ISSN 1870-4069

Generación de un vector característico para la detección de intrusos en redes computacionales

Ana Alcántara-Ramírez, Lourdes López-García, Juvenal Rueda Paz

Universidad Autónoma del Estado de México,
Centro Universitario UAEM Valle de Chalco, Estado de México,
México

aalcantarar@alumno.uaemex.mx,mllopezg@uaemex.mx, jruedap@uaemex.mx

Resumen. El control de acceso no autorizado en redes informáticas es un problema que inicia desde el surgimiento de los sistemas de información computarizados, donde la seguridad y la privacidad de la información son factores importantes. Una solución conveniente para resolver este problema es el uso de un Sistema de Detección de Intrusos (IDS, por sus siglás en inglés). La eficiencia de un IDS está determinada por la certeza en la detección, misma que depende de una correcta clasificación, que tendrá lugar si se cuenta con un vector que contenga las características adecuadas del objeto o entidad a clasificar. En este artículo, se propone la generación de un vector característico a partir de información real proveniente de la red que permita realizar una correcta interpretación sobre el comportamiento de los procesos habituales para los cuales la red fue creada, para así, discernir entre lo autorizado y no permitido en la red. Para comprobar la eficacia de la detección se utilizan 5 clasificadores incluido en ellos una red neuronal y árbol de decisión. Así, la certeza de una evaluación precisa de la red, permitirá protegerla de usuarios maliciosos que intenten invadirla sin ser detectados.

Palabras clave: ataques de intrusión, algoritmos de clasificación, vector característico.

Development of a Characteristic Vector for the Intrusion Detection in Computational Networks

Abstract. Unauthorized access detection in a computer networks is a problem that starts from the beginning of computerized information systems, where the security and privacy of the information are important factors. A good solution to solve this problem is the use of an Intrusion Detection System (IDS). The efficiency of an IDS is measured by the precision in the detection which depends on the an accurate classification, that can be possible, using a vector with the appropriate characteristics of the object or entity to be classified. In this paper, the generation of a

characteristic vector based on real information coming from the network, is proposed. The vector allows classifiers to do a correct interpretation of the behavior of the common processes for which the network was created, in order to discern between what is authorized or non authorized on the network. To verify the effectiveness of the detection, five classifiers are used, including a neural network and decision tree. Thus, the certainty of an accurate evaluation of the network, will protect it from malicious users who try to invade it undetected.

Keywords: intrusion attacks, classification algorithms, characteristic vector.

1. Introducción

La importancia de la comunicación radica en la necesidad de compartir información entre entidades. El canal de transmisión de los datos puede ser público o privado, sin embargo, en ambos casos debe proveerse eficiencia y seguridad ante información importante o secreta. Las redes computacionales son un medio de comunicación que permite compartir información a grandes distancias de manera rápida, fácil y en diferentes formatos.

La mayoría de los usuarios puede acceder a las redes de computadoras, que les permite tener una comunicación desde su ubicación hacia cualquier punto donde la red mantenga conexión. Utilizar este canal, sobre todo si es público, implica tener conocimiento sobre cómo usuarlo, pero no necesariamente sobre cómo funciona. Ésta diferencia, hace que las entidades maliciosas se aprovechen de los usuarios ingenuos para vulnerar el canal de comunicación usado y lograr ataques como una intrusión no deseada en la red.

Para proteger los datos de quienes usan estos canales, se implementan protocolos de seguridad, así como, la aplicación de métodos y herramientas especializadas para ciertas tareas. Los Sistemas de Detección de Intrusiones son útiles en la búsqueda de la seguridad, brindando un medio de detección ante una intrusión no autorizada en una red. En esta herramienta se incluyen métodos para el tratamiento y análisis de los datos que se transmiten, tales como la minería de datos, los algoritmos genéticos, la inteligencia artificial, entre otros. De tal manera, que sea posible clasificar el tipo de tráfico y decidir cuándo se presenta o no, un ataque de intrusión [1].

Contar con procedimientos eficientes de clasificación de patrones es esencial en muchas aplicaciones de gran interés. Una de ellas es el diseño de Sistemas de Detección de Intrusos en sistemas de cómputo. Un factor fundamental para alcanzar la eficiencia en procesos de clasificación, es la ejecución previa de técnicas de selección y la extracción de características sobre el conjunto de datos. Lo cual no sólo mejora la precisión de la clasificación, también mejora la capacidad de generalización en el caso de la clasificación supervisada, o contrarresta el riesgo de una mala clasificación que puede presentarse al usar clasificadores no supervisados.

