The 9th International Conference on Web Research
ICWR2023
May 3th-4th, 2023 ;Tehran , Iran

ENIXMA: ENsemble of EXplainable Methods for Detecting Network Attacks

Seyed Mojtaba Abtahi*, Hossein Rahmani, Milad Allahgholi, Sajjad Alizade

School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
Mojtaba_abtahi@comp.iust.ac.ir, h_rahmani@iust.ac.ir, Milad_allahgholi@comp.iust.ac.ir,
sajjadalizadeh275800@gmail.com

Abstract

Today, the Internet is a major part of society. Given the ubiquity of the Internet, its availability is a must. Attackers, on the other hand, seek to make Internet services inaccessible and exploit Internet service companies. Attackers use various tools and methods to attack the networks and infrastructure of service companies. These attacks are also called network traffic anomalies. In general, malfunctions or attacks are network events that deviate from normal expected behavior and are suspicious of security. In general, anomalies or attacks are network events that deviate from expected normal behavior and are suspicious from a security point of view. Many different methods have been proposed to detect attacks in the network. One of the most important challenges of the previous methods is the low accuracy and lack of interpretability. In this paper, we tried to use a combination of basic methods to detect attacks and achieve 89% attack detection accuracy in the balanced dataset. This accuracy has increased by 3% compared to previous works. In order to solve the challenge of interpretability, we applied SHAP, LIME and decision tree methods and identified the effective features in detecting attacks. The proposed method, in addition to high accuracy and interpretability, has a higher speed than previous works.

Keywords: Anomaly detection, machine learning, network data, botnet, data mining, ensemble learning, interpretability.

The 9th International Conference on Web Research
ICWR2023
May 3th-4th, 2023 ; Tehran , Iran

ترکیبی از روش های تفسیر پذیر برای تشخیص حملات شبکه

سیدمجتبی ابطحی*، دکتر حسین رحمانی، میلاد الهقلی، سجاد علیزاده

دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ، تهران ایران
Mojtaba_abtahi@comp.iust.ac.ir, h_rahmani@iust.ac.ir, Milad_allahgholi@comp.iust.ac.ir,
sajjadalizadeh275800@gmail.com

چکیده

امروزه اینترنت یکی از قسمت‌های اصلی جامعه را تشکیل می‌دهد. با توجه به فراگیر بودن اینترنت، در دسترس بودن آن یک امر ضروری به شمار می‌رود. از طرفی مهاجمان به دنبال از دسترس خارج کردن خدمات اینترنتی و سوءاستفاده از شرکت‌های خدمات اینترنتی هستند. مهاجمان از ابزارها و روش‌های مختلف جهت حمله به شبکه‌ها و زیرساخت‌های شرکت‌های ارائه کننده خدمات استفاده می‌کنند. به آن حملات، ناهنجاری در ترافیک شبکه نیز گفته می‌شود. به طورکلی، ناهنجارها یا حملات، رویدادهای شبکه هستند که از رفتار عادی مورد انتظار، منحرف می‌شوند و از نظر امنیتی مشکوک هستند. روش‌های بسیار متنوعی برای شناسایی حملات در شبکه ارائه شده اند. از مهم‌ترین چالش‌های روش‌های پیشین می‌توان به دقت پایین و عدم تفسیر پذیری[1] اشاره نمود. در این مقاله، ما سعی نمودیم که ترکیبی از روش‌های پایه را برای شناسایی حملات به کار گیریم و دقت شناسایی حملات را در مجموعه‌داده متوازن شده به ۸۹ درصد برسانیم. این دقت در مقایسه با کارهای پیشین ۳ درصد رشد داشته است. به منظور حل چالش تفسیر پذیری، روش‌های SHAP، LIME و درخت تصمیم را اعمال نموده و ویژگی‌های اثرگذار در شناسایی حملات را شناسایی نمودیم. روش پیشنهادی، علاوه بر دقت و تفسیر پذیری بالا، سرعت بالاتری نسبت به روش‌های پیشین دارد.

کلمات کلیدی

تشخیص ناهنجاری، یادگیری‌ماشین، داده‌های شبکه، بات‌نت، داده‌کاوی، یادگیری‌گروهی، تفسیرپذیری

۱- مقدمه

امروزه اینترنت یکی از قسمت‌های اصلی جامعه را تشکیل می‌دهد. با توجه به فراگیر بودن اینترنت، در دسترس بودن آن یک امر ضروری به شمار می‌رود. از طرفی مهاجمان به دنبال از دسترس خارج کردن خدمات اینترنتی و سوءاستفاده از شرکت‌های خدمات اینترنتی هستند [۷]. یکی از متداول‌ترین حملاتی که به این شرکت‌ها صورت می‌گیرد حملات DDoS است که باعث اختلال در ارائه خدمات شرکت‌ها می‌شود. اختلال و قطعی سرویس ضررهای زیادی به شرکت وارد می‌کند تا جایی که ۲۴ ساعت قطعی سرویس در یک شرکت بزرگ تجارت الکترونیک می‌تواند میلیون‌ها دلار ضرر به شرکت برساند [۱].

ترافیک جریان‌های شبکه را می‌توان به دو دسته ترافیک نرمال و ترافیک حمله DDoS تقسیم نمود و براساس ویژگی‌های ترافیک می‌توان متوجه شد که چه زمان به شبکه قربانی حمله صورت می‌گیرد [۶]. حملات DDoS معمولاً مبتنی بر حجم هستند و برای تشخیص این نوع حملات روش جریان‌محور[2] مناسب است. جریان‌ها به عنوان مجموعه‌ای از بسته‌های IP هستند، که در یک بازه زمانی مشخص از یک نقطه مشخص در شبکه عبور می‌کنند. به این ترتیب که بسته‌های متعلق به یک جریان خاص خصوصیات مشترکی دارند. برای تشخیص حملات سه مرحله وجود دارد: در مرحله اول صادرکنندگان جریان بسته‌های خام را دریافت کرده و آن‌ها را جمع‌آوری می‌کنند. در مرحله بعد جمع‌کننده‌های جریان داده‌های جریان را ذخیره و سپس پیش‌پردازش می‌کنند. سرانجام در مرحله آخر، برنامه‌های تجزیه و تحلیل، مانند سیستم‌های تشخیص نفوذ (IDS) ، داده‌های جریان را بازیابی و تحلیل می‌کنند [۲]، [۳].

