IRIS AperTO

UNIVERSITÀ DEGLI STUDI DI TORINO

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

Adversarial Learning e le sue applicazioni nella Cyber Security

This is a pre print version of the following article:

Original Citation:

Availability:

This version is available http://hdl.handle.net/2318/1740303 since 2020-06-03T18:37:39Z

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)

10 November 2026

Adversarial Learning e sue applicazioni nella Cyber Security

Francesco Bergadano

27 febbraio 2020

Dipartimento di Informatica, Università di Torino,
Corso Svizzera 185, 10149 Torino, Italy

francesco.bergadano@di.unito.it; Tel.: +39-011-6706743

1 Security e Machine Learning

Le tecnologie di Machine Learning sono state spesso applicate nel contesto della sicurezza informatica, con numerose aree di applicazione specifica che comprendono il riconoscimento di malware, il rilevamento di intrusioni, i filtri anti-spam, l'autenticazione utente (one-shot e continua), ed i sistemi di password checking.

In generale, l'idea è quella di vedere l'incidente di sicurezza come una anomalia rispetto al funzionamento atteso di un sistema, le cui caratteristiche posso essere apprese da esempi passati. Siamo quindi nel campo della "anomaly detection". Quando il classificatore, ottenuto attraverso tecniche adattive, segnala un'anomalia, scattano azioni correttive o controlli aggiuntivi, quali la generazione di allarmi in un SIEM, il blocco di traffico di rete sospetto, o la richiesta di un ulteriore e più forte livello di autenticazione.

2 Adversarial Learning

Dobbiamo tuttavia considerare un fatto molto importante: è possibile che un avversario sia a conoscenza dei nostri metodi di difesa e della nostra strategia adattativa, e provi quindi ad evitarla o ad interferire. Per esempio potrebbe

modificare i dati che vediamo, e impedirci di apprendere un classificatore, o indurci ad acquisirne uno errato. In altri casi, sarà interessato a conoscere l'output del nostro sistema di Machine Learning, in modo da preparare attacchi che non siano classificati come anomalie. Sulla base di queste osservazioni, negli ultimi anni ha ottenuto sempre maggiore attenzione un settore di ricerca e di applicazione noto come "Adversarial Machine Learning" [1].

Secondo Marco Barreno (Università di Berkeley, e ora Google, Inc.), è possibile classificare in tre macro-ategorie gli attacchi di tipo "adversarial" ad un sistema di sicurezza adattativo [2]:

- Causativo/Esplorativo. Un attacco *causativo* inserisce data manipolati, come esempi classificati erroneamente, o cambierà la frequenza con cui particolari tipologie di dati si presentano. Un attacco *esplorativo* cercherà soltanto di scoprire informazioni. Tipicamente sarà finalizzato a predire la descrizione di "anomalia" che risulterà come output del sistema di Machine Learning.

- Mirato/Indiscriminato. In un attacco *mirato* l'avversario sarà interessato al successo di uno specifico attacco, che gli sarà sufficiente per compromettere is sistemi target. In un attacco *indiscriminato* l'avversario si accontenterà di un qualsiasi successo, con il solo scopo di superare le barriere difensive del sistema.

- Integrità/Disponibilità. Un avversario che vuole mettere in pericolo l'*integrità* del sistema difensivo metterà in atto attacchi che non posso essere rilevati. Se invece ha come obiettivo la *disponibilità* del sistema, cercherà di renderlo instabile o non utilizzabile in pratica. Per esempio, causerà la generazione di un numero molto elevato di allarmi casuali, o realizzerà attacchi DoS per limitare il numero di esempi disponibili.

Nelle applicazioni di Cyber Security, in particolare nell'ambito della Anomaly Detection, il contesto è spesso (1) esplorativo, (2) mirato e (3) interessato all'integrità del sistema. Questo perché il nostro avversario cercherà normalmente di simulare il sistema di Machine Learning, e predire il suo output, con l'obiettivo di attuare specifici attacchi che non possano essere riconosciuti.

