ISBN 978–987–633–574–4

Estrategias de Pre-procesamiento de Datos para el Análisis de Tráfico de Redes como Problema Big Data

Mercedes Barrionuevo[1], María Fabiana Piccoli[1]
[1] Universidad Nacional de San Luis
Ejército de los Andes 950, San Luis, Argentina
{mdbarrio, mfpiccoli}@unsl.edu.ar

Abstract. Detectar posibles ataques a una red de computadoras requiere contar con métodos o estrategias trabajando en conjunto para la clasificación del tráfico. El área constituye un problema básico de amplio interés sobre todo en conceptos emergentes como Big Data, con sus nuevas tecnologías para almacenar, procesar y obtener información a partir de grandes cantidades de datos.

El reconocimiento del tráfico malicioso en una red depende, en primera instancia, de la eficiencia en la recolección de datos y su correcto pre-procesamiento a fin de ser lo más representativo al aplicar el modelo de análisis de datos elegido. Este tema es el abordado en este trabajo, formando parte de un proyecto integral de detección de ataques a redes de computadoras aplicando Computación de Alto Desempeño en GPU, Inteligencia Artificial y Procesamiento de Imágenes

Keywords: Big Data. Tráfico de redes. Normalización y limpieza de datos. Ataques.

1 Introducción

En la actualidad, la información, los sistemas y las redes informáticas brindan un gran apoyo a diversas empresas y organizaciones convirtiéndose en importantes recursos para las mismas. La confidencialidad, integridad y disponibilidad de la información resultan esenciales para mantener la ventaja competitiva, la rentabilidad, el cumplimiento de las leyes y la imagen institucional. Sin embargo, las organizaciones, sus redes y sistemas de información se enfrentan en forma creciente y constante a amenazas, las cuales buscan afectar la seguridad informática [1].

Un procedimiento de detección de anomalías en una red debe ser capaz de hacer frente al constante incremento en el número de ataques y al gran volumen de datos transferidos, buscando dar soporte a las tareas de monitoreo de red y de identificación de comportamiento anómalo o ataques a las redes [2]. Es por ello que el problema se lo considera un problema Big Data o de Datos Masivos.

La tarea de aprender a detectar ataques implica construir un modelo predictivo, un clasificador, capaz de distinguir entre conexiones “malas”, llamadas intrusiones o ataques, y conexiones normales o “buenas” [3].

Cuando hablamos de conexiones hacemos referencia a una secuencia de paquetes [3] que comienzan y terminan en momentos bien definidos, donde los datos fluyen desde una dirección IP origen a una dirección IP destino según un protocolo bien definido. Algunos expertos en intrusiones creen que la mayoría de los ataques novedosos son variantes de ataques conocidos y la "*firma*" de éstos puede ser

suficiente para detectar variantes novedosas [2]. Por lo tanto, tiene sentido seguir analizando sus variaciones.

Para realizar un buen análisis de las conexiones es necesario una adecuada "preparación de los datos" en la que se eliminen o corrijan aquellos incorrectos y se decida la estrategia a seguir con los incompletos o faltantes. Además, se proyectan los datos para considerar únicamente aquellas variables o atributos que van a ser relevantes, con el objetivo de hacer más fácil la tarea propia de análisis de datos. Esta etapa incluye la selección, limpieza y transformación de los datos; y, a su vez, consta de cuatro subfases: selección de datos, limpieza de datos, construcción de datos (atributos derivados, registros generados), y formateo de datos [4].

En [5, 6] se presentaron resultados satisfactorios de un sistema para la detección de ataques usando algunos específicos. Como la base de datos utilizada para la evaluación de los algoritmos de clasificación eran de prueba, estas ya tenían sus datos en el formato requerido. Es por ello que es de interés en este trabajo hacer frente a la Etapa 2 mostrada en la Figura 2.2: Etapa de Pre-procesamiento de Datos.

