مجموعه مقالات پانزدهمین کنفرانس بین‌المللی انجمن رمز ایران
دانشکده مهندسی برق، دانشگاه تربیت دبیر شهید رجایی، تهران، ۶-۷ شهریور ۱۳۹۷

تشخیص بدافزار به کمک یادگیری فعال نیمه‌نظارتی

رضا رحیمیان[۱]، هدی مشایخی[۲]، محسن رضوانی[۳]

[۱] دانشجوی کارشناسی ارشد، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شاهرود، شاهرود
rezarahimian@shahroodut.ac.ir

[۲] استادیار، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شاهرود ، شاهرود
hmashayekhi@shahroodut.ac.ir

[۳] استادیار، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شاهرود ، شاهرود
mrezvani@shahroodut.ac.ir

چکیده

امروزه با توجه به ضرورت استفاده از اینترنت، رشد چشم‌گیر شبکه‌ها و زیرساخت‌های رایانه‌ای و همچنین طراحی بدافزارهای پیچیده و پویایی که دائم در حال به روز رسانی خود هستند، حفظ امنیت و نظارت بر ترافیک شبکه‌ها یکی از مهم‌ترین ملزومات فضای سایبری می‌باشد. به طور کلی بدافزارها پس از ورود به سیستم می‌توانند اقداماتی نظیر سرقت اطلاعات، ایجاد هرزنامه و یا تولید شبکه‌ای از بات‌ها را انجام دهند. بنابراین ایجاد روشی که بتواند به صورت کارا به شناسایی و جلوگیری از نفوذ آنها بپردازد، همواره مورد نیاز خواهد بود. در سال‌های اخیر بات‌نت‌ها به عنوان یکی از خطرناک‌ترین بدافزارهای شناخته شده در بستر اینترنت مطرح می‌شوند که قابلیت تخریب رایانه‌های سالم و تبدیل آنها به بات‌هایی برای انتقال ویروس، اسپم و غیره را دارند. تشخیص بات‌نت‌ها با استفاده از روش های یادگیری چالش‌های متعددی دارد که از میان آنها می‌توان به کمبود داده‌های برچسب‌گذاری شده اشاره نمود. به منظور تخفیف این مشکل می‌توان از روش یادگیری فعال استفاده کرد که کمتر در زمینه تشخیص بات‌نت مورد توجه قرار گرفته است.

در این مقاله یک رویکرد مبتنی بر یادگیری فعال نیمه نظارتی با استفاده از رده‌بندهای لجستیک و ماشین بردار پشتیبان خطی، به منظور تشخیص بات‌نت ارائه شده است. آموزش در این روش به صورت تعاملی انجام شده و سیستم در حین اجرا دائما رده‌بند پایه را با توجه به نمونه‌های انتخابی خود که برچسب آنها درخواست می‌شود، به روز رسانی می‌نماید. برای انجام آزمایشات از مجموعه داده‌ای حاوی انواع مختلف بات‌نت استفاده کرده و پنج مجموعه ویژگی مختلف را استخراج می‌کنیم. نتایج بدست آمده، کارایی مدل را در تشخیص بات‌نت‌های دیده نشده و دقت ۸۹/۸۵ درصد را نشان می‌دهد.

کلمات کلیدی

یادگیری فعال نیمه نظارتی، بدافزار ، بات‌نت، رده بندی ترافیک، رایانش امن، لجستیک، ماشین بردار پشتیبان خطی

۱- مقدمه

در دنیای امروزی ضرورت استفاده از اینترنت و تبدیل شدن آن به عنوان بخش مهمی از زندگی افراد، موضوعی غیر قابل اغماض است. در واقع رشد فزاینده و استفاده‌ی همگانی افراد از اینترنت به عنوان یک سوژه‌ی جذاب برای انجام کارهای خرابکارانه توسط مهاجمان تبدیل شده و انگیزه آنان برای نفوذ در شبکه و ایجاد حملات اینترنتی گسترده‌تر، همواره در حال شدت یافتن است.