Per difendersi da questo tipo di attacchi "adversarial", un sistema di anomaly detection deve mantenere segrete alcune delle sue componenti. Questo viene fatto nascondendo alcuni parametri del processo di Machine Learning

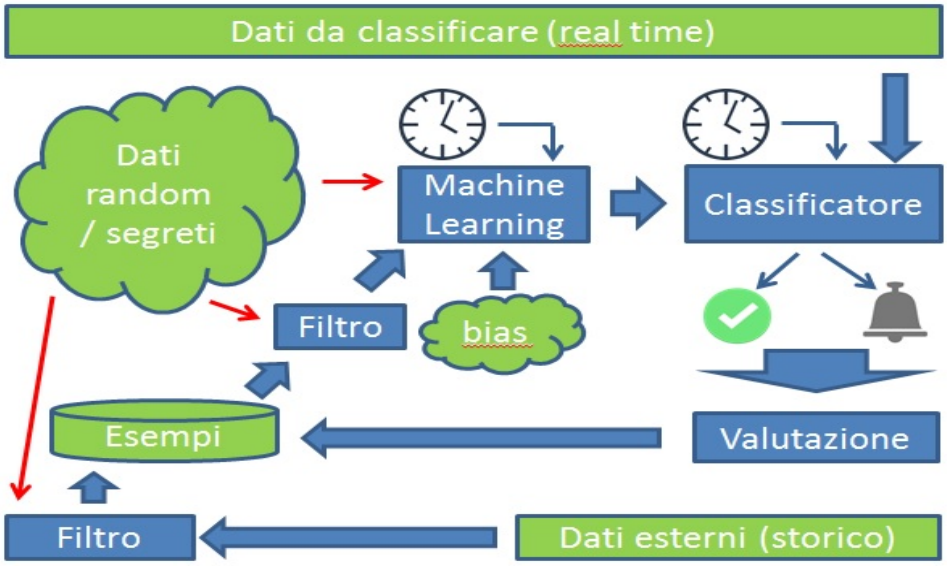

Figura 1: Anomaly detection con randomizzazione

[2, 3]. Per evitare però il rischio di un approccio del tipo *security through obscurity* (STO), sono stati proposti meccanismi di randomizzazione [4] e l'utilizzo di chiavi segrete [5, 6]. Un sistema di anomaly detection che sia "adversarial-aware", può essere rappresentato come in Fig. 1.

L'introduzione di dati random, non predicibili per l'avversario, rende impossibile per quest'ultimo la replicazione esatta del procedimento di Machine Learning, rendendo il suo attacco di tipo esplorativo molto meno efficace. In particolare, l'avversario, nel momento in cui dovrà mettere in pratica un nuovo attacco, non potrà sapere con esattezza se sarà classificato come anomalia, e quindi rilevato dai sistemi difensivi (ad esempio all'interno di un SOC, e con l'ausilio di software SIEM e intrusion detection/prevention che implementino questi principi).

3 Generative Adversarial Networks (GAN)

Consideriamo un sistema di Machine Learning supervisionato, nell'ambito della classificazione, ovvero dove gli esempi disponibili sono correttamente associati ad una classe. Ad esempio abbiamo degli esempi di malware e di applicazioni benigne, e per ognuno di questi esempi sappiamo appunto qual

