Sistema de clasificación de tráfico de red usando autocodificadores para la detección de anomalías

Barra Morales, Kevin Axel

2025

https://hdl.handle.net/20.500.11777/6237

Sistema de clasificación de tráfico de red usando autocodificadores para la detección de anomalías

Barra Morales Kevin Axel (noveno semestre en Ingeniería en Sistemas Computacionales)[1], Lameda Díaz Stevanato Leon (noveno semestre en Ingeniería en Sistemas Computacionales)[1], Suárez Falcón Valeria Samantha (noveno semestre en Ingeniería en Sistemas Computacionales)[1, *], Morúa Álvarez Nora del Rocío (profesor responsable)[1], Pérez Aguirre Rafael (profesor asesor)[1]
[1]*Universidad Iberoamericana Puebla, San Andrés Cholula, Puebla, México*

Palabras clave: Tráfico de red, detección de anomalías, aprendizaje automático
[*]**Autor Corresponsal:** valeriasamantha.suarez@iberopuebla.mx

Introducción

En los últimos años, la frecuencia y sofisticación de los ciberataques ha aumentado. Según el 2023 *Data Breach Report* del *Identity Theft Resource Center*, entre 2021 y 2023 se registró un incremento del 72% en las violaciones de datos en los Estados Unidos, afectando a más de 350 millones de individuos [1]. Además, *Cybersecurity Ventures* estimó que los daños globales causados por ciberataques alcanzaron los 8 billones de dólares en 2023, posicionando al cibercrimen como la tercera economía más grande del mundo, solo por detrás de EE. UU. y China [2].

Este panorama evidencia que los sistemas tradicionales de detección de intrusiones, basados en firmas, resultan insuficientes para identificar amenazas emergentes y ataques previamente desconocidos dado a su dependencia exclusiva de actualizaciones constantes y su incapacidad para reconocer comportamientos nuevos o variantes de amenazas [3].

El presente proyecto tiene como objetivo desarrollar un sistema de clasificación de tráfico de red mediante el uso de autocodificadores para la detección de anomalías. Un autocodificador es un tipo de red neuronal artificial que aprende a representar los datos de entrada en un espacio reducido, llamado representación latente, para luego reconstruirlos lo más fielmente posible, en el ámbito de la ciberseguridad, esta capacidad de reconstrucción permite modelar el comportamiento normal del tráfico de red, así; cuando el autocodificador recibe un flujo de datos que difiere significativamente de lo aprendido, presenta un mayor error de reconstrucción, indicando la presencia de una posible anomalía [4].

Metodología

Para el desarrollo del sistema, se inició con la fase de preprocesamiento de datos. Se utilizó el *dataset CICIDS2017*, reconocido en investigaciones de ciberseguridad por contener una mezcla de tráfico realista y ataques simulados. Se eliminaron columnas no numéricas y redundantes, se fusionaron los conjuntos en un único archivo (*DataFrame*) y se limpiaron eliminando espacios, valores nulos (*NaN*) y filas duplicadas. Posteriormente, se dividió el tráfico en dos grupos: benigno (*BENIGN*) y maligno (cualquier otro tipo de ataque). Los datos fueron normalizados mediante *MinMaxScaler* para que todas las variables estuvieran en un mismo rango (0 a 1), mejorando la estabilidad y el aprendizaje de los modelos.

En la etapa de entrenamiento, se usaron estos datos preprocesados para entrenar dos autocodificadores (*autoencoders*), uno para tráfico benigno y otro para tráfico malicioso. Se calculó el error de reconstrucción de cada muestra, lo que indicó qué tan familiar era el flujo para el modelo. Con estos errores, se definió un umbral de confianza (*trust threshold*) en el percentil 95 para cada modelo, marcando el límite entre flujos conocidos y anómalos. La información de entrenamiento utilizada se guardó para asegurar la consistencia al momento de clasificar nuevos datos.

En la fase de inferencia, se cargaron los modelos y escaladores guardados. Cada nuevo flujo de red fue normalizado y evaluado con ambos autocodificadores. Se calculó su error de reconstrucción y se transformó en una puntuación de confianza (1), según la fórmula:

$$trust\ score = 100 \times \left(1 - \frac{Error\ de\ reconstrucción}{Error\ máximo}\right) \qquad (1)$$

El flujo se clasificó como BENIGNO si el trust score del modelo benigno fue igual o superior al del modelo maligno; de lo contrario, se clasificó como MALIGNO. Finalmente, se generó una matriz de confusión para evaluar el desempeño del sistema.