Por lo tanto, el objetivo planteado en este trabajo es mostrar el procesamiento realizado a los datos recolectados directamente del tráfico de red, previo a la aplicación de los algoritmos paralelos de minería de datos y de visualización mostrados en [5,6] para la detección de ataques o posibles anomalías.

Este documento está organizado como sigue: la siguiente sección describe los conceptos teóricos involucrados en el desarrollo de este trabajo. La sección 3 detalla el preprocesamiento, transformación y limpieza realizado a los datos. Finalmente se muestran los resultados experimentales obtenidos, y se detallan las conclusiones y líneas futuras de trabajo.

2 Marco Teórico

En esta sección se analizan brevemente diferentes conceptos, entre los cuales se destacan distintos aspectos referidos a los datos en contextos de Big Data, su normalización y correlación. Todo relacionado con el problema que nos interesa: el análisis de tráfico de redes para la detección de ataques. Cada uno de estos temas se aborda en las siguientes secciones.

2.1 Big Data

El volumen de datos circulante en la red de redes ha alcanzado niveles inimaginables en la última década y, al mismo tiempo, los dispositivos de almacenamiento han reducido de forma significativa sus precios. Las empresas privadas e instituciones de investigación capturan terabytes de datos de la interacción de los usuarios, redes sociales y de diversos sensores.

La clave en la era de Big data son los datos, los cuales se pueden utilizar para responder a muchas preguntas, pero no a todas. En la actualidad el trabajo con datos presenta ciertos retos a afrontar:

- El aumento masivo del volumen de datos puede implicar una disminución en la calidad del análisis.

■ En los grandes volúmenes de datos no siempre hay contexto, por lo cual se deberá contar con expertos del tema, por ejemplo, con ingenieros en redes de telecomunicaciones.
■ Los datos cambian cada cierto tiempo, por lo cual pueden llegar a generar inconsistencias, si se ha tenido en cuenta sólo un tipo de dato de entrada.
■ Los datos que comprueban las hipótesis planteadas pueden ser difíciles de obtener.

Por ello, el desafío de esta era es darle sentido a este gran conjunto de datos. El análisis de datos en Big Data involucra recolectar datos de diferentes fuentes, unirlos y/o mezclarlos para ser tratados por los analistas, para finalmente, entregar resultados de utilidad a la organización de interés.

El proceso de convertir grandes cantidades de datos no estructurados para ser datos útiles a las organizaciones no es una tarea trivial. Por lo tanto, se deben combinar diferentes estrategias y metodologías para lograr una mejor respuesta.

Al trabajar con datos, uno de los primeros pasos necesarios para hacer un análisis de datos es determinar qué tipo de estrategia usar: descriptiva, exploratoria, inferencial, predictiva, causal o mecánica. Las respuestas a las preguntas mostradas en la Figura 2.1 determinarán el enfoque a usar en la resolución del problema.

En nuestro caso, buscamos predecir si una conexión cumple o no con ser un ataque. Parece un buen inicio utilizar un análisis predictivo, siempre y cuando, los datos del tráfico de una red hayan podido ser puestos en un estado homogéneo. En consecuencia, se debe preparar los datos para un enfoque predictivo.

