ISSN: 2805-1807

ABORDANDO EL DESEQUILIBRIO DE DATOS EN CLASIFICACIÓN DE ATAQUES DE DENEGACIÓN DE SERVICIO DISTRIBUIDO (DDOS)

Acosta-Tejada, Danny
Polytechnic University of Puerto Rico
San Juan, Puerto Rico
acosta_130476@students.pupr.edu

Sanchez-Galan, Javier
Universidad Tecnologica de Panamá
Panamá City, Panamá
javier.sanchezgalan@utp.ac.pa

Torres-Batista, Nelliud
Polytechnic University of Puerto Rico
San Juan, Puerto Rico
ntorres@pupr.edu

Abstract

DDoS attacks pose a substantial menace to organizations and enterprises reliant on interconnected networks for their operational continuity. Distinguishing between malicious assaults and legitimate surges in web traffic represents a formidable challenge, with existing defense mechanisms struggling to precisely identify and counteract such threats. This research delves into the ramifications of data imbalance on the classification of Distributed Denial of Service (DDoS) attacks, presenting a remedy involving synthetic data. The methodology encompasses data acquisition, preprocessing, synthetic data generation, and performance assessment. Leveraging the CICDDoS2019 dataset, comprising 20 million instances characterized by 88 features, for evaluation purposes, synthetic data is artfully crafted through a Generative Adversarial Network (GAN). Specifically focusing on three original dataset attributes: time, attack type, and duration. Classification tasks entail three distinct dataset arrangements, ranging from balanced to imbalanced representations of attack categories. This is achieved by treating the dataset conventionally (imbalanced), subsampling the minority class, and supplementing the dataset with an additional 2 million data points synthesized by GANs. An evaluative comparison between conventional classification methodologies (CNN, KNN, and XGBoost) and GAN utilization demonstrates a substantial performance enhancement. While traditional methods yield accuracy rates

https://doi.org/10.33412/apanac.2023.3922

of 82-86%, GANs consistently achieve 98-99% accuracy. These findings underscore the pronounced impact of data imbalance on classification efficacy and underscore the efficacy of GANs in mitigating this challenge while enhancing accuracy. The research underscores the critical significance of accounting for data imbalance and adopting innovative techniques such as GANs in the realm of cybersecurity.

Keywords: DDoS, Cybersecurity, Machine Learning, Generative Adversarial Networks, Data Generation.

Resumen

Los ataques de denegación de servicio distribuido (DDoS) representan una amenaza significativa para instituciones y empresas que dependen de redes interconectadas. Distinguir entre ataques maliciosos y aumentos legítimos en el tráfico web es un desafío, y los sistemas de defensa existentes luchan por identificar. Este estudio explora el impacto del desequilibrio de datos en la clasificación de ataques DDoS y propone una solución utilizando datos sintéticos. La metodología involucra: recolección de datos, preprocesamiento, generación de datos sintéticos, y análisis de rendimiento. Utilizamos CICDDoS2019 dataset, contiene 22 millones de ejemplos medidos en 88 características. Generamos datos sintéticos utilizando Redes Generativa Antagónica (GANs), centrándonos en tres características del conjunto de datos: tiempo, tipo de ataque y duración. Se trabajó con tres grupos de datos del mismo dataset: manera convencional (desequilibrada), submuestreo de la clase minoritaria y utilizando GANs para generar un total adicional de 2 millones de puntos de datos. Una comparación de rendimiento entre métodos tradicionales de clasificación (CNN, KNN y XGBoost) y el uso de GANs muestra una mejora significativa. Los métodos tradicionales alcanzan tasas de precisión del 82-86%, mientras que las GANs logran consistentemente tasas de precisión del 98-99%. Estos hallazgos resaltan el impacto del desequilibrio de datos en la eficacia de la clasificación y demuestran la efectividad de las GANs para mitigar este desafío mientras mejoran la precisión. El estudio enfatiza la importancia de considerar el desequilibrio de datos y adoptar técnicas innovadoras como las GANs en el campo de la ciberseguridad.