الگوریتم‌های زیادی در سیستم‌های تشخیص نفوذ مانند KNN[3]، SVM[4]، RF[5] استفاده می‌شود که براساس ویژگی‌هایی که از داده‌های ورودی دریافت می‌کنند، تصمیم‌گیری انجام می‌دهد [۴]، [۵]. با توجه به این که در داده‌های مربوط به شبکه تفسیر پذیری داده‌های اهمیت زیادی در تشخیص بهتر حملات دارد، به همین منظور نیاز اساسی برای تفسیرپذیری در مسائل مرتبط با یادگیری ماشین در تشخیص حملات حس شد [۸].

در بخش دوم به مرور کارهای انجام شده در زمینه تشخیص ناهنجاری در داده‌های شبکه با رویکردهای متفاوت و مرور کارهای انجام شده در زمینه تفسیرپذیری و الگوریتم‌های مهم برای تفسیرپذیری خواهیم پرداخت. در بخش سوم به توضیح قسمت‌های مختلف روش پیشنهادی و خروجی الگوریتم‌های تفسیرپذیری خواهیم پرداخت. در بخش آخر به ارزیابی و نتایج روش پیشنهادی و علت برتری مدل پیشنهادی نسبت به کارهای پیشین خواهیم پرداخت.

۲- کارهای پیشین

در این بخش اول به برخی از کارهای انجام شده در زمینه تشخیص ناهنجاری در داده‌های شبکه با استفاده از الگوریتم‌های یادگیری ماشین و معیارهای شباهت خواهیم پرداخت و در بخش دوم به مرور کارهای مرتبط با تفسیر پذیری می‌پردازیم.

۲-۱- تشخیص ناهنجاری در داده‌های شبکه

پژوهش‌های انجام‌شده در این زمینه از دو جنبه بررسی می‌شود ۱- تجزیه و تحلیل آماری، ۲- یادگیری ماشین. شمای کلی از معیارها و روش‌های گفته‌شده در این پژوهش را در شکل (۱) مشاهده می‌کنید.

۲-۱-۱- تجزیه و تحلیل آماری

در این بخش به انواع راه‌حل‌ها برای تشخیص ناهنجاری‌ها در شبکه می‌پردازیم. روش‌های تشخیص مبتنی بر ناهنجاری را می‌توان به دو دسته‌ی تجزیه و تحلیل آماری و یادگیری ماشین تقسیم کرد. روش‌های مبتنی بر تجزیه و تحلیل آماری به قدرت محاسباتی نسبتا کم‌تری نیاز دارد. اگرچه این رویکرد نرخ تشخیص سریع و قابل قبولی را ارائه می‌دهد اما مشکل اصلی آن نرخ مثبت کاذب است.

گیرما و همکاران [۱۱] یک مدل آماری ترکیبی را پیشنهاد می‌کنند که می‌تواند به طور قابل توجهی این حملات را کاهش دهد و می‌تواند راه‌حل جایگزین بهتری برای مشکلات تشخیص فعلی باشد. این طرح ترکیبی براساس ماتریس‌های آنتروپی[۶] و کوواریانس[۷] است. از مزایای این روش می‌توان به دقت بالا و وابسته نبودن به هیچ فرضیه‌ای در بسته‌های شبکه اشاره کرد. از معایب این روش نیز می‌توان به کم شدن تمرکز این طرح در زمان تجمیع محاسبات اشاره کرد.

رودولف و همکاران [۱۲] از یک روش تجزیه و تحلیل آماری که برروی ترافیک چندین لایه شبکه صورت می‌گیرد، استفاده می‌کنند. در این مقاله از دو روش استفاده می‌شود. هر دو روش از مقدار آستانه برای تشخیص ناهنجاری در شبکه استفاده می‌کنند. در روش اول (فرمول ۱) Ssprt[t] متغیر آزمون است. P0(Npk[t]) و P1(Npk[t]) توابع چگالی احتمال قبل و بعد حمله هستند. و اگر Ssprt[t] بزرگ‌تر از آستانه tr باشد احتمال حمله وجود دارد.

$$S_{\text{sprt}}[t] = max\left\{0, \left[S_{\text{sprt}}[t-1] + log(\frac{P_1(N_{\text{pk}}[t])}{P_0(N_{\text{pk}}[t])})\right]\right\}; S[0] = 0 \qquad (۱)$$

در روش دوم (فرمول ۲) اختلاف بین میانگین فعلی و میانگین بلند مدت محاسبه می‌شود. (Npk[t]، میانگین ترافیک در زمان t و m[t] میانگین بلند مدت تا زمان t می‌باشد) وقتی ضریب cusum از آستانه فراتر رود احتمال حمله DDoS وجود دارد.

$$S[t] = max\left\{0, (S[t-1] + N_{\text{pk}}[t] - m[t])\right\}; S(0) = 0 \qquad (۲)$$

سه ویژگی جذاب در این رویکرها وجود دارد. اول، هر دو روش خودآموز هستند و این امر باعث می‌شود که با شبکه و الگوهای آن سازگار شوند. دوم، این روش‌ها میانگین شناسایی حملات را کاهش می‌دهند. سوم، از نظر محاسباتی ساده هستند و از این رو می‌توانند به صورت خطی پیاده‌سازی شوند. چهارم، میزان تشخیص اشتباه یا نرخ مثبت کاذب کم[۸] است [۱۳].