Para desarrollar un sistema clasificador es necesario determinar el conjunto de características que describan la arquitectura de los datos con que se trabaja. La razón de identificar la elección de un subconjunto adecuado de características, es que permite la reducción de la dimensionalidad en el conjunto de datos, lo que contribuye a disminuir la complejidad computacional de la clasificación, mejorando el rendimiento del clasificador y evitando características redundantes o irrelevantes. Aunque la selección de características se puede definir formalmente como un problema de optimización con un solo objetivo, (es decir, la precisión de la clasificación obtenida usando el subconjunto de características seleccionadas), en los últimos años, se han propuesto algunos enfoques multi-objetivo para este problema.

En este trabajo se propone la generación de un vector característico como base fundamental para ser utilizado en un IDS. Dicho vector está conformado por datos reales obtenidos de una red. El escenario propuesto consta de la puesta en marcha de una red, diseñada para analizar su comportamiento en estado normal y bajo ataque por inundación de paquetes. De tal manera que sea posible generar un vector característico con información proveniente de una comunicación cotidiana en la red o en otro caso, bajo un ataque que tiene como objetivo introducir paquetes a la red para saturarla. El resultado es un vector con características distinguibles para cada escenario, lo que implica una clasificación correcta en la toma de decisión sobre si es un ataque o no. Así, la principal contribución de este trabajo es un enfoque empleado para la selección de características y su aplicación a un enfoque supervisado.

Para garantizar la efectividad de nuestra propuesta, los vectores caracteríticos generados son clasificados en una red neuronal y en los algoritmos J48, Random-Forest, Naive Bayes y Decision table, obteniendo una matríz de confusión con un porcentaje mínimo de falsos positivos.

El resto del artículo se organiza de la siguiente forma. En la sección 2 se presentan los conceptos básicos y definiciones necesarias respecto a los IDS y los algoritmos de clasificación. La sección 3 presenta el estado del arte en los trabajos relaciones a la propuesta en este artículo. Posteriormente, en la sección 4 se describe el escenario propuesto y se presenta un análisis de su comportamiento bajo los dos escenarios mencionados. En la sección 5 se detalla la forma en cómo es generado el vector característico y se muestra la efectividad de su clasificación, en la sección 6. En la sección 7 se realiza una discusión de los resultados obtenidos y una comparativa con los trabajos relacionados. Finalmente, en la sección 8, se presentan las conclusiones de este trabajo.

2. Preliminares

Esta sección comprende los conceptos básicos que intervienen y están relacionados con el desarrollo del objetivo principal, que es la creación de un vector característico. Iniciamos con una descripción de los IDS para comprender mejor la aplicación que se buaca alcanzar y la importancia que tiene lograrlo.

Posteriormente, se presenta una breve explicación sobre las características de los algoritmos para el análisis de los datos que componen el vector característico.

2.1. Sistemas de detección de intrusiones (IDS)

Los IDS son herramientas que escuchan el tráfico de la red y son capaces de detectar actividades inusuales, para así, reducir el riesgo de una intrusión no permitida. Los IDS pueden evaluar la información en tiempo real, o que esté contenida en una base de datos. Existen varias clasificaciones para los IDS, el denominado HIDS es un sistema de detección de intrusiones basado en host, que tiene como objetivo identificar ataques con base en la observación de los encabezados de los paquetes, para detectar a una entidad que intenta violar o modificar la seguridad del host. Por otra parte, el NIDS que es un sistema de identificación de intrusiones de red y su detección que se basa en el análisis de los paquetes de red y de los protocolos que se emplean para la transmisión de los datos, ambos pueden ser en tiempo real o no [2].

La Tabla 1 muestra los enfoques que puede tener un IDS, de acuerdo al modo de detección que maneja: comportamiento, firmas, anomalias o heurístico [3,4,5,6].