به طور کلی بدافزارها، به قطعه کدهایی اطلاق می شود که توسط برنامه نویسان تولید شده و هدف آنها از ایجاد این کدها، آلوده کردن، خرابکاری و کارهای مجرمانه بدون اطلاع مالک سیستم خواهد بود. بدافزارها اغلب به منظور تخریب رایانه‌های افراد قربانی از طریق به‌کارگیری نرم‌افزارهای آسیب‌پذیر و یا فریب

کاربران به اجرای کدهای خرابکارنه مورد استفاده قرار می‌گیرند. تشخیص چنین فرآیندی و نحوه استفاده مهاجم از راه‌های نفوذ همانند درهای پشتی، پویشگر کلید، سرقت رمزهای ورود و سایر توابع بدافزاری، همواره پیچیده‌تر و به مسئله ای دشوارتر تبدیل شده است [1].

تاکنون رویکردهای مختلفی به منظور تشخیص حملات و شناسایی بدافزارها ارائه شده است که در این بین، رده‌بندی ترافیک شبکه به عنوان یکی از شناخته شده‌ترین رویکردهای امنیتی مطرح است [2]. به طور کلی این رویکرد با هدف تعیین و طبقه‌بندی کلاس‌های ناشناخته مورد استفاده قرار می‌گیرد که در یک تقسیم‌بندی کلی به دو زیر مجموعه، روش‌های کلاسیک و مدرن طبقه‌بندی می‌شود. در روش رده‌بندی جریان ترافیک کلاسیک، پیش‌بینی مبتنی بر پورت و بررسی مبتنی بر محموله[1]صورت می‌پذیرد. امروزه تکنیک اول، به علت افزایش برنامه‌های نظیر به نظیر[2]که از شماره پورت‌های پویا استفاده می‌کنند دارای محدودیت بوده و کارایی چندان مطلوبی را ندارد. دومین روش این راهکار یعنی مبتنی بر محموله، نیز به علت وجود برنامه‌های مبتنی بر شبکه داده‌ی رمزگذاری شده و استفاده‌ی آنها از تکنیک‌های مختلف رمزنگاری، شکست خورده است و دیگر نتایج مناسبی را به همراه نخواهد داشت [3]. بنابراین روش های مدرن به منظور رفع اشکالات راهکارهای گذشته پیشنهاد شدند که شامل روش‌های مبتنی بر یادگیری ماشین، روش‌های آماری و مبتنی بر رفتار می شوند. روش‌های مبتنی بر یادگیری ماشین که راهکار پیشنهادی این مقاله نیز بر اساس آن ارائه شده است، می‌تواند به خوبی نوع برنامه‌های موجود در جریان ترافیک شبکه را ارزیابی و رده‌بندی نماید [1].

در میان بدافزارهای موجود، ایجاد بات‌نت‌ها روندی رو به رشد داشته و به عنوان یکی از بزرگ‌ترین تهدیدات امنیتی به شمار می‌رود. در واقع بات یک رایانه آلوده شده به بدافزار است که بدون آگاهی و اراده‌ی کاربر و از راه دور توسط یک یا چند عامل انسانی یا ماشین کنترل می‌شود. به این عامل کنترل کننده، بات مرکزی یا چوپان‌بات[3]گفته می شود. علت اصلی خطرناک بودن این گونه شبکه‌ها وجود تعداد زیادی از رایانه‌ها است که چوپان‌بات می‌تواند از پهنای باند، قدرت ذخیره‌سازی و پردازش هر یک از این رایانه‌ها بهره برده و در راستای اهداف مخرب خود، استفاده نماید. حملات ممانعت از سرویس‌دهی توزیع شده[4]، فعالیت‌های فریب‌کارانه همانند تولید هرزنامه، شنود[5]، سرقت هویت[6]و نشر اطلاعات[7]از جمله حملاتی است که توسط بات‌نت‌ها صورت می‌پذیرد [4].