Palabras claves: DDoS, Seguridad de datos, Aprendizaje de Maquinas, Redes Generativa Antagónica (GANs), Generación de Datos.

https://doi.org/10.33412/apanac.2023.3922

1. INTRODUCCIÓN

En el panorama digital actual, las empresas dependen en gran medida de redes interconectadas para agilizar sus operaciones y ofrecer servicios [1]. Sin embargo, esta creciente dependencia de la tecnología conlleva numerosos desafíos, siendo el más destacado la capacidad para distinguir entre los ataques de denegación de servicio distribuido (DDoS) y los aumentos legítimos en el tráfico web [2]. Los ataques DDoS representan una amenaza significativa para las organizaciones al sobrecargar sus recursos de red y perturbar los servicios. Diferenciar entre ataques maliciosos y picos genuinos en el tráfico web es una tarea compleja que requiere mecanismos avanzados de defensa. Los métodos y sistemas de defensa existentes se basan en la identificación de diferencias estadísticas en el tráfico de red para diferenciar entre ataques DDoS y multitudes repentinas [3]. Sin embargo, estas aproximaciones a menudo tienen dificultades para realizar distinciones precisas, lo que conduce a la clasificación errónea. Esta limitación subraya la necesidad urgente de técnicas de ciberseguridad más efectivas. Con su capacidad para reconocer patrones derivados de valores característicos, el aprendizaje automático ha surgido como un camino para combatir diversas amenazas cibernéticas.

Eventos recientes como el de Ticketera [4] en el cual confirma haber sido víctima de un ciberataque destinado a monopolizar la venta de boletos, ejecutado a través de bots que apuntaron a la infraestructura técnica de la empresa, lo que resultó en riesgos financieros y de reputación significativos. Dada la ubicuidad de estas vulnerabilidades en diversos sectores, incluyendo atención médica, automotriz, finanzas, seguridad y la rápida proliferación de dispositivos conectados a Internet [5], abordar los desafíos de seguridad se convierte en una prioridad. Este estudio busca ampliar el análisis de los métodos de clasificación al incorporar datos sintéticos para corregir el problema del desequilibrio de datos. Al explorar la utilidad de los datos sintéticos para abordar problemas de desequilibrio de datos, nuestra investigación pretende realizar contribuciones significativas al avance de los mecanismos de defensa de ciberseguridad, garantizando así un entorno digital más seguro tanto para las organizaciones como para las personas.

Este artículo aborda el desequilibrio de datos en los métodos de clasificación del aprendizaje automático y sus implicaciones para la ciberseguridad. Nuestro objetivo es doble: primero, examinar el impacto del desequilibrio de datos en la clasificación de ataques DDoS utilizando métodos tradicionales y técnicas de aprendizaje profundo de vanguardia; y segundo, proponer un remedio mediante la utilización de datos sintéticos para mitigar las preocupaciones por el desequilibrio de datos. Específicamente, realizamos una comparación de rendimiento entre metodologías tradicionales y enfoques de aprendizaje profundo de última generación utilizando el conjunto de datos CICDDoS2019, diseñado específicamente para la clasificación de ataques DDoS.

https://doi.org/10.33412/apanac.2023.3922

2. MÉTODO

Nuestra metodología se basa en seis pasos: (1) adquisición de datos, (2) luego se realizan los pasos de preprocesamiento para adecuar los datos. (3) Presentamos una estrategia de generación de datos sintéticos debido al desequilibrio en el conjunto de datos. Posteriormente, (4) probamos nuestro modelo; se probaron (5) tres algoritmos de aprendizaje automático y (6) se calcularon métricas de error para determinar cuál proporcionaba un resultado más preciso.