2.2. Algoritmos de clasificación

El proceso de clasificación es uno de los más útiles y comunes en el tratamiento de datos, ya que permite analizar el comportamiento de una o más variables dentro de un conjunto de información. Dicho conjunto es formado por datos agrupados y dependientes del atributo al que pertenecen, los datos son sometidos al sistema clasificador para así, determinar a qué clase corresponden. Los clasificadores requieren una fase de entrenamiento o construcción de la base de conocimientos [7,8]. En este trabajo, se utilizaron cinco clasificadores que se describen a continuación:

- **Red neuronal**: Se compone de varias neuronas (unidad mínima de procesamiento de la información, representa un dato de entrada) que están divididas en varias capas. Las neuronas de una capa se conectan con las neuronas de la capa siguiente y les pasan información. La arquitectura consiste en una capa de entrada que recibe la información del exterior; capas intermedias (ocultas) que realizan el trabajo de la red y una capa de resultados que muestra los resultados de la última capa intermedia [8,9].
- **Algoritmo J48**: Se deriva del algoritmo C4.5 y para la clasificación crea un árbol binario [10]. Se basa en la utilización del criterio *ratio de ganancia* (gain ratio) para evitar que las variables con mayor número de presencia salgan beneficiadas en la selección. Además, el algoritmo incorpora una poda del árbol una vez que éste ha sido inducido [11].
- **Random Forest**: Emplea una selección aleatoria de atributos y genera un conjunto de árboles predictores que serán evaluados posteriormente [12].

Tabla 1. Clasificación de los IDS.

Enfoque	Descripción
Comportamiento	*Funcionalidad*: busca variaciones de costumbres, como un tráfico elevado. *Ventajas*: método simple y efectivo para detectar ataques conocidos. Detalla el análisis contextual. *Desventajas*: inefectivo para ataques no conocidos, o variantes de los conocidos. Difícil mantener las firmas y patrones actualizados. Requiere mucho tiempo para aprender.
Firmas o MD-IDS	*Funcionalidad*: clasifica ataques con base en firmas y auditorias. *Ventajas*: efectivo para detectar vulnerabilidades nuevas, es menos dependiente del sistema operativo y puede detectar el abuso de privilegios. *Desventajas*: los perfiles cambian constantemente y no es efectivo en su reconstrucción.
Anomalías o AD-IDS	*Funcionalidad*: busca elementos fuera de lo común, se centra en patrones de tráfico. *Ventajas*: está basado en comportamiento de protocolos de red, detecta secuencias anormales de comandos. *Desventajas*: no distingue ataques que simulen un comportamiento usual en los protocolos y puede ser incompatible con algunos navegadores.
Heurístico	*Funcionalidad*: emplea algoritmos para analizar el tráfico que pasa por la red. *Ventajas*: puede predecir eventos y ser autodidacta, distingue secuencias de comando. *Desventajas*: consume muchos recursos y es de funcionamiento complejo.

Cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos. Es una modificación sustancial de bagging que construye una amplia colección de árboles no correlacionados y promediados posteriormente [13].

- **Naive Bayes**: Asume que la presencia o ausencia de una característica particular no está relacionada con la presencia o ausencia de cualquier otra característica, variable, tabulador, parámetro o atributo. Se evalúan de modo independiente sin establecer relaciones o coincidencias. Se puede entrenar en un ambiente de aprendizaje supervisado con pocos datos, obteniendo las medias y las varianzas de las variables necesarias para la clasificación. Debido a que las variables independientes se asumen, solo es necesario determinar las varianzas de las variables de cada clase y no toda la matriz de covarianza [14,15].
- **Decision Table**: Llamada DTM (Decision Table Majority), se compone de un conjunto de características que se incluye en la tabla atributos y por instancias etiquetadas (reglas). En su procesamiento cada dato de entrada

se asigna a la clase con la que ha tenido mayor número de correspondencias. De esta forma, a partir de un dato no etiquetado el clasificador busca correspondencias de este dato de entrada con el total de reglas para todos los atributos. Si no se encuentra alguna correspondencia, la tabla DTM asigna el dato a la clase mayoritaria [16,17].

3. Estado del arte

La información generada de procesos de red es cuantiosa, y tiende a crecer en cuanto la arquitectura de la red y los servicios que proporcionan se incrementan. Garantizar la seguridad de esta información obliga a buscar mejores herramientas. La base de un correcto funcionamiento de estas herramientas y su efectividad depende de lo certero que sea el juicio generado, es decir, la capacidad de distinguir correctamente el flujo que circula para así discernir entre lo permitido y lo no permitido.