تاکنون روش‌های مختلفی به منظور شناسایی بات‌نت‌ها ارائه شده است. تکنیک های مبتنی بر یادگیری ماشین که یک فرآیند با ناظر [5] را برای شناسایی ترافیک بات‌نت در نظر گرفته‌اند تا تکنیک‌های بدون ناظر [6] که با درصد کمی از بات‌نت‌های ناشناخته همراه هستند. هرچند که میزان نرخ تشخیص درست این مقالات نزدیک به ۱۰۰ می‌باشد اما همگی بر اساس این فرض ارائه شده‌اند که ما نسبت به زمینه‌ی واقعی داده‌ها (حتی آنهایی که ناشناخته هستند) اطلاع کامل داریم؛ که این موضوع توسعه‌پذیری سیستم‌های تشخیص را با محدودیت همراه می‌کند. یک چالش اساسی دیگر در سیستم‌های تشخیص بات نت، همچون سیستم‌های تشخیص نفوذ؛ توانایی شناسایی بات‌نت‌های جدید و گونه‌های تاکنون مشاهده نشده، خواهد بود.

در این مقاله با هدف بهبود و رفع چالش‌های مطرح شده، رویکردی مبتنی بر یادگیری فعال نیمه‌نظارتی، ارائه می‌دهیم که بعد از مرحله آموزش اولیه، قادر است، نمونه‌های جدید مشاهده شده را بررسی کرده و با انتخاب نمونه‌های مشخص، برچسب آنها را از مدیر شبکه (خبره) درخواست نماید. لازم به ذکر است که نحوه تصمیم‌گیری خبره می‌تواند بر اساس الگوی محموله بسته‌ها، میزان شباهت جریان‌ها به بات‌نت‌های کشف شده، اطلاعات بدست آمده از هانی‌نت و غیره باشد. این سیستم نیاز به داده‌های اولیه برچسب‌گذاری شده با حجم زیاد را از بین برده و به صورت فعال در صورت مواجهه با نمونه‌های جدید بات‌نت یا نمونه‌هایی که دقت تشخیص کافی در باب آنها ندارد، رده‌بند خود را به روز رسانی می‌نماید. بنابراین مدل ایجاد شده قدرت تشخیص بات‌نت‌های جدید را نیز دارا خواهد بود. سیستم پیشنهادی بر مبنای رده‌بندهای لجستیک و ماشین بردار پشتیبان خطی عمل می‌کند. این روش جزء اولین روش‌هایی است که برای تشخیص بات‌نت از یادگیری فعال استفاده می‌کند.

برای ارزیابی روش پیشنهادی از مجموعه داده بات‌نت ISCX استفاده می شود که کامل‌ترین مجموعه داده از نظر تنوع بات‌نت می‌باشد. پنج مجموعه ویژگی مختلف از مجموعه داده استخراج شده و عملکرد روش پیشنهادی با این ویژگی‌ها مقایسه می‌گردد. نتایج حاکی از کارایی مدل پیشنهادی در تشخیص بات‌نت‌های جدید و دقت رده‌بندی ۸۹/۸۵ درصد می‌باشد

۲- کارهای پیشین

به طور کلی بات‌نت‌ها را می‌توان بر اساس دو معیار مرتبط با کانال‌های فرمان و کنترل، به دو بخش ساختار و پروتکل طبقه‌بندی نمود. بر اساس ساختار به سه دسته‌ی متمرکز، غیرمتمرکز و ترکیبی تقسیم‌بندی می‌شوند که تفاوت آنها در نحوه عملکرد بات‌ها در قالب مدل مشتری و سرویس‌دهنده‌های کنترل و فرمان می‌باشد [7]. همچنین بر اساس پروتکل مورد استفاده در این کانال‌ها، به سه نوع، مبتنی بر IRC، مبتنی بر HTTP و نظیر به نظیر دسته‌بندی می شوند که در آنها بات مرکزی سعی در حفظ ارتباط خود با قربانیان و به به روزرسانی برنامه‌ها و دستورات را دارد [8].

