Rilevamento e interpretazione di anomalie a supporto della sicurezza informatica

F. Del Bonifro [1], L. Corradi [1] e D. Del Sorbo [1]

[1] *Lutech s.r.l., Cinisello Balsamo, Italia*

Abstract

L'apprendimento non supervisionato è un insieme di approcci molto potenti che permettono, tra le varie applicazioni, di gestire e analizzare dati non etichettati al fine di identificare eventi anomali rispetto ad un determinato fenomeno. In questo contributo viene presentata l'applicazione di alcuni algoritmi di questo tipo a dati relativi a tentativi di login al fine di identificare eventuali tentativi problematici e/o sospetti. Una seconda analisi descriverà invece un approccio basato sull' utilizzo dei valori di Shapley che permetta l'estrazione di alcune condizioni che hanno portato il modello a valutare un'osservazione come anomalia per garantire un certo livello di explainability delle scelte fatte dall'algoritmo.

Keywords

Unsupervised learning, anomaly detection, explainability

1. Introduzione

La continua crescita del mondo digitale e delle tecnologie sviluppate all'interno di questo consente che un importante aspetto della realtà odierna sia inevitabilmente rappresentato dalla cyber-security. Integrando sempre di più tool digitali nella nostra quotidianità la superficie di attacco e le potenziali vulnerabilità sono in continua crescita ed evoluzione. I cyber-attacchi possono avvenire a diversi livelli e generare diversi tipi di problemi che vanno dall'inaccessibilità di alcune risorse, alla corruzione di sistemi, alla perdita e divulgazione di informazioni riservate e molto altro.

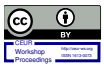
Altrettanto numerosi sono i possibili approcci di difesa da queste minacce. Le metodologie derivanti dall'apprendimento automatico sono anche in questo caso state studiate da diversi punti di vista come in altri ambiti Refs. [1, 2].

Una delle possibili strategie è rappresentata dall'intrusion detection a livello di tentativi di login. In questo contributo viene descritta l'implementazione di un algoritmo data-driven per l'analisi di un dataset contenente lo storico di diversi mesi relativi ai login verso un certo servizio online. Ogni tentativo di login è descritto tramite diverse feature ma non ha etichettatura di tipo anomalo/normale e viene sottoposto ad un primo step di pulizia e preprocessing per poi essere utilizzato da un modello di apprendimento non supervisionato che spesso viene utilizzato nei problemi di rilevazione di anomalie [3]: l'isolation forest [4].

Le anomalie rilevate vengono poi analizzate manualmente per valutare la qualità del modello e, ove possibile, la natura dell'anomalia. L'analisi dei risultati e l'explainability di molti modelli data-drive non è però sempre un compito banale, per questo motivo è stato affiancato al modello di anomaly detection un automatismo in grado di fornire per ogni anomalia, una lista di condizioni (feature-valore) che hanno influenzato maggiormente la decisione del modello.

Il concetto su cui si basa principalmente questo automatismo è quello dei valori Shapley [5] molto utilizzati nel campo dell'Explainable AI [6, 7] che forniscono un indice di importanza per le varie feature sia a livello globale (modello) che locale (singola osservazione). In base a questi valori si cerca l'insieme di condizioni minimo per cui una

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29-31, 2023, Pisa, Italy
EMAIL: f.delbonifro@lutech.it (A. 1); l.corradi@lutech.it (A. 2); d.delsorbo@lutech.it (A. 3)



anomalia sia stata identificata dal modello come tale, ovvero il minimo sottoinsieme delle feature (e relativi valori) che ha spinto il modello ad identificare un'osservazione come anomala.

Questo approccio *non* fornisce un insieme di regole rigide per definire un'anomalia ma è inteso unicamente a supporto della comprensibilità delle scelte fatte dal modello al fine di valutarne più agevolmente la qualità.

2. Dati e modello

Il dataset a disposizione è descritto da diverse feature le cui principali sono:

tipo di utenza
privilegio
IP sorgente
IP destinazione
successo/insuccesso connessione
istante temporale

La feature tipo di utenza si riferisce ad un raggruppamento significativo degli utenti trattati definibile a priori data da conoscenza di dominio.

Nonostante la natura dinamica del fenomeno, questo non viene rappresentato dal punto di vista di una serie storica, bensì come eventi singoli la cui informazione temporale viene codificata in una feature che rappresenta il momento della giornata in cui un'osservazione si è verificata e in alcune feature aggregate costruite in fase di preparazione del dataset in grado di tenere traccia di stagionalità e storico delle connessioni all'interno di una certa finestra temporale.

Ad esclusione di queste feature aggregate, tutte le altre feature utilizzate nell'analisi che segue sono state trattate come categoriche e per la vettorizzazione è stato utilizzato il metodo di one-hot encoding.

La suddivisione in trainset e testset corrisponde ad uno split 80/20% dell'intero dataset.