Dentro de la literatura que contiene los avances en este tipo de trabajos podemos encontrar que siguen distintas vertientes, algunos apuestan por la variación de clases para evaluar correctamente, otros autores mencionan la necesidad de acotar las variables a evaluar y ser más mesurados en la cantidad de clases con las que se trabaja. Otro aspecto que se considera es el enfoque de aprendizaje supervisado, no supervisado, semi-supervisado, entre otros; que garanticen una mejor evaluación de las relaciones entre los datos.

Siguiendo con estos puntos, acciones tales como considerar un pre-procesado en los datos, proponer algoritmos que combinen algoritmos ya existentes, trabajar con bases de conocimientos previamente generadas o proponer el manejo de datos reales, se enfocan a conseguir la muestra apropiada y el evaluador o clasificador preciso que brinde confianza en su predicción.

En [18] analizan el conjunto de datos denominado KDD99 que cuenta con 41 atributos distintos, de los cuales, se seleccionaron 23 para su clasificación. El entrenamiento se realizó con el 10 % de los 51 millones de instancias contenidas en la base de datos, aplicándoles tres variantes de preprocesamiento, para después hacer una comparación basada en el uso de algoritmos representativos del aprendizaje automático. Entre éstos algoritmos se encuentran, una Red Neuronal Perceptron Multicapa (MLP), SMO que es una variante empleada en WEKA del algoritmo de Máquinas de Soporte Vectorial (SVM), el algoritmo J48, Naive Bayes y el algoritmo basado en instancias K con valores 3, 5 y 7.

Los resultados presentados arrojaron porcentaje del 98.14 % para Naive Bayes y un 99.02 % para J48, siendo éste el más preciso. Para la variante de preprocesado 2 se tiene a J48 con un 97.43 % ante lo obtenido con SMO con un 99.23 %. Finalmente, los resultados con la variante 3 presenta al algoritmo J48 con 95.85 % y MLP con 98.4 %.

El [19] se propone un sistema de identificador de intrusiones que use un clasificador basado en aprendizaje semi-supervisado. Los algoritmos empleados para el tratamiento de los datos son J48, Naive Bayes, NB tree, Random Forest, Random tree, Red Neuronal y SVM. Se implementan dos variantes que consisten

en el uso de la base de conocimiento KDDCUP99 con los 41 atributos que la integran y una variación de esta base compuesta por 21 atributos. Los resultados de certeza en la clasificación son SVM con 69.52 % que presenta los valores más bajos con un 42.29 % para la segunda variante de la base de conocimiento que cuenta con 21 atributos.

En [20] se encuentran tres variantes propuestas con distinto enfoque para evaluar los datos, previo al proceso del clasificador. Los datos son obtenidos de las bases KDD99 y Gure KDD. Tiene 6 posibles clases que representan 5 ataques a la red y una clase que describe un comportamiento normal. Dentro de los ataques que se incluyen en el evaluador, se encuentra el de Denegación de Servicio, además, de un algoritmo para clasificar las clases.

Los resultados porcentuales de certeza, se presentan en una tabla que los divide en los tres enfoques de clasificación trabajados. Para el método de clasificación 1 aplicada a 8 algoritmos se obtuvieron los siguentes resultados, 80.67 % para Random Forest y 99.21 % para una variante de su algoritmo propuesto. El segundo método de clasificación se aplicó a tres algoritmos donde su propuesta obtuvo 82.10 % frente al 96.5 % de Naive Bayes. Por último, la tercera clasificación, aplicada a 4 algoritmos, presentaron los valores 98.38 % para Decision Tree Based y 99.27 % para el algoritmo de su propuesta, con lo cual, garantizan una clasificación precisa.

El trabajo realizado en [21] recomienda el uso de Mobile Ad hoc redes (MANET) para la protección de redes inalámbricas. Emplea un modelo probabilístico que tiene la finalidad de reducir los tiempos activos del IDS, centrandose en la conexión realizada entre dispositivos al inicializar un juego. El juego es cooperativo y multijugador que analiza los efectos de un IDS con una actividad reducida en la red. El sistema funciona en redes estáticas y móviles. El algoritmo empleado es LDK para la detección de los vecinos cercanos o jugadores introducidos en el juego. El enfoque principal, entonces, no es diseñar un IDS, más bien es presentar un esquema para un uso eficiente que determina el ahorro de energía de los dispositivos mientras el IDS se ejecuta.

