuna, H. M., Mirta J.; Caballero, Sergio; Jaroszczuk, Susana. (2009). "Procedimientos de la explotación de información aplicados al ámbito bibliotecológico." 7ª Jornada sobre la Biblioteca Digital Universitaria, JBDU2009.

Mahbod Tavallaee, E. B., Wei Lu, and Ali A. Ghorbani (2009). "A Detailed Analysis of the KDD CUP 99 Data Set." IEEE.

Mandeep Kaur, S. K., Swati Sharma (2013). "DoS Flooding Attacks against SIP based VoIP Systems- a Survey." International Journal of Computer Science and Engineering **Volume-2**(Issue-5).

MENDEL, J. M. (March, 1995). "FUZZY LOGIC SYSTEMS FOR ENGINEERING: A TUTORIAL." IEEE **83**.

Mohamed Nassar, R. S., Olivier Festor (April, 2009). "VoIP Networks Monitoring and Intrusion Detection." Universit´e Henri Poincar´.

Moyano Jose, S. L. (2016). "RKEEL: Using KEEL in R Code." 2016 IEEE International Conference on Fuzzy Systems (FUZZ).

Muhammad Ali Akbar , M. F. (2012). "Securing SIP-based VoIP infrastructure against flooding attacks and Spam Over IP Telephony." Springer-Verlag Berlin Heidelberg.

Narendra Shekokar, S. D. (2011). "Anomaly Detection in VoIP System Using Neural Network and Fuzzy Logic " Springer-Verlag Berlin Heidelberg.

Palomo Miñambres, Ó. (2005). "Minería de Datos." Universidad Carlos III de Madrid.