2.1. Modello

Il modello utilizzato per l'identificazione delle anomalie è l'isolation forest, un approccio molto utilizzato nei problemi di anomaly detection. L'obiettivo dell'algoritmo è quello di isolare le osservazioni presenti nel dataset partizionando lo spazio in cui queste osservazioni si trovano. Per fare questo, l'algoritmo sceglie casualmente delle feature e un valore all'interno del range di valori che queste possono prendere e suddivide le osservazioni rispetto al valore che questa feature assume rispetto al valore di split scelto. L'assunzione alla base di questo modello è che un evento anomalo sarà più *facile* da isolare rispetto al resto delle osservazioni, ovvero saranno necessarie meno iterazioni di questo processo di partizionamento per far sì che il punto venga isolato dagli altri punti del dataset.

Il processo di partizionamento viene rappresentato tramite degli alberi binari in cui ogni nodo rappresenta uno split del dataset. In questa rappresentazione la misura utilizzata per determinare se un punto sia o meno anomalo è la profondità nell'albero a cui si arriva per realizzare l'isolamento di quel punto. Più questa grandezza è piccola e meno iterazioni sono state necessarie per isolare il punto, per cui le anomalie sono caratterizzate da bassi valori di questa misura. Questi concetti vengono formalizzati in quella che costituisce la decision function dell'algoritmo e che viene valutata per ogni esempio analizzato dal modello. Questa funzione viene definite in maniera tale che tutte le osservazioni che la rendono negative sono considerate come punti anomaly del campione, tutti gli altri casi sono invece considerati come punti non anomali, la soglia decisionale è quindi impostata sullo zero di questa funzione.

3. Risultati

Il presente caso di studio richiede un approccio di tipo non supervisionato non avendo a disposizione alcun tipo di etichettatura per i dati in esame. Questo non permette di utilizzare le classiche metriche utilizzate nei casi supervisionati per la valutazione della bontà del modello.

Esistono alcune metriche per i modelli di apprendimento non supervisionato, ad esempio basati su varianza intra- ed inter-cluster che sono state utilizzate per la selezione degli iperparametri ma, per la valutazione finale del nostro approccio è stato scelto di considerare il numero di anomalie trovate (rispetto al numero di osservazioni totali) e l'analisi manuale di queste ultime avendo a disposizione della conoscenza di dominio.

Nel modello che è stato selezionato come migliore in base a queste considerazioni di dominio la percentuale di anomalie rilevate risulta essere dell'ordine del $10^{-3}\%$, un valore che

risulta essere accettabile dall'utilizzatore finale. Nel prossimo paragrafo illustreremo un automatismo mirato a migliorare l'explainability del modello scelto per capire se questo stia considerando degli aspetti ritenuti rilevanti dagli esperti o meno.

4. Explainability

Come accennato nel precedente paragrafo, nonostante l'esistenza di alcune misure in grado di fornire informazioni sulla qualità dell'approccio scelto, si è scelto di analizzare manualmente le anomalie identificate dal modello. Per supportare questa analisi manuale ci si è orientati ad approcci utilizzati nel campo dell'Explainable AI, ovvero quell'area che si occupa di sviluppare dei metodi in grado di fornire una certa misura di spiegazione delle scelte fatte da un modello che, nel caso di molti algoritmi di apprendimento automatico, risultano essere delle black box.

All' interno di questo ambito un metodo molto utilizzato è quello basato sui valori di Shapley, ereditato dalla teoria dei giochi e adattato al mondo dell'apprendimento automatico per calcolare il contributo marginale di ogni feature alla predizione fatta dal modello in esame.

Per ogni feature, la misura della sua rilevanza nella scelta fatta dal modello è data dai valori di Shapley che vengono calcolati separatamente per ogni esempio del dataset. Questo permette una spiegazione a livello locale delle scelte prese dal modello. Una feature che ha giocato un ruolo molto importante nella decisione presa dal modello per un certo esempio è caratterizzata da valori assoluti alti dei valori di Shapley. Al contrario, una feature che non ha influito in maniera particolarmente importante nella scelta fatta dal modello per quel determinato esempio assumerà un valore di Shapley prossimo allo zero.

A posteriori, mediando i valori assoluti dei valori di Shapley su tutto il campione e separatamente per ogni feature, è possibile ottenere una stima dell'importanza delle feature a livello globale.

Una proprietà molto importante dei valori di Shapley è la seguente: per ogni osservazione, la somma dei valori di Shapley di tutte le feature in gioco rappresentano la discrepanza tra il valore che il modello ha predetto per quell'osservazione e il valore predetto mediamente sull'intero campione.

L'isolation forest basa le proprie decisioni sul valore preso dalla decision function, in particolare se questa è valutata ad un valore negativo l'osservazione viene considerate un'anomalia, nel resto dei casi è considerata non anomala. Nel caso di anomalie, sono dunque le feature con valori di Shapley più negativi ad aver contribuito maggiormente alla decisione del modello, ovvero ad aver contribuito maggiormente a spostare il valore della decision function al di sotto dello zero rispetto al valore medio delle predizioni, un valore positivo chiamato *base value*.