A. Adquisición de datos

Este estudio construye una metodología de análisis de clasificación en torno a un conjunto de datos reconocido y previamente publicado, el CICD-DoS2019 [6]. El conjunto de datos incluye 88 características y 7 registros, que abarcan diferentes tipos de ataques DDoS y tráfico de red normal. El conjunto de datos seleccionado se utilizó con fines de clasificación para diferenciar y reconocer una variedad de ataques del tráfico normal, es decir, no ataques. Para este estudio, el conjunto de datos se reorganizó de tres formas distintas:
Primer Montaje (Equilibrada mediante submuestreo): análisis seleccionando un subconjunto y asegurando una representación equitativa de todas las clases en las distribuciones de ataques, como se muestra en la Tabla I.
Segundo Montaje (Desequilibrada): se creó seleccionando todo el conjunto de datos sin equilibrar, los detalles sobre el conjunto de datos se encuentran en la Tabla II.
Tercer Montaje(Equilibrada mediante la generación de datos sintéticos): en la Tabla III, se aseguró un equilibrio de clases generando un conjunto de datos utilizando GANs.

Tabla 1. Primer Montaje

Tipo de Ataque	Ataques
LDAP	9.931
MSSQL	205.744
NETBIOS	383.183
Portmap	186.960
SYN	265.120
UDP	278.059
UDPLag	1.873
Total de Ataques	1.330.870

https://doi.org/10.33412/apanac.2023.3922

Tabla 2. Segundo Montaje

Tipo de Ataque	Ataques
LDAP	1.915.122
MSSQL	5.787.453
NETBIOS	3.657.497
Portmap	186.960
SYN	4.891.500
UDP	3.867.155
UDPLag	1.873
Total de Ataques	20.307.560

Tabla 3. Tercer Montaje

Tipo de Ataque	Ataques
LDAP	1.915.122
MSSQL	5.787.453
NETBIOS	3.657.497
Portmap	1.520.290
SYN	4.891.500
UDP	3.867.155
UDPLag	1.285.403
Total de Ataques	22.924.420

B. Preprocesamiento

En este estudio, se realizó un exhaustivo preprocesamiento de datos con el propósito de transformar los datos crudos en un formato estructurado. Esta tarea incluyó la imputación de valores faltantes y la eliminación de datos ruidosos, inconsistentes y redundantes [7]. El objetivo primordial de este proceso fue garantizar la coherencia de los datos y reducir la dimensionalidad del conjunto de datos para facilitar su análisis [8]. La Figura 1 ilustra de manera visual los pasos que se llevaron a cabo en la etapa de preprocesamiento de datos de este estudio.

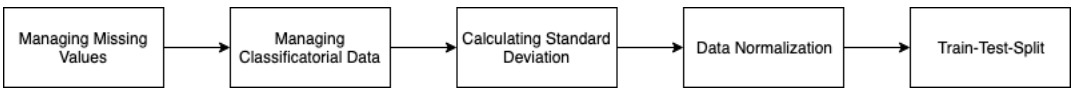

Fig. 1 Preprocesamiento de Datos

En cuanto a la gestión de valores faltantes, se reconoció la importancia de abordar este problema común en conjuntos de datos utilizados en el aprendizaje automático. En estudios previos, como el de Emmanuel et al. [9], se demostró que la omisión de valores faltantes

https://doi.org/10.33412/apanac.2023.3922

puede tener un impacto negativo en la precisión y la puntuación F1 de los modelos de clasificación. En este contexto, se realizó un minucioso análisis en busca de valores faltantes utilizando la potente herramienta '.isnull()' de Pandas. Afortunadamente, nuestro análisis no identificó ningún valor faltante que pudiera afectar nuestro estudio.