در حالت کلی روش‌های تشخیص بات‌نت بر اساس سه معیار طبقه‌بندی می شوند: ۱) موقعیت بات‌نت در چرخه حیات به هنگام تشخیص ۲) رویکرد یادگیری و ۳) میزان سطح همبستگی. بر اساس معیار اول، عملیات تشخیص می‌تواند در مراحل آغازین، یعنی در زمان شکل‌گیری بات‌نت و ایجاد کانال کنترل و فرمان و یا در مرحله‌ی حمله‌ی بات‌نت صورت گیرد. به طور کلی، دقت روش‌هایی که در مرحله حمله، شناسایی را انجام می‌دهند بالاتر است؛ اما از طرفی تشخیص بات‌نت‌ها در مراحل آغازین، از مشارکت آن‌ها در فعالیت‌ها و حمله‌های مخرب، جلوگیری می‌کند [9]. بر اساس معیار دوم، رویکرد یادگیری می‌تواند باناظر یا بدون ناظر باشد؛ که معمولا در یادگیری بر خط، به دلیل محدودیت نیاز به داده‌های برچسب‌گذاری شده، از روش‌های باناظر استفاده نمی‌شود [10]. در نهایت بر اساس معیار سوم نیز، روش تشخیص بات‌نت می تواند بر اساس دو سطح مختلف از تحلیل همبستگی[8]یعنی سطح گروهی و

---

[1] Payload-based Inspection
[2] Peer to peer applications(P2P)
[3] Bot master (Bot herder)
[4] Distributed denial-of-service (DDoS) attacks
[5] Phishing
[6] Identity theft
[7] Information exfiltration
[8] Honeynet
[9] http://www.unb.ca/cic/datasets/botnet.html
[10] Correlation analysis

انفرادی صورت پذیرد. در تحلیل سطح انفرادی، شناسایی بر اساس رفتارهای فردی هر سیستم صورت می‌گیرد و رفتار سایر سیستم‌های آلوده در نظر گرفته نمی‌شود. مزیت این روش‌ها در این است که اگر در شبکه مورد نظر، تنها یک بات وجود داشته باشد، آن را تشخیص دهند. از سوی دیگر روش‌های مبتنی بر تحلیل سطح گروهی بر اساس یافتن الگوی مشابه بین دو یا چند سیستم عمل می‌کنند و آن‌ها را به عنوان اعضای بات‌نت، تشخیص و معمولا دقت در این روش‌ها نسبت به حالت قبل بالاتر ولی تنها قادر به شناسایی بات‌نت‌هایی با عضویت مشترک خواهند بود [8].

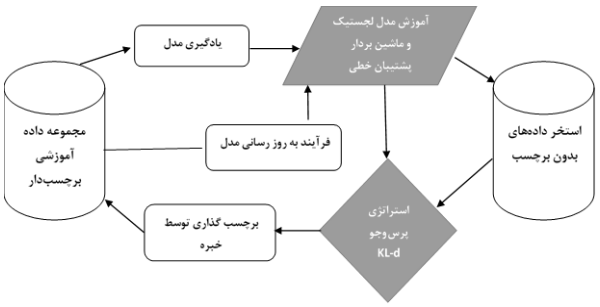
یادگیری فعال نیمه‌نظارتی گونه‌ای از روش‌های یادگیری ماشین است که با هدفی واحد در دو جهت مختلف به یک مسئله یادگیری نگاه می‌کند. به این صورت که روش نیمه‌نظارتی آن به دنبال حدس زدن در خصوص داده‌های بدون برچسب از روی مدل ساخته شده و روش فعال به دنبال کشف جنبه‌های ناشناخته‌ی داده‌های بدون برچسب است تا موثرترین داده‌ها برای فرآیند برچسب‌گذاری انتخاب شوند. این روش در سیستم‌های امنیتی بیشتر به منظور یادگیری و به روز رسانی در مدل فراگرفته شده است که باید با حجم محدودی از داده‌های بر چسب خورده به انجام برسد [11]. یادگیری فعال روشی است که در آن الگوریتم یادگیری قادر به تعامل با خبره[1]یا کاربر، از طریق پرس‌و جو و یا برخی منابع اطلاعاتی دیگر خواهد بود و برای دستیابی به نتایج بهتر، بر اساس نقاط داده‌ای جدید اقدام به کار می‌کند. سه شیوه اصلی یادگیری فعال عبارتند از: ترکیب پرس‌وجوی عضویت[2]، نمونه‌برداری گزینشی مبتنی بر جریان[3] نمونه‌برداری مبتنی بر استخر[4][12]. رویکردهای مبتنی بر یادگیری فعال با هدف اصلاح مشکلات سیستم‌های تشخیص نفوذ مورد استفاده قرار گرفتند. در این سیستم‌ها از طریق انجام یک فرآیند به روز رسانی فعال سعی در کاهش هزینه‌ی برچسب‌گذاری جریان‌های ترافیکی شبکه شد [13]. از سوی دیگر یافتن تعداد محدودی از برچسب‌های جریان حمله با استفاده از یک چارچوب یادگیری فعال و استفاده از آموزش نیمه‌نظارتی با هدف تمیز دادن جریان نرمال از حمله و سپس برچسب‌گذاری کل نمونه‌های ناشناخته نیز در این زمینه صورت پذیرفته که با کمبود تنوع حملات، به خصوص تشخیص بات نت‌های جدید همراه بوده است [14]. در این مقاله رویکرد مبتنی بر یادگیری فعال نیمه‌نظارتی، تحلیل انفرادی و تشخیص در مرحله آغازین است که سیستم در حین اجرا دائما رده‌بند پایه‌ی خود را با توجه به نمونه‌های جدیدی که مشاهده می‌کند، به روز رسانی می‌نماید.