Utilizzando i valori di Shapley, è stata definita una procedura automatica per estrarre le feature di maggiore rilevanza nel caso delle osservazioni che sono state identificate come anomalie per ottenere un insieme ristretto di feature da analizzare per comprendere la decisione fatta dal modello. Il funzionamento di tale automatismo è illustrato dei seguenti punti che vengono ripetuti per ogni anomalia identificata dall'isolation forest:

1) Si sommano tutti i valori di Shapley positivi al base value in modo tale da raggiungere il massimo valore (positivo) della decision function per quella osservazione. Tutte le altre feature, avendo valori di Shapley negativi, tenderanno ad abbassare tale valore fino a farlo scendere al di sotto dello zero
2) Si ordinano i valori di Shapley negativi restanti e si somma al valore calcolato nel punto 1) il valore di Shapley più negativo (ovvero corrispondente alla feature più rilevante nella identificazione dell'anomalia) e memorizziamo la feature a questo corrispondente
3) Si controlla che il valore della somma sia negativo, se questo non è il caso si ripete il procedimento del punto 2) altrimenti, essendo arrivati al di sotto della soglia decisionale, abbiamo selezionato il minimo sottoinsieme di feature che hanno fatto sì che tale osservazione fosse considerata anomala dal nostro modello.

Questo procedimento ci permette quindi di estrarre tutte le più importanti feature necessarie a considerare tale punto come anomalo e focalizzarci su queste e i valori da loro assunti per dare una interpretazione ai risultati ottenuti dal modello di anomaly detection.

Considerando le diverse anomalie rilevate e le feature rilevanti ottenute tramite la metodologia appena illustrata, è stato possibile identificare delle diverse caratteristiche comuni per alcune anomalie, suggerendo sia un sottoinsieme di feature di interesse per le analisi future che la possibilità per l'algoritmo di identificazione di diversi tipi di anomalie.

5. Conclusioni

Il caso d'uso illustrato in questo contributo è un esempio di anomaly detection nel campo della sicurezza informatica che ha il fine di supportare ed affiancare altre metodologie per la gestione della sicurezza a livello di accessi ad un determinato servizio online. Il modello implementato è un algoritmo di isolation forest, spesso utilizzato in contesti simili.

A differenza di quanto viene solitamente fatto per la valutazione di modelli di apprendimento automatico supervisionati e non, si è scelto di affrontare la determinazione della qualità delle anomalie identificate in maniera manuale. Per fare questo si è implementato un automatismo basato sui valori di Shapley che sfrutta alcune delle loro proprietà per poter estrarre dall'insieme di tutte le feature un sottoinsieme costituito dalle più rilevanti ai fini della identificazione come anomalia degli esempi rilevati come tali. Questo approccio è stato ideato per poter analizzare ed interpretare le scelte fatte dall'algoritmo in maniera più puntuale e focalizzata su quelli che il nostro algoritmo ritiene più importanti piuttosto che indagare l'intero set di feature che avrebbe portato una gestione poco efficiente e in alcuni casi impraticabile delle informazioni contenute in una rappresentazione ad alta dimensionalità. Infatti, soprattutto a causa della trasformazione dovuta al metodo di one-hot encoding, il numero delle feature è diventato dell'ordine di 10^2 che, considerando anche tutte le possibili combinazioni tra feature, porta a una dimensionalità di difficile gestione.

Questo approccio non è da considerarsi come un metodo per l'estrazione di regole rigide a partire dalle predizioni dell'algoritmo ma un supporto alla spiegazione delle scelte fatte dal modello che altrimenti sarebbe stata difficilmente affrontabile.

6. Riferimenti

[1] Dylan Chou and Meng Jiang. 2021. A Survey on Data-driven Network Intrusion Detection. ACM Comput. Surv. 54, 9, Article 182 (December 2022), 36 pages. https://doi.org/10.1145/3472753

[2] A. B. Nassif, M. A. Talib, Q. Nasir and F. M. Dakalbab, "Machine Learning for Anomaly Detection: A Systematic Review," in *IEEE Access*, vol. 9, pp. 78658-78700, 2021, doi: 10.1109/ACCESS.2021.3083060

[3] Verkerken, M., D'hooge, L., Wauters, T. *et al.* Towards Model Generalization for Intrusion Detection: Unsupervised Machine Learning Techniques. *J Netw Syst Manage* **30**, 12 (2022). https://doi.org/10.1007/s10922-021-09615-7

[4] F. T. Liu, K. M. Ting and Z. -H. Zhou, "Isolation Forest," *2008 Eighth IEEE International Conference on Data Mining*, Pisa, Italy, 2008, pp. 413-422, doi: 10.1109/ICDM.2008.17

[5] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 4768–4777

[6] Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). christophm.github.io/interpretable-ml-book/

[7] Kamath, U., Liu, J. (2021) Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning, Springer International Publishing