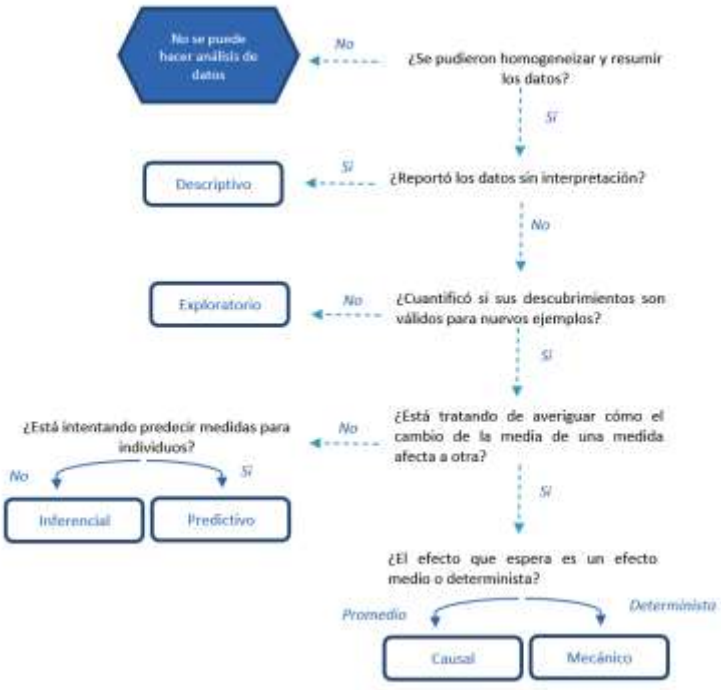

Fig 2.1 Diagrama de flujo del tipo de análisis de datos según la pregunta

En las siguientes secciones detallamos el ciclo de vida de los datos en entornos Big Data, la metodología a implementar y las estrategias empleadas para realizar el pre-procesamiento de los datos.

2.1.1 Ciclo de vida de Big Data

Existen metodologías como SEMMA [11] que son incompletas dado que ignoran las etapas de recopilación de datos. Estas etapas generalmente constituyen la mayor parte del trabajo en un proyecto exitoso de Big Data.

En la figura 2.2 se muestran las etapas por las que tienen que pasar los datos en un proceso de Big Data, siendo la Etapa 2 de Pre-procesamiento de los datos nuestro punto de interés. Este ciclo comienza luego de definir el problema y evaluar correctamente cuánto potencial de ganancia tiene para una organización, siendo buena estrategia investigar y/o analizar lo que otras organizaciones han implementado en la misma situación estudiando soluciones que sean razonables para su compañía.

Lo primero a considerar en un ciclo de análisis de Big Data es “adquirir y recopilar los datos” definiendo cuáles datos serán de relevancia: Etapa 1.

Fig 2.2: Etapas de Big Data

Una vez que los datos son recuperados, es necesario almacenarlos en un formato fácil y apto de usar. Luego es necesario almacenarlos en una base de datos siendo una de las alternativas más comunes a utilizar el sistema de archivos Hadoop [7], Spark [8], entre otros; “preprocesar los datos”, implica remodelar los datos limpios recuperados previamente y usar estadísticas para la imputación de valores perdidos, detección de valores atípicos, normalización, extracción de características y selección de funciones: Etapa 2.

“Modelar y analizar datos” implica probar diferentes modelos y esperar resolver el problema empresarial en cuestión. En la práctica, normalmente se desea que el modelo proporcione algunos conocimientos del negocio seleccionando el mejor modelo o combinación de modelos evaluando su rendimiento en un conjunto de datos excluido: Etapa 3; y finalmente, “implementar y evaluar” esta etapa implicaría aplicar el modelo a nuevos datos y una vez que la respuesta esté disponible, evaluar el modelo: Etapa 4.

2.1.2 Metodología

En términos de metodología, el análisis de Big Data difiere significativamente del enfoque estadístico tradicional. El análisis comienza con los datos, luego se los modela para obtener una respuesta.

En aplicaciones de análisis a gran escala, se necesita una gran cantidad de trabajo (normalmente el 80% del esfuerzo) sólo en la limpieza de los datos, para ser utilizado posteriormente por modelos de aprendizaje automático.

Una vez que los datos se pre-procesan están disponibles para el modelado, los resultados de las evaluaciones en los diferentes modelos deben ser los razonables y/o esperados. Finalmente, una vez implementado el modelo, se deben informar tales evaluaciones y resultados adicionales al experto en el problema, para que analice si la información obtenida le aporta conocimiento o no.

Si bien no se tiene una metodología única a seguir en aplicaciones reales a gran escala, normalmente una vez definido el problema se aplican estas pautas generales en la mayoría de los problemas.