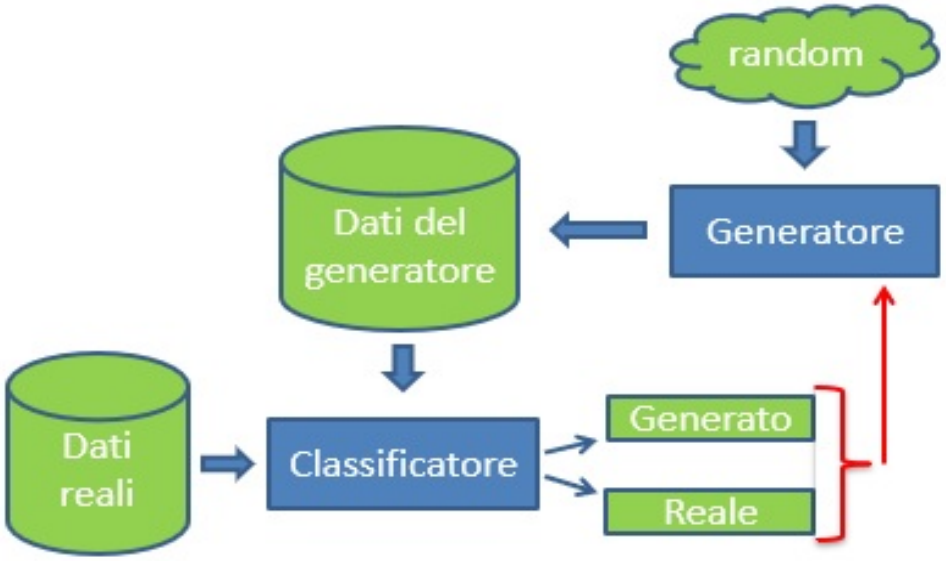

Figura 2: Generative Adversarial Network (GAN)

è la loro natura. Il sistema di Machine Learning produrrà un classificatore capace di etichettare esempi futuri, nell'esempio quindi con la conseguente possibilità di rilevare la presenza di malware.

Una "Generative Adversarial Network" (GAN) è un sistema capace di utilizzare il classificatore appreso come "black box" e di generare esempi che rispecchiano la distribuzione e la natura degli esempi utilizzati in fase di Machine Learning. Le GAN sono state proposte in [7] da Ian Goodfellow (ora alla Apple). Il funzionamento può essere sintetizzato come in Fig. 2:

1. il generatore inizialmente non sa nulla e genera esempi partendo da dati random;

2. il classificatore invece, utilizzando Machine Learning su dati reali, è in grado di distinguere i dati "fake" generati da quelli reali.

3. La decisione del classificatore però ritorna come feedback al generatore, il quale comincia ad apprendere come generare dati più credibili.

4. Dopo un sufficiente numero di iterazioni i dati generati saranno di fatto indistinguibili da quelli veri.

Per una descrizione dei principi matematici e degli algoritmi utilizzati in una GAN si faccia riferimento alla pubblicazione originaria di Goodfellow [7].

Secondo Yann LeCun, professore alla New York University e Chief Scientist per l'Intelligenza Artificiale di Facebook, i GAN sono "la più interessante idea nell'ambito del Machine Learning negli ultimi 10 anni" [8]. In effetti le GAN hanno avuto straordinario successo nelle applicazioni di image processing, e sono in grado di generare immagini artificiali perfettamente realistiche, compresa la controversa possibilità di generare situazioni e immagini false di persone conosciute o che si vogliano utilizzare come target.

Più recentemente è cresciuto l'interesse per l'utilizzo delle GAN nell'ambito della Cyber Security. Il collegamento è ovvio: se posso generare dati falsi, che però sembrano veri, potrò facilmente ottenere la cosiddetta "evasion" di un sistema di sicurezza, evitando la generazione di allarmi e riuscendo a battere i sistemi di autenticazione biometrica.

4 Applicazioni nella Cyber Security

Le tecnologie "adversarial" sono in primo luogo un pericolo per i sistemi difensivi, in quanto permettono di aggirare allarmi e strumenti di controllo di accesso che sono stati appresi in base ai dati disponibili. Di conseguenza esse diventano anche uno strumento difensivo, nel senso che servono a progettare sistemi di anomaly detection e di autenticazione che siano più robusti rispetto a questo tipo di attacchi. In particolare, le possibili aree di applicazione comprendono gli ambiti descritti sotto.

Network Intrusion Detection
Un avversario può aggirare un sistema di intrusion detection (evasion) se riesce ad ottenere un classificatore simile a quello del sistema IDS. A quel punto potrà attuare intrusioni che sa non essere rilevabili. Per difendersi da questa situazione sono stati realizzati sistemi di intrusion detection "adversarial", basati su chiavi segrete [5].

