Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Vicepresidencia Primera del Gobierno · Ministerio de Asuntos Económicos y Transformación Digital · Secretaría de Estado de Digitalización e Inteligencia Artificial
Plan de Recuperación, Transformación y Resiliencia
España | digital 2026
incibe_ Instituto Nacional de Ciberseguridad

Alicia Mayor Bal
presenta el resumen del trabajo

Aplicación de aprendizaje trasnferido a la asignación de maliciosidad de IPs

David Escudero García, Noemí DeCastro-García
Actas VIII JNIC, atlanTTic, 85-92, (2023).

10 de julio de 2024

SERIE DE RESÚMENES EN ESPAÑOL

Palabras clave: Aprendizaje automático, Aprendizaje transferido, Ciberataques, Deriva conceptual, Direcciones IP.

Introducción

El uso de técnicas de aprendizaje automático es muy frecuente para la detección y prevención de ciberataques. Sin embargo, se suele suponer una distribución de datos estacionaria, lo cual no es realista. Estos modelos cuentan con **deriva conceptual**, lo cual compromete su eficacia con datos nuevos.

En este trabajo se plantea el uso de técnicas de **aprendizaje transferido** (rama del aprendizaje automático) para contrarrestar la degradación producida en los modelos por esta deriva conceptual en problemas de clasificación multiclase para grados de maliciosidad en IPs, permitiendo aprovechar datos de un conjunto fuente para mejorar su aprendizaje y capacidad predictiva.

1. Metodología

Se utiliza un **conjunto de datos** de 99720 eventos asociados a direcciones IP, divididos en 4 niveles de maliciosidad (1, 3, 6 y 9). El conjunto está **desbalanceado**, con más del 50 % de muestras pertenecientes a la clase 6.

Se mide la magnitud de la **deriva conceptual** de los datos, obteniendo un valor en torno a 0.4. Es una cifra moderadamente alta por lo que se concluye la existencia de deriva conceptual.

Los **modelos usados** son el XGBoost y la librería autosklearn, cuya optimización de hiperparámetros se realiza de forma conjunta mediante el algoritmo SMAC.

Se aplican 10 **algoritmos de aprendzaje transferido** implementados con Pyton: EasyTL, ACM, TCA, MCS, CORAL, SDA, VSA, MSLDA, DTFC y SRARA. Cada uno cuenta con sus propios hiperparámetros que también son optimizados.

El **conjunto de características** utilizado cuenta con: dirección IP de origen, timestamp del evento y geolocalización (país de la dirección IP, su latitud y su longitud). Al tener un número pequeño de características, se añaden algunas otras derivadas de literatura relacionada.

Se define una red como una **red CIDR/24** para encontrar direcciones IP pertenecientes a cada clase de maliciosidad.

La parte experimental se divide en las siguientes tareas:

- Dividir conjunto de datos en entrenamiento y test.
- Entrenar los modelos y algoritmos de aprendizaje transferido con el conjunto de entrenamiento.
- Predecir sobre el conjunto de datos de test.

Este proceso se repite para cada combincación de conjuntos de características, modelo y algoritmo.

2. Resultados

Se utiliza como métrica el **coeficiente de correlación de Matthews**, con valores de -1 (predicciones completamente opuestas a las etiquetas reales) a 1 (predicciones perfectas).

Los algoritmos ACM, EasyTL y MCS cuentan con un rendimiento particularmente malo, que no supera en ningún caso el 0.322, debido a su planteamiento en centroides de clases.

DTFC, TCA y SRARA son los algoritmos más **consistentes** para ambos modelos, sin registrar empeoramiento para ningún conjunto de características y contanto con valores de hasta 0.793.

A modo general, se observa que la aplicación de algoritmos de aprendizaje transferido a este problmea **no ofrece mejoras notables de forma consistente**. Solamente se perciben mejoras superiores al 20 % en casos concretos con ciertos conjuntos de características y algoritmos.

3. Discusión/Conclusiones

Tras la evaluación experimental de varios algoritmos de aprendizaje transferido al problema de asignación de maliciosidad en IPs, los resultados de **mejoras no son consistentes** para diferentes conjuntos de datos y algoritmos.

Por ello, la aplicación de aprendizaje transferido puede no resultar eficaz, siendo mejores modelos para esta problemática los de tipo **ensablado**, que son más robustos y resistentes ante cambios en la distribución.

4. Valoración del documento original

Este artículo cuenta con una sección inicial en la que se explican de forma concisa los conceptos más importantes para el desarrollo del estudio (como la deriva conceptual, el dominio o la tarea), lo cual hace que desde un primer momento se entienda el proceso de forma sencilla en comparación con otros estudios similares. Además, cabe destacar que los resultados se ofrecen de forma desagregada para cada modelo y sus algoritmos más destacados, tanto positiva como negativamente.