4. Escenario propuesto

La Fig. 1 muestra el escenario definido para el análisis de tráfico. Como puede observarse, consiste de dos redes conectadas a través de un ruteador. La red 2 se compone de un servidor web, el cual será atacado, y varios host, mientras que la red 1 contiene por lo menos una computadora que fungirá como adversario.

El ataque aplicado a la red definida es el de denegación de servicio (DoS), el cual busca la interrupción del flujo de datos y reduce la disponibilidad que otorga un servicio activo. El modo en que opera consiste en enviar paquetes con formato permitido en grandes cantidades para lograr la saturación del servidor, de tal manera que ya no le sea posible atender las solicitudes. Para lograr la saturación del servidor web se transmitieron paquetes del protocolo ICMP con carga elevada. Es importante mencionar que, el interés en este artículo es detallar una solución al ataque de DoS por inundación y no describir cómo se efectúa.

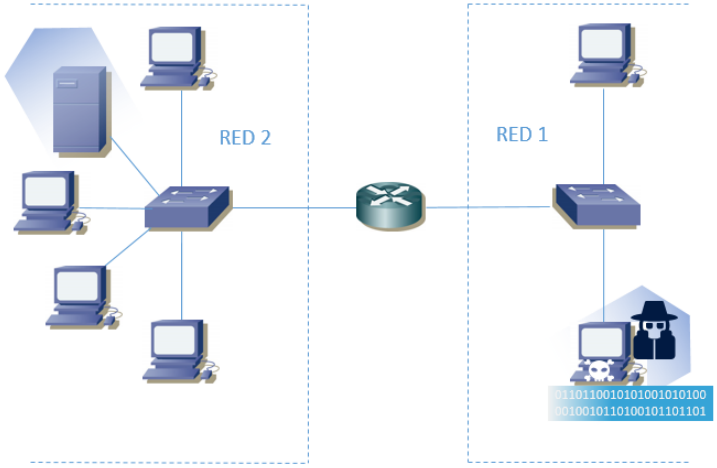

Fig. 1. Topología de la red.

La información resultante del ataque es obtenida a través de Whireshark, que es una aplicación para la escucha de la red, que nos permite guardar en un archivo de formato de texto, el cual será utilizado para obtener el vector característico.

5. Creación del vector característico

Los IDS protegen a un conjunto de computadoras de usuarios no autorizados, incluyendo, posiblemente, a entidades internas. Durante el periodo de entrenamiento, el IDS construye un modelo predictivo (un clasificador) capaz de distinguir entre las conexiones normales y las conexiones anormales, denominadas intrusiones o ataques. Como se mostró en la Fig. 1, el escenario establecido contempla dos redes que mantienen comunicación a través de un ruteador. El paso de información entre ambas redes, es generada para evaluar el tráfico en los estados *normal* y de *ataque*.

Los datos transmitidos en la conexión son generados aleatoriamente. Llamamos conexión a una secuencia de paquetes que fluyen desde una dirección IP en la red 1 hacia una dirección IP de la red 2, bajo algún protocolo bien definido como TCP, ICMP, entre otros. La captura del tráfico de red se tomo con un periodo de 5 segundos para cada lectura realizada. Se obtuvieron 45 lecturas de la red en estado normal y 45 en estado de ataque.

Ya que es un ambiente controlado, cada conexión es etiquetada ya sea como *normal* o como *ataque* y los registros resultantes tienen un tamaño entre 7.5KB y 2.5MB. Como la finalidad de este trabajo es distinguir cuándo es un ataque o no, las lecturas son divididas en dos clases: a0 y a1, respectivamente.

Como es de esperarse, las lecturas indican información proveniente de la red: IP origen, IP destino, tipo de puerto, descripción de la tarea como solicitud, acuse, transmisión del paquete, etc., lo que comúnmente se llama el volcado TCP sin procesar para una red de área local (LAN). Esta información no puede ser ingresada a los clasificadores, tal cual se encuentra en el registro, por lo que es necesario procesarla y obtener las características del archivo resultante de la lectura.