۲-۱-۲- یادگیری ماشین

دسته بعدی سیستم‌های مبتنی بر یادگیری ماشین می‌باشد که از تکنیک‌های داده‌کاوی برای کشف الگوریتم‌های ناشناخته در حجم زیاد استفاده می‌کند. سیستم یادگیری ماشین از رویکردهای مختلفی در سیستم‌های خود استفاده می‌کند که می‌توان به رویکرهای بانظارت، نیمه‌نظارت و بدون‌نظارت اشاره کرد. در رویکرد بانظارت، داده آموزشی باید به عنوان موارد حمله و غیر حمله برچسب‌گذاری شود که این کار خیلی وقت‌گیر است و ممکن است با خطای ناخواسته برخورد کنیم. در رویکرد نیمه‌نظارتی، مجموعه‌داده‌های آموزشی نیازی به برچسب‌گذاری کامل ندارد. اگرچه پیچیدگی برچسب‌زدن را کاهش می‌دهد، اما ابهام مدل ارائه‌دهنده ترافیک سیستم یا شبکه را افزایش می‌دهد. رویکرد بدون نظارت نیازی به برچسب ندارند. این سیستم‌ها الگوها و رفتارهای مشابه را خوشه‌بندی می‌کنند.

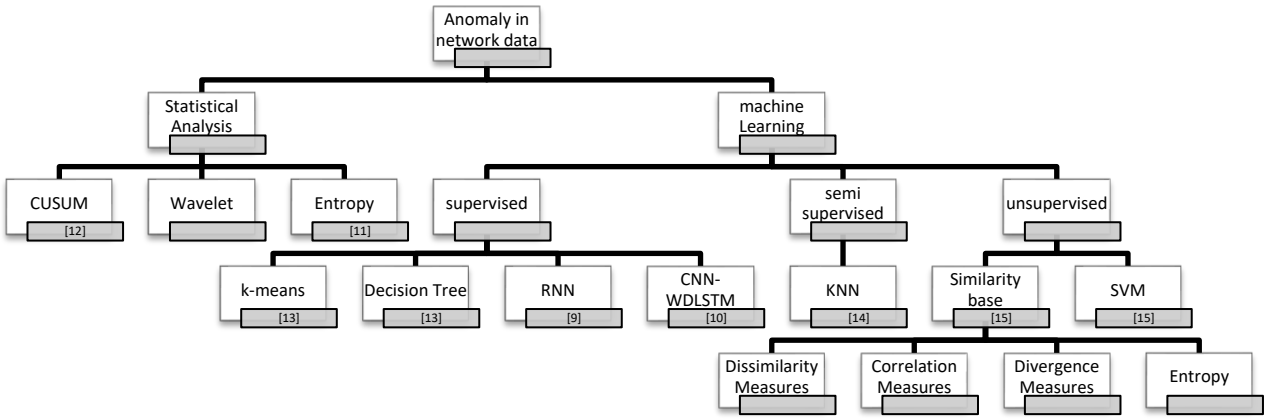
چوآن‌لانگ و همکاران [۹] از شبکه‌های عصبی بازگشتی برای تشخیص ناهنجاری‌ها استفاده می‌کنند. آن‌ها در متدولوژی خود از دو روش انتشار رو به جلو و انتشار رو به عقب استفاده می‌کنند.

آزمایشات آن‌ها برروی مجموعه‌داده NSL-KDD[۹] انجام گرفته‌است. رده‌بندی براساس نرمال بودن یا نبودن حملات می‌باشد. آن‌ها در آزمایشات خود ویژگی‌ها را از ۴۱ ویژگی به ۱۲۲ ویژگی افزایش دادند، بنابراین مدل RNN-IDS دارای ۱۲۲ گره ورودی و ۲ گره خروجی در آزمایشات رده‌بندی دودویی است. تعداد دوره‌ها[۹] نیز ۱۰۰ دور می‌باشد. آزمایش‌ها را با تعداد گره‌های پنهان شده ۲۰، ۶۰، ۸۰، ۱۲۰، ۲۴۰ و نرخ یادگیری ۰/۰۱، ۰/۱، ۰/۵ انجام دادند و بالاترین دقت برای تعداد گره‌های پنهان شده ۸۰ و نرخ یادگیری ۰/۱ است.

آبهیجیت داس و همکاران [۱۰] از یک راهکار ترکیبی استفاده کردند که بر پایه سه مدل Balanced Bagging، XGBoost، RF-HDDT استفاده می‌کنند. پارامترهای Balanced Bagging و XGBoost برای داده‌های نامتعادل تنظیم شده‌اند و معیار هلینگر مکمل جنگل تصادفی می‌شود تا محدودیت‌های معیار فاصله پیش‌فرض را برطرف کند. دو الگوریتم جدید برای رسیدگی به مسئله همپوشانی کلاس در مجموعه‌داده پیشنهاد می‌کنند و در طول آموزش اعمال می‌کنند. این دو الگوریتم برای کمک به بهبود عملکرد مجموعه داده آزمایشی با تأثیرگذاری بر تصمیم رده‌بند نهایی که توسط سه رده‌بند پایه به عنوان بخشی از طبقه‌بندی گروه، که از ترکیب کننده اکثریت رأی استفاده می‌کند، استفاده می‌شوند.

طرح پیشنهادی آن‌ها با حاشیه قابل توجهی برای موارد طبقه‌بندی دودویی و چند دسته‌ای، از طرح‌های گزارش‌شده بهتر عمل می‌کند.