۳- روش پیشنهادی

در این مقاله یک رویکرد جدید با استفاده از یادگیری فعال نیمه‌نظارتی برای تشخیص بات‌نت‌ها و شناسایی جریان ترافیک نرمال از حمله، ارائه شده است. پس از استخراج ویژگی‌های مبتنی بر جریان‌های ترافیک، ابتدا مدل رده‌بندی داده‌ها با تعداد کمی جریان‌های برچسب‌دار ساخته می‌شود. سپس با مشاهده جریان‌های جدید، میزان مفید بودن اطلاعات آنها را به کمک استراتژی‌های موجود در یادگیری فعال، اندازه‌گیری کرده و برچسب جریان‌های انتخاب شده از فرد خبره استعلام می‌شود. جریان‌های مفروض به صورت یک سری بردارهای ویژگی‌های استخراج شده هستند و با اعمال فرآیند برچسب‌گذاری، سیستم در حین اجرای، عملیات به روز رسانی رده‌بند را با توجه نمونه‌های جدیدی که مشاهده می‌کند، انجام می‌دهد. در ادامه، ابتدا به شرح روش یادگیری و سپس مجموعه ویژگی‌های استفاده شده، می‌پردازیم.

۳-۱- یادگیری فعال نیمه‌نظارتی

سیستم یادگیری فعال پیشنهاد شده در این مقاله دارای فرایندی است که در شکل (۱) مشاهده می‌شود. در این سیستم که بر اساس سناریوی مبتنی بر استخر پیاده‌سازی شده است، در ابتدا ۱۰درصد از داده‌ها برای آموزش اولیه‌ی رده‌بند استفاده شده و بقیه داده‌ها بدون برچسب باقی می‌مانند. علت اصلی این امر کمبود داده‌های برچسب خورده و حجم بالا نمونه‌های بدون برچسب در کاربردهای دنیای واقعی خواهد بود. رده‌بند مورد استفاده در این مقاله به صورت تجمیعی[1] مبتنی بر رده‌بندهای لجستیک و ماشین بردار پشتیبان خطی می باشد. به طور کلی با توجه به نوع ویژگی‌های استخراج شده که حاصل یک سری عملیات احتمالاتی هستند، نحوه‌ی انتخاب رده‌بندها بر اساس ساز وکار فرآیند رده‌بندی و همچنین خاصیت پارامتریکی آنها می‌باشد؛ که سیستم در نهایت با مشاهده هر نمونه داده جدید، با توجه به نوع استراتژی یادگیری فعال، می‌تواند برچسب داده‌ی منتخب را از فرد خبره استعلام نماید؛ در واقع داده‌ی منتخب ارزش اطلاعاتی بالایی را برای بهبود عملیات رده‌بندی در پی خواهد داشت.