2.2 Aspectos Generales del Pre-procesamiento de Datos

La etapa de preparación de los datos consiste en aplicar limpieza, normalización de los datos y la selección de características. Cada una de estas fases se describen brevemente en las siguientes secciones.

2.2.1 Limpieza de datos

Una vez que los datos son recolectados, pueden existir diversas fuentes de datos con diferentes cantidades de atributos. Es importante preguntarse si es práctico homogeneizarlos.

Si las fuentes de datos son completamente diferentes, la pérdida de información puede resultar muy grande al homogeneizarlas. En este caso, podemos pensar en alternativas. ¿Puede una fuente de datos ayudarnos a construir un modelo de regresión y la otra un modelo de clasificación? ¿Es posible trabajar con la heterogeneidad a nuestro favor en lugar de simplemente perder información? Tomar estas decisiones es lo que hace que la analítica sea interesante y desafiante.

En este punto necesitamos limpiar los datos no estructurados, convertirlos en una matriz de datos para aplicar algún algoritmo como así también eliminar o reemplazar datos con valores nulos. Particularmente, para el problema que nos convoca los datos pueden ser recolectados por los administradores de red usando distintas herramientas, las cuales generan datos con distintos formatos y variados atributos.

2.2.2 Normalización de Datos

En muchos algoritmos basados en distancias es necesario escalar los datos, es decir normalizar el rango de valores numéricos, las distancias debidas a diferencias de un

atributo que van entre 0 y 1000 serán mucho mayores que aquellas debidas a diferencias de un atributo variando entre 0 y 10.

Como consecuencia, es necesario aplicar alguna función de normalización a los datos. Para ésto existen muchos métodos, siendo la técnica *z-score* la más utilizada por su sencillez en el cálculo. Este método conserva el rango (máximo y mínimo) e introduce la dispersión de la serie (desviación estándar/varianza), transformando linealmente los valores de tal manera que el valor medio de los datos transformados es igual a 0 mientras que su desviación estándar es igual a 1. La fórmula de transformación es la correspondiente a la ecuación (1).

$$x = (x_1 - \mu) / \sigma \; . \qquad (1)$$

Donde x es la muestra actual, x_1 es la muestra transformada, μ denota la media de los datos y σ representa la desviación estándar.

2.2.3 Selección de Características

La selección de características es fundamental para la detección de ataques o anomalías. Este proceso consiste en dar un peso a cada característica para determinar cuál de ellas es la que tiene mayor impacto.

La ponderación de características mejora la precisión, logrando un mayor rendimiento. Las métricas comúnmente conocidas para la selección de características son *chi-cuadrado* (CHI), *ganancia de información*, *coeficiente de correlación* y *razón de probabilidades* (OR)[9].

Por tratarse de valores numéricos y por su simplicidad en el cálculo, la métrica utilizada en este trabajo es el *Coeficiente de Correlación* entre variables.

La correlación, también conocida como *Coeficiente de Correlación Lineal* (de Pearson), es una medida de regresión que pretende cuantificar el grado de variación conjunta entre dos variables. Es una medida estadística que cuantifica la dependencia lineal entre dos variables, es decir, si se representan en un diagrama de dispersión los valores que toman dos variables, señalará lo bien o lo mal que el conjunto de puntos representados se aproxima a una recta. Formalmente, la podemos definir como el número que mide el grado de intensidad y el sentido de la relación entre dos variables, ver ecuación **(2)** [10].

$$\rho(x,y) = \mathrm{cov}(x,y) \; / \; \sigma x * \sigma y \; . \qquad (2)$$

Siendo la covarianza entre dos variables definida como:

$$\mathrm{cov}(x,y) = (\Sigma \; (x_i - x)(y_i - y))/n \; . \text{ para } i=1...n$$

Los valores que puede tomar la correlación son: $\rho = -1$ para la correlación perfecta negativa, $\rho = 0$ cuando no existe correlación y $\rho = +1$ para la correlación perfecta positiva.