هوآ–وو و همکاران [۱۴] ابتدا معماری DDoS را بررسی می‌کنند و جزئیات مراحل را به‌دست می‌آورند. سپس رویه‌های حملات DDoS را بررسی می‌کنند و متغیرها را براساس این ویژگی‌ها انتخاب می‌کنند. در نهایت از روش



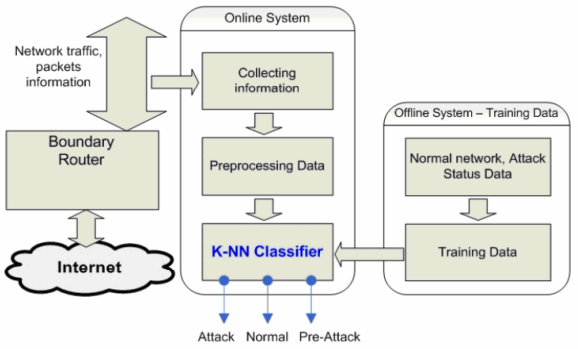
شکل ۱ . شمای کلی از دسته‌بندی کارهای انجام شده در زمینه تشخیص ناهنجاری در داده‌های شبکه

K نزدیک‌ترین همسایه برای رده‌بندی وضعیت شبکه در هر مرحله از حمله DDoS استفاده می‌کنند. همان‌گونه که در شکل (۲) مشاهده می‌کنید، روند انجام کار به این صورت است که پس از آموزش الگوریتم K نزدیک‌ترین براساس ۹ ویژگی انتخاب شده داده‌ها به صورت برخط[۱۰] جمع‌آوری می‌شود، سپس پیش‌پردازش شده و در مرحله پایانی به سه کلاس ترافیک نرمال، ترافیک حمله و ترافیک پیش از حمله تبدیل می‌شود. پس از انجام آزمایش بر روی این الگوریتم دقت این الگوریتم برروی مجموعه‌داده DARPA ۲۰۰۰، ۹۱ درصد است.

لازارویچ و همکاران [۱۵] به مقایسه تکنیک‌های تشخیص ناهنجاری در الگوریتم‌های بدون نظارت پرداختند. آن‌ها در کار خود به طرح‌های مختلف برای تشخیص داده‌های پرت برای تشخیص ناهنجاری خود اقدام کردند. اکثر الگوریتم‌های تشخیص ناهنجاری برای آموزش مدل، به مجموعه‌ای از داده‌های کاملا عادی نیاز دارد و به طور ضمنی فرض می‌کنند که ناهنجاری‌ها را می‌توان به عنوان الگوهایی که قبلا مشاهده نشده‌است شناسایی کرد. از آن‌جا که ممکن است یک داده پرت برروی اندازه‌گیری‌ها و مدل‌سازی‌ها تاثیر بگذارد باید طرح‌های مختلفی برای استخراج این داده‌ها در نظر بگیریم تا بفهمیم کدام یک به طور موثر کار می‌کند [۱۶].

۲-۲- تفسیرپذیری

از جمله روش‌های مهم در تفسیرپذیری روش LIME [11] و SHAP[12] هستند که بسیار در کارهای مختلف استفاده شده‌اند [۱۷]–[۲۰]. این روش‌ها دارای تفسیرپذیری غیرذاتی، محلی و مستقل از مدل هستند. ریزی و همکاران [۱۸] در پژوهشی از روش LIME برای بررسی عملکرد پیش‌بینی مدل LSTM پرداختند. آن‌ها به استخراج ویژگی‌های مهم در نمونه‌های با پیش‌بینی اشتباه پرداخته و نهایتا با تغییر اثر پارامترهای منفی، دقت مدل خود را افزایش دادند. سیندگاتا و همکاران [۱۹] در پژوهشی مشابه به بررسی ویژگی‌های مهم در خروجی مدل XGBoost پرداختند. آن‌ها نتایج خود را بر روی مجموعه‌داده لاگ کاربران اجرا نمودند. اما نکته اساسی که در همه مقالات مشابه وجود دارد

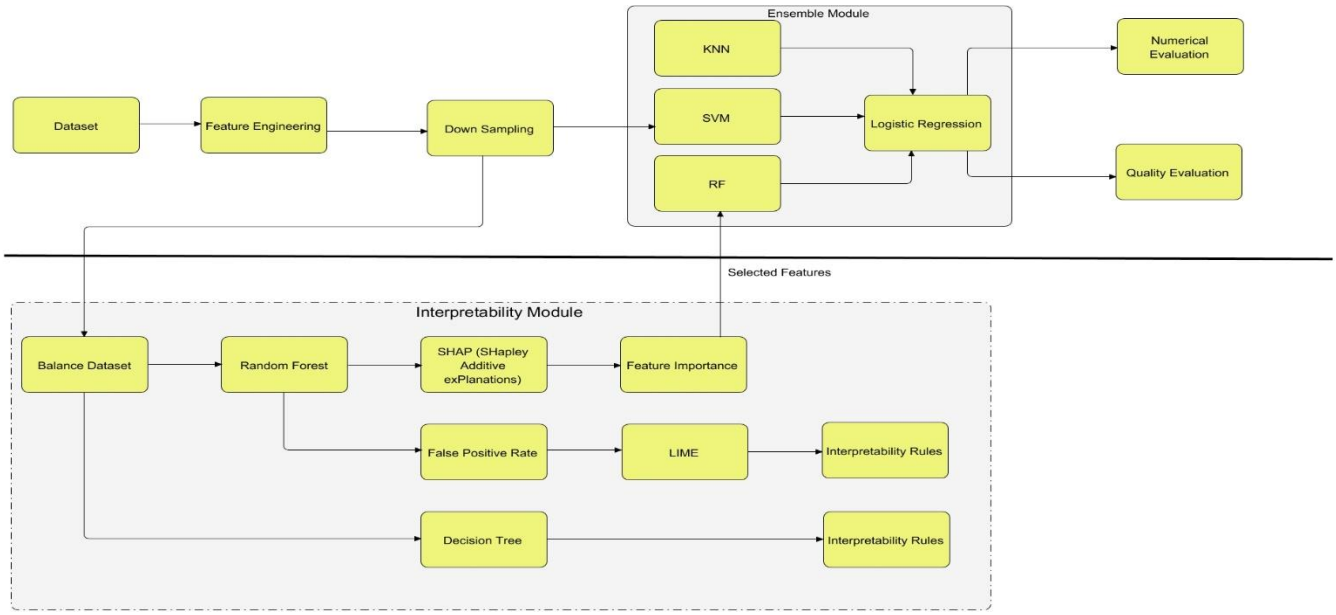


شکل ۲ . شمای کلی روند رده‌بندی با الگوریتم k نزدیک‌ترین همسایه[۱۳]