شکل (۱) : یادگیری فعال نیمه‌نظارتی مبتنی بر استخر

یادگیری فعال دارای استراتژی‌های پرس‌وجوی مختلفی می‌باشد که از مهمترین آنها می‌توان به نمونه‌برداری مبتنی بر آنتروپی، نمونه‌برداری تصادفی، آنتروپی رای‌گیری، واگرایی Kullbak-Liebler[6](KL)، اتلاف لگاریتم مورد انتظار[7]نمونه‌برداری بیشترین شباهت و غیره اشاره کرد که هرکدام به گونه‌ی خاصی داده‌های بدون برچسب را برای پرسش از خبره انتخاب می‌کنند [12]. ما در این مقاله از راهکار واگرایی KL استفاده کردیم که به عنوان یکی از استراتژی‌های پرس‌وجوی هیئت بررسی می‌باشد و بر اساس اختلاف میزان اطلاعات خروجی چند رده‌بند نسبت به انتخاب داده‌ها اقدام می‌کند. در واقع این راهکار میانگین اختلاف هر توزیع احتمالاتی را با میانگین کل توزیع داده‌ها اندازه‌گیری می‌نماید (بین رای هر عضو هیئت بررسی با اجماع کلی) [15]. مدل یادگیری برای به کارگیری در استراتژی واگرایی KL و همچنین آموزش داده ها بر اساس رده‌بندهای لجستیک و مدل خطی ماشین بردار پشتیبان صورت پذیرفت که در ادامه شرح داده می‌شوند.

1 Oracle
1 Membership Query Synthesis
1 Stream Based Selective Sampling
1 Pool Based Sampling
1 Ensemble
1 Kullback-Leibler divergence (KL-d)
1 Expected Log-Loss
1 Query-By-Committee

۲-۳- رگرسیون لجستیک

رگرسیون لجستیک یک رده‌بند دو کلاسه خطی است که احتمال کلاس را بر پایه‌ی تابع سیگموئید تخمین می‌زند:

$$P(Y=1|y)=\frac{1}{1+e^{-y}} \qquad (۱)$$

که در آن y به صورت تابع خطی از ورودی x در نظر گرفته می شود:

$$y=w_0+w_1x_1+w_2x_2+\cdots+w_nx_n \qquad (۲)$$

پارامترهای مدل لجستیک با روش ML[19] محاسبه می‌شوند [16].

۳-۳- ماشین بردار پشتیبان خطی

ما به منظور استفاده از یک روند افزایشی برای همگرایی خطی داده‌ها و استفاده از رده‌بند ماشین بردار پشتیبان نسبت به کمینه ساختن تابع اتلاف آن (تابع hinge) از طریق الگوریتم گرادیان نزولی تصادفی اقدام کردیم. به طور کلی مشکل رده‌بندی و رگرسیون داده‌های بزرگ معمولا با استفاده‌ی صحیح از روش مرتبه‌ی دوم گرادیان تصادفی[20] و تکنیک‌های گرادیان تصادفی متوسط برطرف می‌شود. در این روش تابع هزینه با استفاده از الگوریتم گرادیان نزولی تصادفی به صورت رابطه‌ی (۳) کمینه می‌گردد.

$$w_{k+1}=w_k-\mu\nabla_w Q(x_k,w_k) \qquad (۳)$$

که در آن μ به عنوان نرخ یادگیری الگوریتم SGD و $Q(x_k,w_k)$ به عنوان تخمین‌زننده آنی تابع اتلاف[23] $Q(x,w)$ است که از بردار ورودی x_k برای نمونه‌ی k ام استفاده می‌کند. بر همین اساس، مدل پارامترها یا ویژگی‌ها (w_k) با استفاده از هر بردار ورودی، به صورت افزایشی اقدام به روز رسانی می‌نماید. در نهایت زمانی که μ به اندازه‌ی کافی کوچک شد، الگوریتم SGD یک همگرایی خطی برای داده‌ها به دست می‌آورد. [17].