Asimismo, se abordó la gestión de datos categóricos, siendo las direcciones IP de origen y destino, la marca de tiempo y las etiquetas las variables no numéricas detectadas. Para lidiar con estas variables categóricas, se aplicaron técnicas específicas. En primer lugar, las direcciones IP se convirtieron en números enteros mediante el uso de la biblioteca netaddr. IPAddress, lo que facilitó su análisis y modelado. Posteriormente, la variable de marca de tiempo se transformó en un número entero, preservando el formato original de cadena, eliminando los dos puntos de la marca de tiempo y manteniendo la secuencia temporal de los datos. Por último, se reemplazaron los nombres de los ataques con valores numéricos, lo que simplificó la manipulación y el análisis de las etiquetas, proporcionando una representación numérica que se emplearía en los métodos de clasificación.

La gestión de la desviación estándar (Standard Deviation) en los conjuntos de datos utilizados en el aprendizaje automático se consideró crucial. Por lo tanto, se llevó a cabo una normalización de datos para reducir la desviación estándar y garantizar que las características del conjunto de datos tuvieran una escala similar. Tras esta gestión, se eliminaron con éxito todas las columnas donde la desviación estándar era igual a cero, lo que resultó en un conjunto de datos con 75 columnas para análisis posteriores.

La normalización de la puntuación Z se utilizó para estandarizar los datos, manteniendo las relaciones originales entre los puntos de datos y eliminando la influencia de escalas y distribuciones variables. Por último, se aplicó la técnica de división de entrenamiento-prueba, para garantizar un rendimiento confiable del modelo en datos no vistos. Nuestro estudio utilizó una división de entrenamiento-prueba del 75% para entrenamiento y del 25% incluyendo los datos generados sintéticamente mediante GANs.

C. Generacion de Datos con GANs

Los Generative Adversarial Networks (GANs), introducidos por [10], presentaron el concepto de GANs, que constan de un generador y una red neuronal discriminadora que se entrenan de manera competitiva. El generador produce muestras de datos sintéticos, mientras que el discriminador intenta distinguirlas de las muestras de datos reales. La Figura 2 describe la arquitectura de GANs utilizada en este artículo:

Datos Reales: Empleamos CICDDoS19, que está compuesto por una colección las solicitudes realizadas al servidor del conjunto de datos de evaluación de CICDDoS2019.

Penalización del Gradiente: Se utilizó la Penalización del Gradiente (GP) [11]. En términos generales, esto ocurre cuando se agrega ruido a los datos de entrada en proporción al gradiente de la salida del discriminador con respecto a los datos de entrada.

https://doi.org/10.33412/apanac.2023.3922

Red del Generador: El generador se entrena minimizando la diferencia entre las muestras de datos generadas y reales (error de clasificación).
Red del Discriminador: Entrenamos al discriminador para distinguir entre datos reales y falsos.

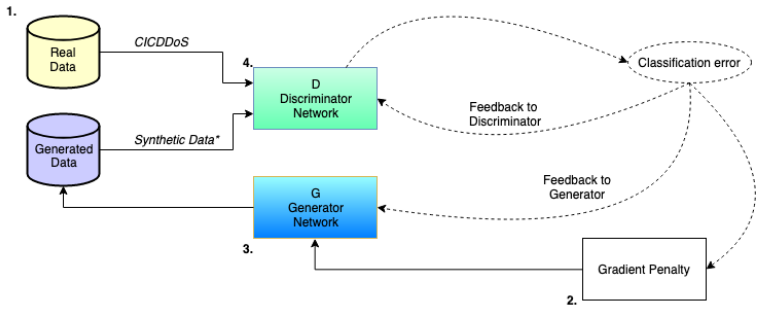

Fig. 2 Arquitectura GANs

D. Métodos de Clasificación

CNN: Las Redes Neuronales Convolucionales (CNN, por sus siglas en inglés) son una clase de redes neuronales profundas adecuadas para procesar datos de imágenes. Las CNN están compuestas por capas de convolución que aprenden características locales de los datos de entrada [12], seguidas de capas de agrupación que reducen la dimensionalidad espacial y capas completamente conectadas que realizan la clasificación final.
KNN: K-nearest neighbors (KNN) es un algoritmo de clasificación no paramétrico ampliamente utilizado en reconocimiento de patrones y aprendizaje automático [13]. La idea básica de KNN es clasificar un punto de datos desconocido basado en las etiquetas de clase de sus k vecinos más cercanos en el conjunto de entrenamiento [14].
XGBoost: XGBoost (Extreme Gradient Boosting) es un método de conjunto que combina múltiples aprendices débiles para formar un aprendiz fuerte [15]. XGBoost es una variante del algoritmo de aumento de gradiente introducido por Friedman [15]. El algoritmo fue posteriormente desarrollado y optimizado por Chen et al. [16].
Hardware
La configuración experimental para esta investigación implicó la implementación de los modelos descritos utilizando Python 3.10. El equipo utilizado estaba equipado con un procesador AMD Ryzen 5 3600XT, 32GB de memoria RAM DDR4 y una tarjeta de video EVGA NVIDIA 3080. La implementación del código se basó en las bibliotecas TensorFlow 2.8 y scikit-learn 1.1.

3. RESULTADOS

Por medio de esta investigación se pudo identificar que los ataques UDPLag y LDAP están subrepresentados. La Tabla 1 muestra que se obtuvo un total de 191,694 de los

https://doi.org/10.33412/apanac.2023.3922

cuales se identificaron que estos ataques tienen una representación de 0.746% y 0.141% equitativamente.

El segundo montaje, denotado en la Tabla 2, mostró que al utilizar el conjunto de datos completo se obtuvo un desbalance mayor. Al analizar los ataques, se pudo denotar que los ataques UDPLag (0.009%) y Portmap (0.921%) obtuvieron un desbalance mayor, el cual exigía un nivel de datos mayor para alcanzar un balance.

Una vez generado los datos sintéticos con GANs, se pudo presentar un conjunto de datos mejor representado. Tabla 3 muestra un conjunto de datos con una distribución más equitativa entre las diferentes categorías. Esta distribución equilibrada sugiere que el conjunto de datos abarcó una gama más amplia de categorías con magnitudes similares.

Después de generar datos con una GAN, el conjunto de datos se normalizó y se combinaron los datos sintéticos con el conjunto de datos reales. Se entrenaron modelos de clasificación (KNN, CNN y XGBoost). El rendimiento se evaluó utilizando la precisión y el puntaje F1, que se pueden ver en la Tabla 4 y la Tabla 5, respectivamente.

Los resultados se pueden resumir de la siguiente manera: independientemente del método de clasificación (CNN, KNN o XGBoost), la precisión y puntaje F1, los mejores resultados se encontraron en el tercer conjunto de datos, es decir, utilizando el conjunto de datos más equilibrado obtenido a través de una GAN. En algunos casos, mejorando la evaluación del modelo en más del 10%.

Tabla 4. Comparación de Precisión

Classification Method			
Dataset	**CNN**	**KNN**	**XGBoost**
Mismo Tamaño	82,42%	83,87%	85,66%
Dataset Completo	88,46%	91,49%	92,57%
GANs	97,58%	98,15%	99,44%

Tabla 5. Comparación de Puntaje F1

Classification Method			
Dataset	**CNN**	**KNN**	**XGBoost**
Mismo Tamaño	82,42%	83,87%	85,66%
Dataset Completo	88,46%	91,49%	92,57%
GANs	97,58%	98,15%	99,44%

https://doi.org/10.33412/apanac.2023.3922

4. CONCLUSIONES

En conclusión, nuestro estudio empleó tres clasificadores diferentes: dos clasificadores tradicionales, CNN y KNN, y un clasificador novedoso, XGBoost, para evaluar la efectividad de incorporar datos sintéticos en el proceso de clasificación. Los tres modelos se entrenaron utilizando una combinación de datos reales y sintéticos. El uso de un conjunto de datos combinado demostró una mejora en términos de precisión y puntaje F1. Al aprovechar las representaciones sólidas aprendidas a partir del conjunto de datos aumentado, todos los clasificadores exhibieron capacidades de generalización mejoradas, especialmente cuando se enfrentaban a datos reales limitados y difíciles de obtener. Este artículo también aborda el problema del desequilibrio de datos en los métodos de clasificación de aprendizaje automático y sus consecuencias para la ciberseguridad. Por lo tanto, se cumplieron nuestros dos objetivos. Nuestros hallazgos resaltan los beneficios de utilizar datos sintéticos en la fase de entrenamiento de clasificadores de aprendizaje automático para reconocer ataques DDoS. Al aumentar los conjuntos de datos reales limitados de ataques con muestras generadas de manera sintética, observamos un mejor rendimiento de clasificación en los tres clasificadores.

REFERENCIAS

[1] S. Randhawa, B. Turnbull, J. Yuen, and J. Dean, “Mission-centric automated cyber red teaming,” in Proceedings of the 13th International Conference on Availability, Reliability and Security, 2018, pp. 1–11.

[2] C. Cangea, P. Velic˘kovic´ , N. Jovanovic´ , T. Kipf, and P. Lio`, “Towards sparse hierarchical graph classifiers,” arXiv preprint arXiv:1811.01287, 2018.

[3] D. Sun, K. Yang, Z. Shi, and C. Chen, “A new mimicking attack by LSGAN,” Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI, vol. 2017-Novem, pp. 441–447, 2018.

[4] R. Times, “Ticketera confirms that it was the victim of a new cyberattack,” Jul 2022. [Online]. Available: https://rivaltimes.com/ ticketera- confirms- that- it- was- the- victim- of- a- new-cyberattack/

[5] N. F. Syed, Z. Baig, A. Ibrahim, and C. Valli, “Denial of service attack detection through machine learning for the IoT,” Journal of Information and Telecommunication, vol. 4, no. 4, pp. 482–503, 2020.

[6] I. Sharafaldin, A. H. Lashkari, S. Hakak, and A. A. Ghorbani, “Devel- oping realistic distributed denial of service (ddos) attack dataset and taxonomy,” in 2019 International Carnahan Conference on Security Technology (ICCST). IEEE, 2019, pp. 1–8.

[7] S.-A. N. Alexandropoulos, S. B. Kotsiantis, and M. N. Vrahatis, “Data preprocessing in predictive data mining,” The Knowledge Engineering Review, vol. 34, p. e1, 2019.

[8] M. Kang and J. Tian, “Machine learning: Data pre-processing,” Prog- nostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things, pp. 111–130, 2018.

https://doi.org/10.33412/apanac.2023.3922

[9] [9] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, “A survey on missing data in machine learning,” Journal of Big Data, vol. 8, no. 1, pp. 1–37, 2021.
[10] I.J.Goodfellow,“Ondistinguishabilitycriteriaforestimatinggenerative models,” arXiv preprint arXiv:1412.6515, 2014.
[11] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved training of wasserstein gans,” Advances in neural information processing systems, vol. 30, 2017.
[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” Communications of the ACM, vol. 60, no. 6, pp. 84–90, 2017.
[13] T. Cover and P. Hart, “Nearest neighbor pattern classification,” IEEE transactions on information theory, vol. 13, no. 1, pp. 21–27, 1967.
[14] M. Mustaqeem and M. Saqib, “Principal component based support vec- tor machine (pc-svm): a hybrid technique for software defect detection,” Cluster Computing, vol. 24, no. 3, pp. 2581–2595, 2021.
[15] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” Annals of statistics, pp. 1189–1232, 2001.
[16] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

AUTORIZACIÓN Y LICENCIA CC

https://doi.org/10.33412/apanac.2023.3922