این است که ارزیابی برای تفسیرپذیری وجود ندارد. مارسیلو و همکاران [۲۲] از تفسیرپذیری برای استخراج ویژگی استفاده کرده‌اند. آن‌ها از روش SHAP استفاده کرده و براساس امتیازی که این روش برای هر ویژگی درنظر می‌گیرد، به انتخاب ویژگی‌های مهم پرداختند. سپس پس از انتخاب ویژگی، مسئله طبقه‌بندی را اجرا نمودند. در این پژوهش نتایج SHAP با سایر روش‌های موجود انتخاب ویژگی مانند ANOVA[14] مقایسه شده است. نتایج نشان از این دارد که SHAP در انتخاب ویژگی عملکرد بهتری را نسبت به سایر روش‌ها دارد.

۳- روش پیشنهادی

در این بخش به بررسی روش پیشنهادی ما در این مقاله می‌پردازیم. همان طور که در شکل (۳) مشاهده می‌کنید ما از یک روش ترکیبی برای تشخیص نوع حملات استفاده کردیم و علاوه بر آن با استفاده از الگوریتم‌های تفسیر پذیری برای تشخیص ویژگی‌های مهم، ویژگی‌های متمایز کننده برای تشخیص نوع حمله و استخراج قوانین مهم برای تشخیص نوع حمله استفاده می‌کنیم.



شکل ۳ مراحل مختلف فرآیند تفسیر پذیری در روش پیشنهادی

۳-۱- مجموعه داده

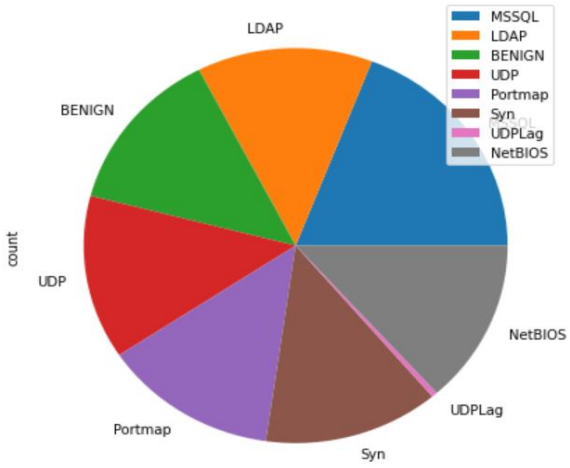
در این مقاله ما از مجموعه داده CICDDoS2019 استفاده می‌کنیم. این مجموعه داده شامل ۷ نوع حمله DDoS با نام‌های LDAP، MSSQL، NetBIOS، Portmap، Syn، UDPLag، UDP در قالب csv است. همچنین این مجموعه داده شامل ۸۸ ویژگی می‌باشد. این مجموعه داده شامل ۲۰,۳۶۴,۵۳۲ رکورد داده هست [۲۱].

۳-۲- پیش پردازش

ابتدا مقادیر گمشده مجموعه داده حذف می‌کنیم، سپس ویژگی‌هایی که در همه داده‌ها یک مقدار مشخص دارند، را حذف می‌کنیم بعد از آن مقادیر inf^{15} را به دلیل این‌که باعث درست آموزش ندیدن مدل می‌شود با مقدار بیشینه آن ویژگی جایگزین می‌کنیم. در مرحله بعد با استفاده از الگوریتم z-score داده‌های پرت مجموعه داده را حذف می‌کنیم و در آخر برای آموزش بهتر مدل داده‌ها را نرمال‌سازی[۱۶] می‌کنیم.

۳-۳- متوازن کردن مجموعه داده

به دلیل نامتوازن بودن داده‌ها و بالا بودن نسبت داده‌ها با برچسب حمله نسبت به داده‌ها با برچسب غیرحمله از داده‌های موجود نمونه‌گیری شد و توزیع برچسب‌ها به صورت شکل (۴) شد. به دلیل پایین بودن تعداد داده‌ها با برچسب UDPLag بعد از متوازن کردن مجموعه داده نیز باز هم تعداد این داده‌ها به نسبت بقیه برچسب‌ها پایین‌تر است.



شکل ۳. توزیع برچسب‌ها بعد از متوازن کردن مجموعه‌داده

۳-۴- پودمان تفسیر پذیری

ما در این قسمت از سه الگوریتم SHAP، LIME و درخت تصمیم برای تفسیرپذیری بهتر تشخیص حملات و تشخیص ویژگی‌های مهم برای تشخیص حملات استفاده کردیم.

۳-۴-۱- الگوریتم SHAP

الگوریتم SHAP یک روش تفسیر مدل است که برای تحلیل تاثیر ویژگی‌ها بر پیش بینی‌های یک مدل استفاده می‌شود. این الگوریتم بر اساس تئوری بازی‌ها و تخصیص بهره (cooperative game theory and Shapley value) ساخته شده است [۱۷].

در این الگوریتم، برای هر نمونه از داده، تاثیر هر ویژگی بر پیش بینی خروجی مدل محاسبه می‌شود. برای محاسبه این تاثیر، ابتدا برای هر ترکیب از ویژگی‌ها، Shapley value محاسبه می‌شود. سپس با استفاده از این مقادیر، تاثیر هر ویژگی بر پیش بینی مدل محاسبه می‌شود [۲۲].

در روش ما ابتدا مجموعه داده متوازن شده را به الگوریتم جنگل تصادفی می‌دهیم سپس با استفاده از الگوریتم SHAP ویژگی‌های مهم آن را استخراج می‌کنیم.