En la tabla 2 se muestran los atributos que conforman el vector característico, que están divididos en la información de la red como las IP de origen y la de destino; el tiempo de lectura (5 segs) y la información contenida en el archivo de registro como el número de patrones totales, número de patrones distintos, densidad léxica, etc. La tabla indica el tipo de información, si sus valores son constantes o variables.

El resultado es un vector característico de 17 elementos, que son normalizados y clasificados como $a0$ (estado normal) y $a1$ (estado de ataque).

Tabla 2. Lista de atributos.

Atributo	Descripción	Valor
a0	Ip origen	constante
a1	Ip origen	constante
a2	Ip origen	constante
a3	Ip origen	constante
a4	Ip destino	constante
a5	Ip destino	constante
a6	Ip destino	constante
a7	Ip destino	constante
a8	Tiempo de transmisión	constante
a9	Patrones totales	variable
a10	Número de patrones distintos	variable
a11	Densidad léxica	variable
a12	Total de sentencias	variable
a13	Longitud promedio comando	variable
a14	Longitud máxima de comando	constante
a15	Longitud mínima de comando	constante
a16	Legibilidad1	variable
a17	Legibilidad2	variable

6. Resultados de la clasificación

Para garantizar la efectividad del vector característico propuesto, se utilizaron los clasificadores descritos en la sección 2.2. De la información resultante de los clasificadores, tomamos al porcentaje de clasificaciones correctas, el tiempo

de ejecución y la matriz de confusión como ponderadores de los vectores caracteristicos puestos a prueba.

La Fig. 2 muestra la matriz de confusión para cada clasificador. Su interpretación es a través de la diagonal donde se muestra que los datos pertenecen a una clase (para este estudio clase a y b), separándolos en los que fueron clasificados correctamente y cuales incorrectamente. Como puede observarse, las diagonales de cada matriz muestran que los clasificadores realizaron una distribución correcta de la muestra, de acuerdo a la clase establecida por cada vector. Por ejemplo, para la Red Neuronal se tiene que de los 44 registros pertenecientes a la clase a todos fueron clasificados correctamente, ya que ninguno se catalogó como clase b, en tanto, para los 46 registros que pertenecian a la clase b, 45 fueron clasificados correctamente, es decir, sólo hubo un falso positivo para la clase a.

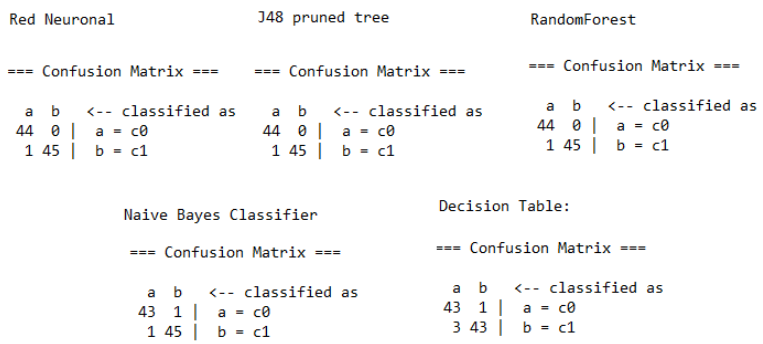

Fig. 2. Matriz de confusión de los resultados de cada clasificador.

La Tabla 3 presenta los valores restantes de la clasificación, referente a los porcentajes de la clasificación correcta y el tiempo de ejecución del clasificador en segundos. El algoritmo de *Decision table* fue el que reportó menor precisión, sin embargo, aún es un porcentaje elevado de eficacia. Por otro lado, el algoritmo J48 es el que reporta más eficiencia y eficacia, para este caso de estudio.

Tabla 3. Resultados obtenidos por los clasificadores que tienen como entrada el vector caraterístico propuesto.

Clasificador	Instancias Correctas	Tiempo de ejecución
Red Neuronal	98.8 %	0.21 segs.
J48	98.8 %	0.02 segs.
Random Forrest	98.8 %	0.07 segs.
Naive Bayes	97.7 %	0.001 segs.
Decision Table	95.5 %	0.04 segs.

7. Discusión

Los resultados que los clasificadores reportan al usar el vector característico propuesto, reportan porcentanjes elevados de precisión. Para hacer un análisis de su efectividad, es necesario realizar una comparación con respecto a los trabajos publicados en el estado del arte.

