# Utilizzo di Soft-Brownian-Offset per la generazione di attacchi ai fini dell'addestramento di rilevatori di intrusioni

Guglielmo Bartelloni
Relatore: Andrea Ceccarelli — andrea.ceccarelli@unifi.it

Partiamo con l'introduzione dei rilevatori di intrusioni o Intrusion Detection Systems (IDS). Un è un applicazione volta a monitorare continuamente la rete per identificare attività malevole.

Questo tipo di strumenti sono ormai essenziali per individuare tempestivamente gli attacchi informatici.

Per migliorare questi sistemi, nell'ultimo decennio, si è iniziato ad utilizzare algoritmi di Intelligenza Artificiale/Machine Learning.

Il machine learning e' la branca che si occupa di sviluppare modelli che permettano alle macchine di imparare dai dati.

Questi modelli però sono molto sensibili ai dati di addestramento e, generalmente, i dati o dataset che rappresentano traffico di rete sono perlopiù composti da pacchetti di rete rappresentanti del traffico comune mentre per restante parte da pacchetti di attacchi. Quindi i dataset sono generalmente sbilanciati. Inoltre non vanno a coprire molti casi di attacco come gli zero day cioè attacchi mai visti prima.

- ADFA: attacchi (40961) e normali (91045) (132k)
- CICIDS: attacchi (132503) e normali (67494) (199k)

Un metodo per sopperire a questa mancanza è quello di generare nuovi dati a partire da quelli già esistenti (Data Augmentation) per poi raffinare i modelli utilizzando questi nuovi elementi. Nella tesi esploreremo un algoritmo di Data Augmentation chiamato Soft-Brownian-Offset (SBO) e lo utilizzeremo per generare nuovi dati di attacco per l'addestramento di un modello di classificazione dei dati ossia XGBoost che rappresenta il nostro Intrusion Detection System. In particolare Soft-Brownian-Offset permette di generare campioni così detti out-of-distribution (OOD), cioè campioni che non fanno parte della distribuzione dei dati di partenza. Questo è coerente col fatto che molti pacchetti di attacchi sono molto "diversi", in termini di caratteristiche, dai pacchetti normali e perciò fuori dalla distribuzione. Soft-Brownian-Offset è un algoritmo di generazione di dati sintetici inizialmente creato per la generazione di eventi anomali nel caso dei cyberphycical systems, si pensi per esempio a macchine a guida autonoma dove il rilevamento di casi limite come un ciclista che attraversa improvvisamente la strada è di vitale importanza. In questo caso i dati necessitano di essere non troppo dissimili da quelli originali per rappresentare degli scenari il più reali possibile. Abbiamo quindi utilizzato Soft-Brownian-Offset per il nostro caso cioè la generazione di nuovi pachetti di attacco.

In particolare ci siamo serviti di due dataset uno complesso "CIC-IDS" e, uno semplice "Adfa-Net" per mettere in evidenza eventuali divergenze. Abbiamo poi esplorato vari approcci di generazione per ogni dataset a seconda dei dati di partenza, cioè generando i dati a partire da:

- Solo pacchetti normali
- Solo pacchetti di attacchi
- Dataset completo (pacchetti normali e di attacchi)

Prima dell'addestramento del modello sono stati creati dei grafici per poter fare un'analisi qualitativa delle generazioni.

Generazione da pacchetti normali:

In figura si nota, come questo metodo non sia molto efficace infatti, ci sono delle evidenti sovrapposizioni tra pacchetti normali e pacchetti generati, questo potrebbe provocare dei problemi perché i nostri pacchetti che dovrebbero rappresentare degli attacchi sono molto simili a quelli normali mentre, in corrispondenza degli attacchi, questo non accade. Le figure successive sono invece i test eseguiti su CICIDS che ottengono i medesimi risultati. E' evidente che i pacchetti generati, siano troppo simili ai pacchetti normali e troppo dissimili da quelli di attacco. Questa valutazione qualitativa potrebbe gia darci un'idea sull'efficacia di addestramento del modello.

Generazione da pacchetti di attacco: Questa e la modalita opposta rispetto a quella precedentemente citata, dove si generano i dati sintetici a partire dai pacchetti di attacco. Come vediamo in figura 11, in questo caso i pacchetti OOD, sono separati dai dati normali ma vicini a quelli di attacco. Questa, osservando i grafici, sembra essere una buona soluzione di generazione.

Passiamo adesso all'addestramento del modello, in particolare abbiamo addestrato un modello di classificazione chiamato XGBoost, si poteva in realtà utilizzare un qualsiasi modello di classificazione perchè quello che ci interessa è vedere la differenza tra un modello addestrato con dati generati e un modello addestrato senza dati generati.

Abbiamo qua utilizzato tre modalità di addestramento per ogni dataset:

- Addestramento con dataset completo e dati sintetici
- Addestramento con pacchetti normali e dati sintetici
- Addestramento senza dati generati

Dove i dati sintetici possono essere generati secondo le tre modalita precedentemente descritte.

Quindi alla fine per ogni dataset si hanno 7 casi differenti di test.

Vediamo per il dataset semplice ADFANET:

Il test di riferimento è quello senza generazione di dati, come si vede si ottiene il 99.842.

Si nota come l'approccio peggiore è quello dove il modello è stato addestrato con i pacchetti normali e quelli generati (senza quindi i pacchetti di attacco originali).

L'approccio migliore invece è quello che utilizza il dataset completo insieme ai nostri pacchetti generati a partire da i pacchetti di attacco come era stato visto in precedenza dai grafici.

Vediamo ora il dataset complesso CICIDS:

In questo dataset invece non otteniamo miglioramenti, il caso peggiore rimane sempre quello ce non utilizza i pacchetti di attacco iniziali, ma l'addestramento tramite il dataset completo con i nostri pacchetti di attacco non risulta produttiva come si vede. Il caso migliore è quello senza generazione

Scenario differente lo si ha invece nel caso del dataset semplice dove si ottengono dei miglioramenti grazie alla generazione.

In conclusione, la tecnica di Data Augmentation utilizzata da SBO risulta essere produttiva solo nel caso di dati di partenza non complessi, nel caso in cui invece si abbia un dataset con molte caratteristiche, è necessario rivolgersi ad altre tecniche.