۲-۴-۳- الگوریتم LIME

الگوریتم LIME یک الگوریتم تفسیر پذیری مدل است که برای فهمیدن عملکرد مدل های یادگیری ماشین استفاده می‌شود. این الگوریتم به طور خاص برای مدل هایی که به طور غیر خطی کار می‌کنند. LIME مخفف Local Interpretable Model-agnostic Explanations است و در واقع با استفاده از یک تابع خطی محلی[۱۷] برای تفسیر تصمیمات مدل، برای هر داده ورودی توضیحاتی را ارائه می‌دهد که بیان می‌کند کدام ویژگی‌ها برای تصمیم گیری مهم هستند [۲۰].

در روش ما ابتدا مجموعه داده متوازن شده را به الگوریتم جنگل تصادفی می‌دهیم سپس داده‌های مثبت کاذب را جدا می‌کنیم و به الگوریتم LIME می‌دهیم و سپس ویژگی‌های مهم و تاثیرگذار را جدا می‌کنیم.

۳-۴-۳- درخت تصمیم

استخراج قوانین از درخت تصمیم، می‌تواند بسیار مفید و حیاتی در بسیاری از مدل یادگیری ماشین باشد. در ادامه توضیحاتی درباره اهمیت استخراج قوانین از درخت تصمیم آورده شده است:

۱- شفافیت و قابلیت تفسیر: درخت تصمیم، به عنوان یک مدل یادگیری ماشین، می‌تواند بسیار پیچیده و ساختاری باشد. استخراج قوانین از درخت تصمیم می‌تواند به کاربران کمک کند تا ساختار و عملکرد درخت را به طور کامل درک کنند و قادر باشند تصمیمات بهتری بگیرند.

۲- کاهش هزینه‌ها و زمان: استخراج قوانین از درخت تصمیم می‌تواند به کاربران کمک کند که در برنامه‌های بزرگ و پیچیده‌تر، زمان و هزینه‌های بیشتری را برای آموزش مدل یادگیری ماشین صرف کنند. با داشتن قوانین استخراج شده، می‌توان به راحتی تعداد داده‌های بیشتری را پردازش کرد و به مدل آموزش داده شده اعتماد بیشتری داشت.

۳- افزایش دقت و بهبود عملکرد: قوانین استخراج شده از درخت تصمیم می‌توانند به کاربران کمک کنند تا عملکرد مدل یادگیری ماشین را بهبود بخشند. با استفاده از این قوانین، می‌توان به طور مستقیم به پیش‌بینی‌ها و تصمیم‌گیری‌های بهتری دست یافت.

۴- کمک به ادمین شبکه: قوانین استخراج شده می‌تواند به ادمین‌های شبکه کمک کند که با استفاده از الگوها و قوانین حملات را راحت‌تر شناسایی کنند.

در روش ما ابتدا درخت تصمیم را برروی داده‌های خود پیاده‌سازی می‌کنیم سپس الگوهای تکرار شونده را از درخت خود استخراج می‌کنیم.

۴- ارزیابی

در این مقاله ما یک روش ترکیبی مدل‌های تفسیر پذیری ارائه دادیم که و براساس این الگوریتم‌های SHAP و LIME بهترین ویژگی‌ها را انتخاب کردیم که ترکیبی از ویژگی‌های حجمی بسته‌ها و ویژگی‌های تعداد پرچم‌ها می‌باشد.

در این مقاله ما دو دسته ارزیابی انجام دادیم: ارزیابی کمی[۱۸]، ارزیابی کیفی[۱۹].

۱-۴- ارزیابی کمی

در این بخش به ارزیابی کمی روش پیشنهادی می‌پردازیم. به همین منظور در ابتدا به بررسی ویژگی‌های مهم الگوریتم‌های استفاده شده در روش پیشنهادی خواهیم پرداخت. ۱۵ ویژگی‌های مهم براساس الگوریتم SHAP به شرح زیر است(جدول (۱)).

بیشتر ویژگی‌هایی که براساس الگوریتم SHAP انتخاب شده است ویژگی‌های حجمی هستند.

۱۵ ویژگی‌های مهم براساس الگوریتم LIME در جدول (۲) آمده است. بیشتر ویژگی‌هایی که براساس الگوریتم LIME انتخاب شده است، ویژگی‌های براساس تعداد پرچم‌ها[۲۰] هستند.

جدول ۱ . ویژگی‌های مهم مجموعه‌داده براساس الگوریتم SHAP

نام ویژگی	توضیح ویژگی
Fwd Act Data Pkts	تعداد بسته‌هایی با حداقل ۱ بایت بار داده TCP در جهت جلو
Total Fwd Packet	مجموع بسته‌ها در جهت جلو
Flow Bytes/s	تعداد بایت‌های جریان در ثانیه
Fwd Packet Length Max	حداکثر اندازه بسته در جهت جلو
Fwd Packet Length Std	انحراف معیار اندازه بسته در جهت جلو
Packet Length Mean	میانگین طول یک بسته
Total Bwd packets	مجموع بسته‌ها در جهت عقب
Flow duration	مدت زمان جریان در میکروثانیه
Bwd IAT Std	زمان انحراف استاندارد بین دو بسته ارسال شده در جهت عقب
Bwd Packet Length Max	حداکثر اندازه بسته در جهت عقب
Bwd Packet Length Mean	اندازه متوسط بسته در جهت عقب
Total Length of Bwd Packet	اندازه کل بسته در جهت عقب
Total Length of Fwd Packet	اندازه کل بسته در جهت جلو
Protocol	نوع پروتکل را مشخص می‌کند
Fwd Packet Length Min	حداقل اندازه بسته در جهت جلو

از نظر کمی همانطور که جدول (۳) مشاهده می‌کنید از نظر کمی روش ما باعث افزایش دقت می‌شود و از همچنین هم روش ما به علت کاهش تعداد ویژگی باعث کاهش زمان اجرا می‌شود.

۲-۴- ارزیابی کیفی

در این بخش به سناریوهای پرتکرار در درخت تصمیم می‌پردازیم.

سناریو ۱:

```
| | | |--- act_data_pkt_fwd <= 0.50
| | | | |--- Total Backward Packets <= 33.50
| | | | | |--- class: BENIGN
| | | | |--- Total Backward Packets > 33.50
| | | | | |--- class: Syn
| | | |--- act_data_pkt_fwd > 0.50
| | | | |--- Fwd Packet Length Std <= 0.28
| | | | | |--- class: Syn
| | | | |--- Fwd Packet Length Std > 0.28
| | | | | |--- class: BENIGN
```

در سناریو ۱ تاثیر سه ویژگی act_data_pkt_fwd، Total Backward Packets و Fwd Packet Length Std در تشخیص حمله syn از داده‌های غیرحمله مشاهده می‌کنید.

سناریو۲:

```
| |--- Fwd Packet Length Min > 118.50
| | |--- Fwd Packet Length Max <= 319.50
| | | |--- Flow Duration <= 44.50
| | | | |--- Flow Bytes/s <= 276000000.00
| | | | | |--- class: Portmap
| | | | |--- Flow Bytes/s > 276000000.00
| | | | | |--- class: Portmap
| | | |--- Flow Duration > 44.50
| | | | |--- Flow Bytes/s <= 2083567.06
| | | | | |--- class: BENIGN
| | | | |--- Flow Bytes/s > 2083567.06
| | | | | |--- class: NetBIOS
```

در سناریو ۲ دو ویژگی جریانی Flow Duration و Flow Bytes/s در تشخیص حمله protmap و NetBIOS تاثیرگذار هستند.

سناریو ۳:

```
| | |--- Fwd Packet Length Max > 319.50
| | | |--- Total Length of Fwd Packets <= 737.00
| | | | |--- Fwd Packet Length Max <= 364.50
| | | | | |--- class: UDP
| | | | | |--- Fwd Packet Length Max > 364.50
| | | | | | |--- class: UDPLag
| | | | |--- Total Length of Fwd Packets > 737.00
| | | | | |--- Fwd Packet Length Mean <= 399.50
| | | | | | |--- class: UDP
| | | | | |--- Fwd Packet Length Mean > 399.50
| | | | | | |--- class: UDP
```

در سناریو ۳ مشاهده می‌کنید بالا بودن ویژگی‌های حجمی علت تشخیص حملات UDP می‌باشد.

از نظر کیفی توانستیم سناریوها و ویژگی‌های مهم برای تشخیص نوع حملات را تشخیص دهیم که این امر به کارشناسان شبکه کمک می‌کند تا حملات مختلف را بهتر شناسایی و تشخیص دهند.

جدول ۲. ویژگی‌های مهم مجموعه‌داده براساس الگوریتم LIME

نام ویژگی	توضیح ویژگی
Packet Length Std	انحراف معیار طول یک بسته
Init Win bytes forward	تعداد کل بایت‌های ارسال شده در پنجره اولیه در جهت جلو
Packet Length Variance	واریانس طول یک بسته
Min Packet Length	حداقل طول یک بسته
ACK Flag Count	تعداد بسته‌های دارای ACK
URG Flag Count	تعداد بسته های دارای URG
Fwd Packet Length Min	حداقل اندازه بسته در جهت جلو
Total Backward Packets	مجموع بسته‌ها در جهت عقب
Subflow Bwd Packets	میانگین تعداد بسته‌ها در یک زیر جریان در جهت عقب
Bwd Packets/s	تعداد بسته های در جهت عقب در ثانیه
SYN Flag Count	تعداد بسته‌های دارای SYN
Fwd Packet Length Mean	اندازه متوسط بسته در جهت جلو
Fwd PSH Flags	تعداد دفعاتی که پرچم PSH در بسته هایی که در جهت جلو حرکت می‌کنند تنظیم شده است (۰ برای UDP)
RST Flag Count	تعداد بسته‌های دارای RST
Average Packet Size	اندازه متوسط بسته

جدول ۳۰ ارزیابی کمی الگوریتم‌های پیاده شده مقاله

نام الگوریتم	Accuracy	Precision	Recall	F1-score
جنگل تصادفی	٪۸۴	٪۸۳	٪۸۳	٪۸۲
یادگیری گروهی (RF-SVM-KNN)	٪۸۷	٪۸۸	٪۸۷	٪۸۶
یادگیری گروهی + SHAP	٪۸۷	٪۸۷	٪۸۸	٪۸۶
یادگیری گروهی + (ENIXMA)	**٪۸۹**	**٪۹۰**	**٪۸۹**	**٪۸۸**

۵- نتیجه‌گیری

با گسترش روزافزودن استفاده از اینترنت، شناسایی و سپس مقاوم‌سازی شبکه در مقابل حملات DDoS، یکی از مهم‌ترین اهداف مسئولین شبکه است. در این مقاله، ضمن بررسی کارهای پیشین در زمینه شناسایی حملات، به ارائه یک روش جدید به نام انیکسم[۲۱] برای شناسایی و تفسیر حملات انجام شده در شبکه پرداختیم. انیکسما از لحاظ عددی، باعث ۳ درصد بهبود در دقت روش‌های پیشین شده است. همچنین انیکسما، با بهره‌جویی از روش‌های تفسیرپذیری در یادگیری ماشین، نه تنها حملات را شناسایی می‌کند بلکه به دلایل رخداد حمله و ویژگی‌های اثرگذار بر آن نوع حمله را تشریح می‌نماید. تفسیرپذیری حمله، امکان اقدام مناسب در جهت مقاوم‌سازی را برای مسئول شبکه فراهم می‌آورد. در راستای کارهای آتی، شناسایی حملات دیگر شبکه نظیر پورت–اسکن و بهره جویی از روش‌های پیشرفته‌تر تفسیرپذیری پیشنهاد می‌گردد.