La Limpieza, Normalización y Selección de Características forman parte de la Etapa 2 mostrada en la Figura 2.2 referida al preprocesamiento de los datos. Llevar adelante estas etapas dan origen a la propuesta de este trabajo.

3 Pre-procesamiento del Tráfico de Redes

Este trabajo forma parte de un proyecto integral, el cual aplica un modelo de aprendizaje predictivo, donde se combinan técnicas de clasificación, análisis por similitud, visualización de datos y Computación de Alto desempeño en su solución.

El problema a afrontar implica reconocer en un tiempo razonable ataques a una red, y de una manera lo más confiable posible. Para iniciar con esta tarea se define a $X = \{x_1, x_2, \ldots, x_n\}$ como una conexión de red, donde cada atributo representa los valores intervinientes en una comunicación. Por cada conexión, se evalúa intentando predecir si es normal, un ataque o una anomalía teniendo en cuenta los valores de cada uno de los atributos y sus relaciones.

Luego de definir el problema, se debe contar con los datos a analizar, ésto se logra mediante la recolección de los datos de la red. A continuación, se describen los pasos realizados para la normalización, utilizando la técnica z-score, y la selección de características, según la correlación de datos, ambas descritas anteriormente. Las tareas a desarrollar son:

- **Eliminación de datos Nulos y Anómalos:** Una vez que los datos son recolectados pueden haber valores fuera de lo normal o valores faltantes en algunos de los atributos recolectados. Por lo tanto, se deben eliminar aquellos datos donde no existe ni dirección IP origen ni destino, o son direcciones de multicast o broadcast limitado. Éstas últimas no aportan información útil para las reglas en la clasificación de los datos.

Aproximadamente el 20% de los datos son eliminados por ser del tipo multicast, broadcast ilimitado o poseer valores nulos. Las conexiones con valores nulos deben analizarse por separado para evaluar cuáles son los datos faltantes y si son anomalías.

- **Selección de Características:** La normalización de los datos es un proceso costoso desde el punto de vista computacional, más cuando se trabaja con mucha cantidad de datos como es este caso. Por ello, es necesario, previo a la normalización, determinar cuáles serán las características con las que se trabajará para determinar si existe un ataque, anomalía o no. Como se mencionó anteriormente, nosotros seleccionamos aquellas características independientes entre sí, en consecuencia, su determinación será según el coeficiente de correlación, el mismo se obtiene aplicando (2) en cada una de las características. Serán seleccionadas aquellas que estén más cercanas a cero o no superen un umbral de correlación definido.
- **Normalización de los Datos:** Al trabajar con la gran cantidad de datos circulantes en una red, existen muchos atributos con distintos rangos de valores. Por ejemplo, al convertir una dirección IP a un número decimal, el valor máximo es 4.294.967.295 si es un broadcast (255.255.255.255), sin embargo, otros atributos pueden tomar valores entre 0 a 1024 si se trata de evaluar puertos bien conocidos.

Para cada uno de los atributos seleccionados, se procede con su normalización aplicando la ecuación (1). Para la misma se debe calcular previamente la media y la varianza del conjunto de datos.

Una vez que los datos han sido recolectados, transformados, normalizados y seleccionados, se continúa con la siguiente etapa del proceso, por ejemplo, aplicar reglas de clasificación para determinar cuáles de esas conexiones son ataques conocidos generando *firmas*, y luego, establecer similitudes entre las conexiones y las *firmas* para determinar potenciales ataques.