Resultados y Discusión

El presente proyecto evaluó la detección de tráfico anómalo en redes, enfocándose en identificar ataques de denegación de servicio distribuido (*DDoS*) a partir de flujos de red procesados del conjunto de datos *CSE-CIC-IDS2018*.

Como primer paso, se analizó la distribución de los flujos de red contenidos en el conjunto de prueba. Tras el preprocesamiento de los datos y la generación de un gráfico de distribución, se observó que el 42.62% de los flujos correspondía a tráfico BENIGNO y el 57.38% a tráfico MALIGNO (Fig 1).

Posteriormente, los datos preprocesados fueron sometidos a inferencia mediante el *pipeline* desarrollado. El modelo predijo que **50.01%** de los flujos pertenecían a la categoría ***BENIGN*** y **49.99%** a ***MALIGN***, mostrando una clasificación general equilibrada, aunque con una ligera desviación respecto a la distribución real de las muestras (Fig 2).

Para evaluar la precisión de las predicciones, se realizó una comparación directa entre los resultados inferidos y las etiquetas reales mediante una matriz de confusión. De un total de 223,083 flujos analizados, los resultados obtenidos fueron (Fig 3):

- 95,040 flujos BENIGNOS fueron correctamente clasificados (verdaderos positivos).
- 28 flujos MALIGNOS fueron erróneamente clasificados como BENIGNOS (falsos negativos).
- 111,501 flujos MALIGNOS fueron correctamente clasificados (verdaderos positivos).
- 16,513 flujos BENIGNOS fueron erróneamente clasificados como MALIGNOS (falsos positivos).

A partir de estos resultados, se calcularon las siguientes métricas de evaluación (Tabla 1):

Tabla 1: La tabla presenta las métricas de evaluación del modelo: *precision*, *recall* y *accuracy*, que indican la calidad de las predicciones en términos de aciertos y errores

Métrica	MALIGN (%)	BENIGN (%)
Sensibilidad (*Recall*)	99.97	85.15
Precisión (*Precision*)	87.11	99.97
Exactitud general (*Accuracy*)	92.75	-

Estas métricas indican que el modelo tiene un desempeño sobresaliente en la identificación de tráfico benigno, alcanzando una sensibilidad (*recall*) del 99.97%. La clasificación de tráfico maligno también es sólida, pero ligeramente inferior en *recall* (87.11%), sugiriendo que parte del tráfico maligno es confundido como benigno, aunque de forma contenida

La diferencia en el desempeño puede explicarse, en parte, por el desbalance en la cantidad de datos utilizados durante el entrenamiento: mientras que el modelo de tráfico benigno fue entrenado con 2,095,057 muestras, el modelo de tráfico maligno sólo contó con 425,741 muestras.

En comparación con investigaciones similares en el campo, este patrón coincide con estudios que evidencian que un mayor volumen y diversidad de datos de entrenamiento contribuye significativamente a mejorar la capacidad de generalización de los modelos de detección de anomalías [5].

Conclusiones

Los principales hallazgos indican que el modelo fue capaz de identificar el tráfico benigno con una sensibilidad (*recall*) del 99.97%, y logró una exactitud global del 92.75% sobre el conjunto de datos de prueba basado en tráfico real. Estos resultados son especialmente relevantes, ya que validan que un enfoque de aprendizaje automático no supervisado, cuando es correctamente entrenado y balanceado, puede alcanzar niveles de detección comparables o superiores a los métodos tradicionales basados en firmas o reglas específicas.

Desde una perspectiva de ingeniería, el estudio aporta una metodología robusta y replicable que integra técnicas de preprocesamiento automatizado de datos de red, entrenamiento de modelos diferenciados para tráfico benigno y maligno y generación de umbrales de confianza específicos para tomar decisiones de clasificación. Esto permite a los profesionales del área diseñar sistemas adaptables a distintas topologías de red y perfiles de tráfico, contribuyendo directamente al fortalecimiento de las estrategias de ciberseguridad.

En cuanto a las posibles líneas futuras de investigación, se propone incrementar y balancear las muestras de tráfico maligno para mejorar la capacidad de detección y reducir los falsos positivos y aplicar técnicas de *transfer learning* para adaptar modelos previamente entrenados a nuevas condiciones de tráfico sin necesidad de reentrenar desde cero.