Es importante no perder de vista, que la información de comparación tiene variación para cada trabajo publicado, considerando el enfoque de aprendizaje empleado, los ataques de intrusión a los que está dirigido el detector, la base de datos de conocimiento usada, los atributos y las clases empleadas o si es supervisado o no. Por lo anterior, primero presentamos una lista de las coincidencias y otra de las diferencias, para que con ello, se tomen los elementos más importantes para la comparación.

- Diferencias:
  1. El origen de los datos empleados en la generación del vector característico propuesto proviene de información real de la red, mientras que los trabajos reportados en el estado del arte utilizan, principalmente, la base de datos KDD99. El número de variables
  2. Los trabajos relacionados usan multivariables con 23 atributos, en contraste con nuestra propuesta que tiene sólo 1 con 17 atributos.
  3. Los ataques de intrusión son diferentes, en este trabajo nos enfocamos al de denegación de servicio.
- Similitudes de comparación
  1. La precisión con la que el clasificador reporta resultados.
  2. Los algoritmos de clasificación J48, Redes neuronales y Naive Bayes y Randiom Forest.
  3. La variación más precisa reportada en cada trabajo relacionado.

La Tabla 5 presenta la relación entre los algoritmos empleados y el porcentaje de precisión que presentan en la clasificación. Como se puede apreciar, nuestra propuesta reporta un elevado porcentaje de precisión en todos los clasificadores, muy cercano a lo propuesto en [8] y en contraste con [19], que reporta un 81.05 % con el algoritmo J48, sin embargo, con el algoritmo Naive Bayes cae hasta el 76.56 %. Finalmente, en la Tabla 4 se presenta el tipo de información utilizada, en donde, claramente se identifica que el vector característico propuesto usa información obtenida de un escenario de ataque real a la red y en estado normal, mientras que los restantes, se apoyan de información generalizada, almacenada en la base de datos DKK99 y sus variaciones.

8. Conclusiones

En este artículo se presenta un método de clasificación para la detección de intrusos en una red, particularmente, del ataque de denegación de servicio, provocado por inundación de paquetes. La propuesta consta de la generación de un vector caraterístico a partir de información obtenida de una red. Este método

Tabla 4. Tabla de comparación en el porcentaje de precisión.

	Red Neuronal	Alg. J48	Naive Bayes	Random Forest
Rivero [18]	98.52 %	99.02 %	98.14 %	NA
Ashfaq [19]	77.41 %	81.05 %	76.56 %	80.67 %
Zhu [20]	NA	NA	76.56 %	80.67 %
Vector Propuesto	98.8 %	98.8 %	97.7 %	98.8 %

Tabla 5. Tabla de comparación usando tipos de muestra y enfoque.

	Muestra	Enfoque
Rivero [18]	KDD99	Supervisado
Ashfaq [19]	KDDCUP99	Semi-supervisado
Zhu [20]	KDD99 y Gure KDD	Supervisado
Vector Propuesto	Lectura directa de la red	Supervisado

permite diferenciar entre los registros provenientes de un ataque y los registros provenientes de un flujo normal.

El vector característico está compuesto por 17 atributos que corresponden a información propia de la red como las IP de origen y destino, el tiempo de captura y toda la información que participa en la transmisión, de tal manera, que permita una clasificación correcta.

Para comprobar que la información contenida en el vector característico permite distinguir claramente un ataque o no, se probó en 5 clasificadores, tales como, una red neuronal, el algoritmo J48 y el Naive Bayes, entre otros. Los resultados reportados por los clasificadores indican que el vector característico permite una categorización precisa y eficiente.

En comparación con los trabajos relacionados, nuestra propuesta consigue obtener un porcentaje elevado de precisión con diferentes tipos de clasificadores, usando información real proveniente de la red, en un enfoque supervisado, al estar constituido con únicamente 17 atributos para su caracterización.

Referencias

1. Debar, H., Becker, M., Siboni, D.: A neural network component for an intrusion detection system. In: Proceedings on Research in Security and Privacy. IEEE Computer Society Symposium, pp. 240–250 (1992)
2. Horng, S., Su, M., Kao, T., Chen, R., Lai, J., Perkasa, C.: A novel intrusion detection system based on hierarchical clustering and support vector machines. Expert Systems with Applications, 38(1), pp. 306–313 (2011)
3. Liao, H., Lin, C., Lin, Y., Tung, K.: Intrusion detection system: A comprehensive review. Journal of Network and Computer Applications, 36(1) pp. 16–2 (2012)
4. Capraru, C.: Detección de anomalías HTTP trazando la sesión web de un usuario. Tesis de Maestría en Seguridad de las Tecnologías de la Información y las

Comunicaciones (MISTIC), Universidad Oberta de Catalunya, España, pp. 1–30 (2016)
5. Rivero, J.: Técnicas de aprendizaje automático para la detección de intrusos en redes de computadoras. Revista Cubana de Ciencias Informáticas, 8(4), pp. 52–73 (2014)
6. Diaz, G., Flores, R., Silva, V.: Sistema Monitor Detector de Intrusos usando TRIPLE-DES96. Tesis de Maestría en Tecnología de Cómputo, Centro de Innovación y Desarrollo Tecnológico en Cómputo, Instituto Politécnico Nacional, México (2014)
7. Flores, J., Lara, Pedro., Gutierrez, M., De los Cobos Silva, S. , Rincón, E.: Un sistema clasificador utilizando coloración de gráficas suaves. Revista de Matemática: Teoría y Aplicaciones, 24(1), pp. 129–156 (2017)
8. Silva, E., Chura, E.: Clasificación de dígitos manuscritos de imágenes digitales. Revista Ciencia & Desarrollo, 1(19), pp. 61–67 (2017)
9. Rubio, J., Hérnandez-Aguilar, J., Stein-Carrillo, J., Ávila-Camacho, F., Meléndez-Ramírez, A.: Sistema sensor para el monitoreo ambiental basado en redes Neuronales. Ingeniería, Investigación y Tecnología, 17(2), pp. 211–222 (2016)
10. Patil, T., Sherekar, S.: Performance analysis of Naive Bayes and J48 classifiation algorithms for data classification. International Journal of Computer Science and Applications, 6(2), pp. 256–261 (2013)
11. Salazar, C.: Generación de Modelos Predictivos de Satisfacción Transaccional para un Centro de Atención a Clientes. Tesis de Maestría en Ciencias Computacionales con Especialidad en Redes y Seguridad Informática, Tecnológico de Monterrey Campus Estado de México, México (2016)
12. Bai, S.: Growing random forest on deep convolutional neural networks for scene categorization. Expert Systems with Applications, 71(1), pp. 279–28 (2017)
13. Tang, F., Ishwaran, H.: Random forest missing data algorithms. Journal arXiv preprint, eprint: 1701.05305 (2017)
14. García, A., Camacho, O., Yáñez, C.: Clasificador de Heaviside. Nova scientia, 7(14), pp. 365–397 (2015)
15. Krishnan, D., Balasubramanian, K.: A Fusion of Multiagent Functionalities for Effective Intrusion Detection System. Security and Communication Networks, 2017(1), pp. 1–15 (2017)
16. Berdun, F., Armentano, M., Amandi, A.: Inferencia de roles de equipo a partir de conductas colaborativas detectadas en 5 interacciones textuales. En: Simposio Argentino de Inteligencia Artificial (ASAI 2016), Buenos Aires, Argentina, Febrero 3, pp. 78–85 (2016)
17. Univaso, P., Ale, J.,Gurlekian, J.: Data Mining applied to Forensic Speaker Identification. IEEE Latin America Transactions, 13(4), pp. 1098–1111 (2015)
18. Rivero Pérez, J. L., Ribeiro, B., Ortiz, K. H.: Comparación de algoritmos para detección de intrusos en entornos estacionarios y de flujo de datos. Universidad y Sociedad, 8(4). pp. 32–42 (2016)
19. Ashfaq, R. A. R., Wang, X. Z., Huang, J. Z., Abbas, H., He, Y. L.: Fuzziness based semi-supervised learning approach for intrusion detection system. Information Sciences, 378(1), pp. 484–497 (2017)
20. Zhu, Y., Liang, J., Chen, J., Ming, Z.: An improved NSGA-III algorithm for feature selection used in intrusion detection. Knowledge-Based Systems, 116(1), pp. 74–85 (2017)
21. Marchang, N., Datta, R., Das, S.: A Novel Approach for Efficient Usage of Intrusion Detection System in Mobile Ad Hoc Networks. IEEE Transactions on Vehicular Technology, 66(2), pp. 1684–1695 (2017)