4 Resultados Experimentales

En esta sección se presentan los experimentos realizados en el Laboratorio de Redes de la Universidad Nacional de San Luis y el análisis de los resultados obtenidos. Cada una de las etapas consideradas se realizaron de la siguiente manera:

- **Recolección de Datos.**
En nuestro caso usamos la herramienta *tshark* durante un día generando 24 archivos (1 por hora), donde los atributos recuperados son: direcciones IP origen y destino, puertos origen y destino, y protocolo utilizado en la comunicación. A partir de ellos se generan nuevos atributos tales como: clase, número de red y host de cada dirección IP. Esto es necesario para aplicar las reglas de clasificación a cada conexión.
El comando utilizado es:
tshark -f \tcp or udp or icmp{T elds |E separator=, -e frame.time relative -e ip.src -e ip.dst -e tcp.srcport -e tcp.dstport -e udp.srcport -e udp.dstport -nni eth0 > trafico_a_analizar.txt

- **La correlación entre los datos**
En la Tabla 1 se puede observar el nivel de correlación de los atributos donde las direcciones IP están fuertemente relacionadas con sus respectivas clases, red y host con valores cercanos a 1. Mientras que existe muy baja o nula dependencia con el resto de los atributos.

Tabla 1. Correlación entre los atributos de las conexiones.

	clase_o	red_o	host_o	pto_o	pto_d	prot	ip_d	red_d	host_d	clase_d
ip_o	0,75	0,75	-0,64	-0,15	-0,11	-0,46	-0,11	-0,08	-0,05	-0,05
ip_d	-0,05	-0,05	-0,08	-0,15	-0,11	-0,48	1,00	0,69	-0,66	0,69
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
prot	0.30	0.20	0.10	0.40	0.40	1.00	0.30	0.20	0.10	0.01

- **Selección de características:**

De las pruebas realizadas anteriormente se pudo determinar que el análisis de correlaciones entre los atributos nos permite decidir cuáles son los parámetros correlacionados o dependientes, estableciendo cuáles son los atributos significativos a ser considerados, por ejemplo, en el cálculo de la función euclidiana utilizados por el algoritmo *k-nn* para la evaluación de la similitud de las conexiones con las firmas de ataques conocidos. En este caso, de los 11 parámetros utilizados para las reglas de clasificación sólo son considerados 5 *(dirección IP origen y destino, puerto origen y destino y protocolo)* los que nos interesan. Esta reducción de atributos permite mitigar el alto costo computacional involucrado en el cálculo de la función euclidiana para cada una de las conexiones.

- **La normalización de los datos:**

Se realizó tomando los atributos seleccionados de cada conexión: (dirección IP origen y destino, puertos origen y destino) y aplicándoles la función z score a cada uno de ellos, dando como resultado los valores como se muestran en la Tabla 2.

Tabla 2. Normalización de los datos.

IP origen	IP destino	pto origen	pto destino	prot
-0,3152596	-0,33519186	-0,3274522	-0,48115293	-1,4115906
-0,3152595	-0,33519186	-0,4152103	-0,47791469	0,290621
……..	……..	……..	…….	……..
-0,31525969	-0,335191866	-0,42403127	-0,48569784	1,99283384

Estos valores se corresponden para el análisis de un día del tráfico de la red del Laboratorio, según los paquetes obtenidos y seleccionados después de la limpieza.

Una vez realizada la etapa de pre-procesamiento de los datos, se comprobaron los datos obtenidos aplicando los modelos que incluyen las reglas de clasificación utilizadas y el algoritmo de los *k*-nn más cercanos para la detección de valores similares a ataques conocidos. En la Tabla 3 se muestra el valor del cálculo de la función para distintas 5-uplas. Se incluye en dicha tabla una columna donde se especifica si es un ataque o no y otra columna con el valor de la función euclidiana.

Tabla 3. Valor de la función euclidiana

IP origen	IP destino	pto origen	pto destino	prot	es_ataque	f. euclidiana
-0,315	-0,335	-0,327	-0,481	-1,411	si	0
-0,315	0,652	2,937	-0,460	-1,411	no	0,987
-0,315	-0,335	-0,415	-0,477	0,290	no	1,702
-0,315	-0,335	-0,415	-0,477	0,290	no	1,702

En este caso el algoritmo *k-nn* se ejecuta con k=4, devolviendo las 4 filas mostradas en la Tabla 3. Donde si se realiza la inversa de la normalización, se puede observar que para los casos en que la función era 0, la conexión analizada es exactamente igual a la firma considerada como ataque, mientras que los otros valores muestran ser anomalías. En estos casos se observa que son ataques a otros puertos, a otro protocolo o son ataques sin puertos especificados. Esto se muestra en la Tabla 4 en un formato entendible para el experto del dominio de redes.