La principal contribución original de este estudio radica en la implementación práctica de un sistema de detección de anomalías capaz de funcionar de forma autónoma en entornos de red reales, combinando eficiencia computacional, facilidad de despliegue y alta precisión de detección. Esto abre nuevas posibilidades para el diseño de sistemas de defensa cibernética inteligentes, basados en la analítica avanzada de tráfico de red.

Referencias

[1] Identity Theft Resource Center, "2023 Data Breach Report," ITRC, 2023. [Online]. Available: https://www.idtheftcenter.org/post/2023-data-breach-report/

[2] Cybersecurity Ventures, "2023 Official Cybercrime Report," 2023. [Online]. Available: https://cybersecurityventures.com/cybercrime-damages-6-trillion-by-2021/

[3] S. Bhuyan, D. Bhattacharyya, and J. Kalita, "Network Anomaly Detection: Methods, Systems and Tools," IEEE Communications Surveys & Tutorials, vol. 16, no. 1, pp. 303–336, 2014. doi: 10.1109/SURV.2013.052213.00046

[4] D. Bergmann and C. Stryker, "Autoencoder," *IBM Think*, IBM. [Online]. Available: https://www.ibm.com/think/topics/autoencoder. [Accessed: 25-Apr-2025].

[5] K. Aslansefat, W. Bridges, and Y. Papadopoulos, "*SafeML and Intrusion Detection Evaluation Dataset (CICIDS2017)*," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/code/kooaslansefat/cicids2017-safeml

Anexo

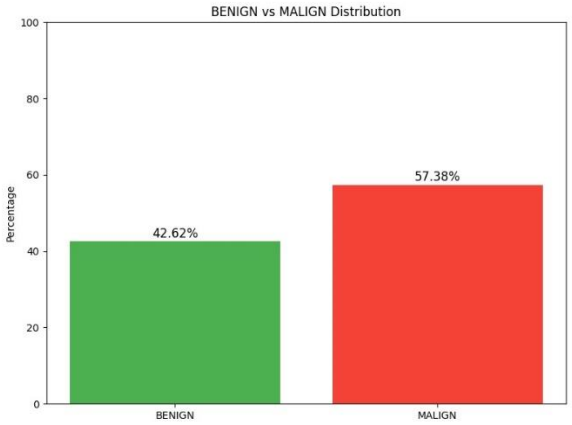

Fig. 1. Distribución porcentual de flujos de red clasificados como tráfico BENIGNO y MALIGNO en el conjunto de datos de prueba, posterior al preprocesamiento.

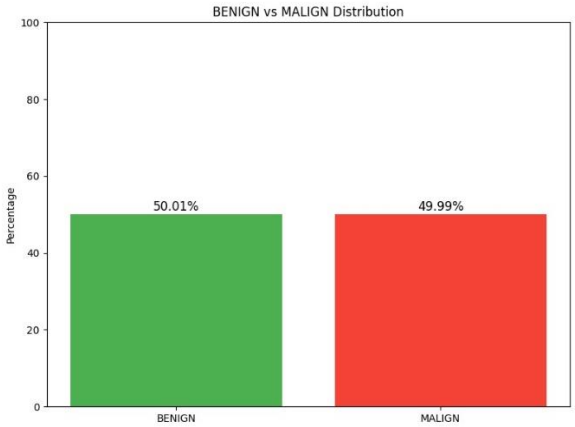

Fig. 2. Distribución porcentual de las predicciones realizadas por el modelo de inferencia sobre el conjunto de datos de prueba, diferenciando tráfico BENIGNO y MALIGNO.

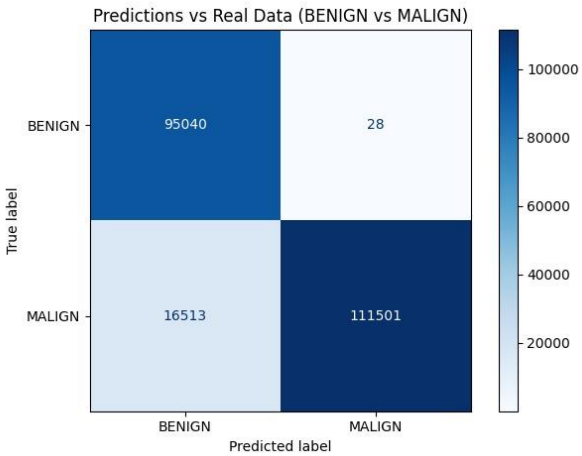

Fig. 3. Matriz de confusión que compara las predicciones del modelo con las etiquetas reales del conjunto de prueba, mostrando el número de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos.