Comparación de técnicas de *machine learning* para la detección de *phishing*

Andrés Eduardo Moncada-Vargas

El *phishing* es el robo de datos personales a través de una página web falsa donde se le pide al usuario ingresar sus datos para validar su identidad frente a una supuesta entidad legítima. En este trabajo, se compararon técnicas de *machine learning* y se determinó que la técnica bosque aleatorio es la más efectiva en casos donde las características de las páginas tengan un valor exacto, y árbol de decisión es la más efectiva en casos donde las características hayan sido analizadas y se haya determinado una clasificación en base a dicha característica.

Comparison of Machine Learning Techniques for Phishing Detection

Phishing is the act of stealing personal data through a false Web page. Users are asked by a supposedly legitimate company to enter their private information in order to verify their identity. This research work compared different machine learning techniques and determined that the random forest technique is the most effective one when the characteristics of the pages have an exact value, and the decision tree is the most effective one when the characteristics of the pages have been analyzed and a classification has been determined based on such characteristics.

Comparación de técnicas de *machine learning* para la detección de *phishing*

Andrés Eduardo Moncada-Vargas
20152102@aloe.ulima.edu.pe

Resumen: El *phishing* es el robo de datos personales a través de una página web falsa donde se le pide al usuario ingresar sus datos para validar su identidad frente a una supuesta entidad legítima. En este trabajo se compararon técnicas de *machine learning* y se determinó que la técnica bosque aleatorio es la más efectiva en los casos en que las características de las páginas tengan un valor exacto y árbol de decisión es la más efectiva en aquellos casos en los cuales las características hayan sido analizadas y se haya determinado una clasificación en base a dicha característica.

Introducción

El *phishing* es una de las mayores amenazas en la actualidad, siendo la causa de la mayoría de los robos y estafas cibernéticas en los últimos años. Por dicho motivo, se desarrollaron varias técnicas para detectar páginas *phishing*. Con el paso del tiempo, los métodos de crear páginas *phishing* se fueron innovando, por lo que se empezó a usar *machine learning* para la detección de *phishing*. En este trabajo se desarrollará una metodología de comparación de técnicas de *machine learning* para determinar cuál es la mejor en distintos casos. Por ese motivo, se buscará replicar experimentaciones realizadas en el pasado por otros autores, para determinar si los resultados coinciden y realizar una comparación basada en los resultados de otros autores y los obtenidos en esta experimentación.

Materiales y métodos

Clasificadores de *machine learning*: árbol de decisión, regresión logística, redes neuronales (perceptrón de multicapas), máquina de soporte vectorial y bosque aleatorio.

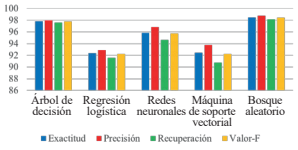
Datasets, Phishing_BestFirst y Phishing_Infogain, obtenidos de la experimentación de Islam *et al.* (2016), Phishing_Legitimate_Full, obtenido de la experimentación de Cuzzocrea *et al.* (2018).

Proceso. Los hiperparámetros de los clasificadores fueron refinados, optimizando los resultados, y se compararon los resultados obtenidos con los vistos en los trabajos de los autores mencionados, para determinar el mejor clasificador en cada caso.

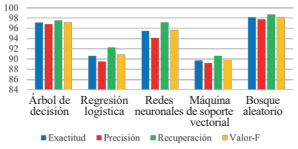
Resultados

Los resultados obtenidos se muestran en los siguientes gráficos:

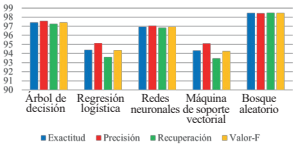
Phishing_BestFirst



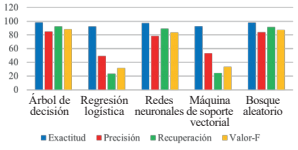
Phishing_Infogain



Phishing_Legitimate_Full



PhishingData



Tras analizar los resultados, se observó lo siguiente:

- Los clasificadores de árboles (árbol de decisión y bosque aleatorio) fueron los más efectivos.
- Bosque aleatorio obtuvo mejores resultados que árbol de decisión con cada *dataset*, a excepción del último, PhishingData.
- Regresión logística y máquina de soporte vectorial fueron los menos efectivos al momento de clasificar páginas *phishing*, en comparación con los otros tres clasificadores.

Conclusiones

Tras investigar técnicas de *machine learning* y distintos casos de uso, se concluye que los clasificadores de árboles tienen un mejor rendimiento en distintas métricas para la clasificación de páginas *phishing*. Aunque se ha comprobado que los métodos de aprendizaje automático pueden ser apropiados para la identificación de sitios web *phishing*, debe considerarse que los procedimientos de análisis deben ser ampliados a nuevos conjuntos de datos para afianzar el universo de casos de uso examinados. En trabajos futuros se espera examinar ese comportamiento.

Referencias

Chiew, K. L., Tan, C. L., Wong, K. S., Yong, K. S., y Tiong, W. K. (2019). A New Hybrid Ensemble Feature Selection Framework for Machine Learning-Based Phishing Detection System. *Science Direct*, 14. doi: 10.1016/j.ins2019.01.064

Cuzzocrea, A., Martinelli, F., y Mercaldo, f. (2018). Applying Machine Learning Techniques to Detect and Analyze Web Phishing Attacks. *ACM Digital Library*, 5. doi: 10.1145/3282373.3282422

Islam Mamun,M. S., Rathore, M. A., Lashkari, A. H., Stakhanova, N., y Ghorbani, A. A. (2016) Detecting Malicious URLs Using Lexical Analysis. *Springer Link*, 16. doi: 10.1007/978-3-319-46298-1_30

UNIVERSIDAD DE LIMA