Tabla 4. Inversa de la Normalización

IP origen	IP destino	pto origen	pto destino	prot
10.230.34.73	10.255.255.255	1500	80	tcp
10.230.34.146	10.255.255.255	137	137	udp
10.230.34.146	10.255.255.255	137	137	udp
10.230.34.12	10.255.255.255			icmp

5 Conclusiones y Trabajos Futuros

Este trabajo presenta una metodología a aplicar en la etapa de pre-procesamiento de los datos en un sistema de Big Data, particularmente aplicado al dominio de seguridad en el tráfico de Redes de Computadoras. Para ello proponemos llevar a cabo la selección de características y la normalización de los datos mediante funciones estadísticas bien conocidas.

La aplicación de esta propuesta fue evaluada en una red, mostrando resultados satisfactorios, no sólo respecto a las respuestas obtenidas sino también al desempeño del sistema en general al realizarse la limpieza de datos.

Como líneas futuras, se propone aplicar técnicas de programación paralela en la normalización y análisis de correlación, como así también en la secuenciación de las etapas de manera de crear estructuras similares a arquitecturas de pipeline y *overlaping* de etapas, particularmente en la Etapa 1 y 2. Además, se prevé ampliar los conocimientos usando diversas técnicas de aprendizaje de máquina y/o redes neuronales a fin de comparar los modelos utilizados en este trabajo y, de ser necesario, mejorar los existentes.

Referencias

1. Tulasi.B, R. S. Wagh, and B. S., “High performance computing and big data analytics-paradigms and challenges,” International Journal of Computer Applications, vol. 116, Abril 2015.
2. Terzi,D. S., Terzi, R. and Sagiroglu, S. “Big data Analytics for Network Anomaly Detection from Netflow Data,” IEEE, 2017.
3. Ghimes, A. M., and Patriciu,V. V. “Neural network models in big data analytics and cyber security,”in 2017 9th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), pp. 1–6, June 2017.
4. Hernández Orallo, J.; Ramírez Quintana, M. J.; Ferri Ramírez, C. “Introducción a la Minería De Datos” ISBN eBook: 978-84-8322-558-5.
5. Barrionuevo M., Lopresti M., Miranda N., Piccoli F.. “Secure Computer Network: Strategies and Challenges in Big Data Era”. JCC&BD 2018. VI Jornadas de Cloud Computing & Big Data. La Plata (Buenos Aires), 25 al 29 de junio de 2018. ISBN 978-950-34-1659-4
6. Barrionuevo M., Lopresti M., Miranda N., Piccoli F.. “An Anomaly Detection Model in a LAN using K-NN and High Performance Computing Techniques”. Congreso Argentino de Ciencias de la Computación. CACIC 2017. http://sedici.unlp.edu.ar/handle/10915/63951
7. White, T.. “Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale”. O'Reilly Media, Inc. ISBN 1491901713, 9781491901717. 2015
8. Perrin, J. “Spark in Action, Second Edition: Covers Apache Spark 3 with Examples in Java, Python, and Scala”. 2do Edition. ISBN 1617295523, 9781617295522. Simon and Schuster, 2020.
9. Ikram,S., Kumar, C. “Intrusion detection model using fusion of chi-square feature selection and multi class SVM.” J. King Saud Univ. Comput. Inf. Sci. 29(4): 462-472. 2017.
10. Peiro Ucha, A.. “Coeficiente de correlación lineal”. Economipedia.com. 2015.
11. Azevedo, A., Santos, M. “KDD, SEMMA and CRISP-DM: a parallel overview”. IADIS European Conf. Data Mining. Pp 182-185. 2015.