Elaborato Sicurezza Informatica

TECHINT: Analisi statica dataset

Introduzione

Lo scopo dell'elaborato è di effettuare l'analisi statica di un dataset attraverso Machine Learning. In particolare i dati a disposizione rappresentano il traffico di rete raggiunto da una Darknet honeypot, che in quanto tale non poteva rilevare alcun traffico genuino. Ai fini dell'analisi sono state ingegnerizzate delle nuove feature per identificare i potenziali attaccanti e sulla base di queste (oltre al resto del dataset originale) sono stati applicati diversi algoritmi di clustering per rilevare possibili modus operandi. Infine sono state utilizzate le label assegnate dagli algoritmi di clustering e le feature create per provare ad effettuare early classification. Per limitazioni riguardo al costo computazionale dell'elaborazione l'analisi effettuata riguarda 100.000 istanze del dataset originario.

Feature Engineering

Per identificare possibili attaccanti sono state utilizzate le feature "src_port" che identifica la porta sorgente del traffico e "src_ip" che rappresenta l'indirizzo ip della sorgente del traffico, queste due feature sono state combinate in un unica feature a cui è stato assegnato lo stesso valore per tutte le istanze con la stessa coppia src_ip/src_port osservate in intervalli di 30 minuti. Per ogni attaccante così identificato sono state create altre tre feature: la prima porta a cui tenta di collegarsi, la lista di porte a cui tenta di collegarsi in ordine temporale ed infine la lista di porte a cui si è collegato ignorando l'ordine temporale.

Clustering

Durante la fase di clustering sono stati utilizzati tre algoritmi in particolare:

- **DBSCAN**: Non fa assunzioni sulla forma dei cluster e sulla loro dimensione relativa. Vede i cluster come aree ad alta densità separate da aree a bassa densità, e supporta un elevato numero di sample.

- **Agglomerative**: Applica il clustering gerarchico e permette di utilizzare diversi metodi di linkage. Supporta un elevato numero di cluster e sample.
- **Kmeans**: Assume che i cluster abbiano una forma convessa, uguale dimensione e separa i dati in n gruppi con uguale varianza.

Le performance di ogni algoritmo sono state valutate utilizzando due metriche diverse: Silhouette coefficient e Calinski-Harabasz Index. Il Silhouette coefficient va da un minimo valore di -1 (clustering non corretto) fino ad un massimo di +1 (clustering ad alta densità) ed è generalmente più alto per cluster convessi, il Calinski-Harabasz Index ha valori alti per cluster densi e ben separati e richiede un basso tempo di calcolo.

Risultati:

	Silhouette coefficient	**Calinski-Harabasz Index**
DBSCAN (5 cluster)	0.4136	9419.35
Agglomerative (5 cluster)	0.6727	6653.52
Kmeans (5 cluster)	0.2707	29639.11

I cluster identificati da DBSCAN e Agglomerative sono simili in quanto raggruppano la maggior parte delle istanze in un solo cluster mentre Kmeans suddivide il dataset in modo più bilanciato tra diversi cluster.

Early classification

La possibilità di effettuare early classification è stata testata usando la feature “1_port” che rappresenta la prima porta a cui si è tentato di collegarsi da parte di un attaccante. Per fare ciò è stato creato un dataset contenente solo questa feature e le label identificate dagli algoritmi di clustering come classi, questi dati sono poi stati forniti in input ad una rete neurale con un solo livello nascosto costituito da 5 unità con funzione di attivazione “ReLU” e il livello di uscita con funzione di attivazione softmax.

Risultati:

	Accuracy Training set	**Accuracy Test set**
DBSCAN	92.71%	92.69%
Agglomerative	98.50%	98.52%
Kmean	79.52%	79.76%

Gli elevati valori di accuracy riscontrati per DBSCAN e Agglomerative sono probabilmente dovuti al fatto che, come già descritto, entrambi gli algoritmi hanno raggruppato quasi tutte le istanze in un solo cluster e ciò spiega anche perché l'accuracy ottenuta con le label generate da Kmean è inferiore (in quanto i cluster che trova sono più bilanciati).