مراجع

[1] M. Aamir and S. M. Ali Zaidi, “Clustering based semi-supervised machine learning for DDoS attack classification,” *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 4, May 2021, doi: 10.1016/j.jksuci.2019.02.003.

[2] S. Zavrak and M. Iskefiyeli, “Anomaly-Based Intrusion Detection From Network Flow Features Using Variational Autoencoder,” *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3001350.

[3] R. Bhatia, R. Sharma, and A. Guleria, “Anomaly Detection Systems Using IP Flows: A Review,” 2021. doi: 10.1007/978-981-16-0235-1_80.

[4] M. M. Hassan, A. Gumaei, A. Alsanad, M. Alrubaian, and G. Fortino, “A hybrid deep learning model for efficient intrusion detection in big data environment,” *Information Sciences*, vol. 513, Mar. 2020, doi: 10.1016/j.ins.2019.10.069.

[5] S.-T. Chiu and F.-Y. Leu, “Detecting DoS and DDoS Attacks by Using CuSum Algorithm in 5G Networks,” 2021. doi: 10.1007/978-3-030-57811-4_1.

[6] M. Nooribakhsh and M. Mollamotalebi, “A review on statistical approaches for anomaly detection in DDoS attacks,” *Information Security Journal: A Global Perspective*, vol. 29, no. 3, May 2020, doi: 10.1080/19393555.2020.1717019.

[7] S. Hosseini and M. Azizi, “The hybrid technique for DDoS detection with supervised learning algorithms,” *Computer Networks*, vol. 158, Jul. 2019, doi: 10.1016/j.comnet.2019.04.027.

[8] M. Du, N. Liu, and X. Hu, “Techniques for interpretable machine learning,” *Commun. ACM*, vol. 63, no. 1, pp. 68–77, 2020, doi: 10.1145/3359786.

[9] C. Yin, Y. Zhu, J. Fei, and X. He, “A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks,” *IEEE Access*, vol. 5, pp. 21954–21961, Oct. 2017, doi: 10.1109/ACCESS.2017.2762418.

[10] M. M. Hassan, A. Gumaei, A. Alsanad, M. Alrubaian, and G. Fortino, “A hybrid deep learning model for efficient intrusion detection in big data environment,” *Information Sciences*, vol. 513, pp. 386–396, Mar. 2020, doi: 10.1016/j.ins.2019.10.069.

[11] A. Girma, M. Garuba, Jiang Li, and Chunmei Liu, “Analysis of DDoS Attacks and an Introduction of a Hybrid Statistical Model to Detect DDoS Attacks on Cloud Computing Environment,” Apr. 2015. doi: 10.1109/ITNG.2015.40.

[12] R. B. Blažek, H. Kim, B. Rozovskii, and A. Tartakovsky, “A novel approach to detection of ‘denial-of-service’ attacks via adaptive sequential and batch-sequential change-point detection methods,” 2001.

[13] S. R. Gaddam, V. v Phoha, and K. S. Balagani, “K-Means+ID3: A Novel Method for Supervised Anomaly Detection by Cascading K-Means Clustering and ID3 Decision Tree Learning Methods.”

[14] Hoai-Vu Nguyen and Yongsun Choi, “Proactive Detection of DDoS Attacks Utilizing k-NN Classifier in an Anti-DDos Framework,” *World Academy of Science, Engineering and Technology* , 2010.

[15] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, and J. Srivastava, “A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection,” May 2003. doi: 10.1137/1.9781611972733.3.

[16] C.-K. Han and H.-K. Choi, “Effective discovery of attacks using entropy of packet dynamics,” *IEEE Network*, vol. 23, no. 5, Sep. 2009, doi: 10.1109/MNET.2009.5274916.

[17] C. Di Francescomarino and C. Ghidini, “Predictive Process Monitoring,” in *Lecture Notes in Business Information Processing*, 2022, vol. 448, pp. 320–346. doi: 10.1007/978-3-031-08848-3_10.

[18] W. Rizzi, C. Di Francescomarino, and F. M. Maggi, “Explainability in predictive process monitoring: When understanding helps improving,” in *Lecture Notes in Business Information Processing*, 2020, vol. 392 LNBIP, pp. 141–158. doi: 10.1007/978-3-030-58638-6_9.

[19] R. Sindhgatta, C. Ouyang, and C. Moreira, “Exploring interpretability for predictive process analytics,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020, vol. 12571 LNCS, pp. 439–447. doi: 10.1007/978-3-030-65310-1_31.

[20] D. Adi and N. Nurdin, “Explainable Artificial Intelligence (XAI) towards Model Personality in NLP task,” *IPTEK J. Eng.*, vol. 7, no. 1, p. 11, 2021, doi: 10.12962/j23378557.v7i1.a8989.

[21] I. Sharafaldin, A. H. Lashkari, S. Hakak, and A. A. Ghorbani, “Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy,” in Proc. 53rd International Carnahan Conference on Security Technology, Chennai, India, 2019.

[22] W. E. Marcilio and D. M. Eler, “From explanations to feature selection: Assessing SHAP values as feature selection mechanism,” in *Proceedings - 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images, SIBGRAPI 2020*, 2020, pp. 340–347. doi: 10.1109/SIBGRAPI51738.2020.00053.

زیرنویس

[1] Interpretability
[2] Flow Base
[3] K-Nearest Neighbors
[4] Support Vector Machine
[5] Random Forest
[6] Entropy
[7] Covariance
[8] False positive rate
[9] Epoch
[10] Online
[11] Local Interpretable Model-Agnostic Explanations
[12] SHapley Additive exPlanations
[13] K-Nearest Neighbors
[14] Analysis of variance
[15] Infinity
[16] Normalization
[17] local linear model
[18] Numerical Evaluation
[19] Quality Evaluation
[20] Flag
[21] ENIXMA