Autenticazione biometrica.
Le GAN rappresentano un gigantesco problema per i sistemi di autenticazione biometrica. Ad esempio uno dei migliori sistemi di Face Recognition, il Deepface di Facebook, con una accuratezza dichiarata del 97,35%, può essere facilmente ingannato con una GAN. Allo stesso modo i sistemi di riconosci-

mento utente basati ad esempio su impronta digitale, keystroke dynamics, o riconoscimento della voce non possono prescindere dall'utilizzo di tecniche adversarial, e devono impedire l'utilizzo del classificatore come black box, perché potrebbe essere utilizzato in modo malevolo per allenare il generatore di una GAN.

Rilevamento di Defacement.
Siccome il rilevamento di Web e Web Application Defacement è soggetto ad "evasion" mediante tecniche *adversarial*, sono stati proposti metodi per ridurre l'efficacia di questi attacchi, utilizzando randomization e chiavi segrete [3, 6].

Riconoscimento di Malware.
Il sistema MalGAN [9] descrive una GAN che è in grado di generare malware non rilevabile da un classificatore reso disponibile anche come *black box*. Di nuovo, anche in questo ambito, i sistemi di malware detection devono introdurre randomization o informazioni segrete nella fase di Machine Learning, in modo da evitare l'uso *adversarial* dei classificatori appresi.

Proactive password checking.
La verifica di password scelte dall'utente può essere basata su semplici regole, o, meglio, sull'appartenenza ad un dizionario di password. Ancor meglio, la password deve essere rifiutata se è "simile" a parole presenti in un dizionario dato [10]. Questo perché, ad esempio attraverso una GAN, è possibile generare un gran numero di password a partire da un dizionario, come dimostrato dal sistema PassGAN [11]. PassGAN, allenandosi sul dizionario "rockyou", ha dimostrato di poter riconoscere il 24,2% del password leak di Linkedin, che non ha relazione con il dataset di partenza.

Riferimenti bibliografici

[1] L. Huang, A. D. Joseph, B. Nelson, B. Rubinstein and J. D. Tygar, "Adversarial Machine Learning", ACM W. on AI and Security, 2011.

[2] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, J. D. Tygar. "Can Machine Learning be Secure?", Proc. ACM Symp. on Information, Computer and Communications Security (AsiaCCS), 2006.

[3] G. Davanzo, E. Medvet and A. Bartoli, "Anomaly detection techniques for a web defacement monitoring service", Expert Systems with Applications, vol. 38, number 10, 2011.

[4] S. Rota Bulò, B. Biggio, I. Pillai, M. Pelillo, F. Roli, "Randomized Prediction Games for Adversarial Machine Learning", IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 11, 2017.

[5] J. E. Tapiador, A. Orfila, A. Ribagorda, and B. Ramos, "Key-Recovery Attacks on KIDS, a Keyed Anomaly Detection System", IEEE Trans. On Dependable and Secure Computing, vol. 12, no. 3, 2015.

[6] F. Bergadano, F. Carretto, F. Cogno, D. Ragno, "Defacement Detection with Passive Adversaries", Algorithms, 12:8, n. 150, 2019.

[7] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, Y. Bengio, "Generative Adversarial Networks", ArXiv, abs/1406.2661, 2014

[8] Y. LeCun, "Recent and potentially upcoming breakthroughs in deep learning" https://www.quora.com/What-are-some-recent-and-potentially-upcoming-breakthroughs-in-deep-learning, 2016.

[9] W. Hu, Y. Tan, "Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN", arXiv:1702.05983, 2017.

[10] F. Bergadano, B Crispo, G Ruffo, "High dictionary compression for proactive password checking", ACM Trans. Inf. and System Security, 1998.

[11] B. Hitaj, P. Gasti, G. Ateniese, F. Perez-Cruz, "PassGAN: A Deep Learning Approach for Password Guessing", arXiv:1709.00440, 2017.