۴-۳- مجموعه ویژگی‌ها

در این مقاله از پنج مجموعه ویژگی مختلف استفاده شده است. برای ساخت مجموعه ویژگی اول که آن را Qiu2017 [18] می‌نامیم، از هر جریان، ۱۰ بسته‌ی نخست آن (بعد از عملیات دست‌تکانی) را به عنوان بسته‌های مطلوب انتخاب کردیم. سپس با توجه به اینکه بسته‌های موجود بین مشتری به سرور و سرور به مشتری متناوب هستند؛ برای هر بسته اندازه آن را در نظر گرفته و در صورت عدم وجود بسته بعدی از سمت مقابل، یک بسته با اندازه صفر وارد فضای ویژگی می‌کنیم؛ بنابراین بردار ویژگی به ابعاد ۲۰ بر روی هر ۱۰ بسته ی نخست که توسط دو ویژگی اندازه بسته‌ها و جهت ارسال مشخص شده‌اند، در نظر گرفتیم. این بازنویسی اندازه بسته‌ها سبب حفظ اطلاعات جریان‌های دو جهته به منظور تشخیص بات‌نت و ترافیک نرمال شبکه از هم خواهد شد. در ادامه مقادیر این بردارهای ویژگی باینری شده و به ورودی یک شبکه بیزین اعمال می‌شود و با محاسبه‌ی بیشترین مقدار راست‌آزمایی[24] بر روی تابع مدل ساخته شده، تمام بسته‌های تکی و همچنین جفت شده ، از طریق شمارش تکرار نمونه‌ها انجام می‌پذیرد. در مرحله بعد با استفاده از تک ویژگی و جفت ویژگی‌هایی که دارای مقادیر مثبت و مخالف صفر هستند، نسبت به مدل‌سازی داده‌ها به کمک مدل مخلوطی گوسی[25] اقدام می‌کنیم. (لازم به ذکر است که برای جفت ویژگی‌ها از یک مدل گوسی مخلوطی دو مقداره[26] استفاده می‌شود). در نهایت میزان انحراف احتمالاتی ویژگی دو بعدی یا تک بعدی ورودی، نسبت به پراکندگی و چگالی کل ویژگی‌ها محاسبه می‌شود. لازم به ذکر است که ما در تولید این مجموعه ویژگی ابتدا به اندازه‌ی ۱۰ درصد کل داده‌ها (همانند فرآیند یادگیری سیستم پیشنهادی) نسبت به شمارش تکرار نمونه‌ها و انجام محسابات احتمالاتی اقدام کردیم و سپس برای مابقی داده‌ها، عملیات شمارش را نسبت به همان مجموعه اولیه (۱۰درصد کل داده‌ها) انجام دادیم؛ البته با فرض اینکه نمونه جدیدی به مجموعه اولیه اضافه نمی‌شود. در پایان با انجام مراحل گفته شده، مجموعه ویژگی ۲۳۰ بُعدی استخراج می‌گردد.

مجموعه ویژگی دوم Milcom2015 نام دارد. در مقاله [19] نویسندگان تعداد ۱۰ ویژگی موثری که در تشخیص بات‌نت تاثیرگذار هستند را از بین ۲۴۸ ویژگی که در مقاله [20] ارائه شده است، انتخاب کردند. ویژگی‌های مشتق شده، حاصل اعمال فیلتر میزان همبستگی[27] داده‌ها بر روی مجموعه داده‌های مختلف بوده که به عنوان مجموعه ویژگی‌های نهایی در نظر گرفته شده است. نام هر ویژگی به همراه توضیحات مرتبط با آن در جدول (۱) مشاهده می‌شود.

جدول (۱) : مجموعه ویژگی MilCom2015

نام ویژگی	توضیحات
Flow duration	طول مدت زمان جریان بر حسب میکرو ثانیه
Cnt_data_pkt	تعداد کل بسته‌ها با حداقل یک بایت از محموله (رو به جلو)
Min_data_size	کمترین اندازه محموله مشاهده شده (رو به جلو)
Mean_Bytes	میانگین بایت داده‌های رو به عقب (داده بر حسب بایت به تعداد کل بسته ها)
Int_data_Len	تعداد کل بایت‌های ارسالی قبل از مشاهده‌ی اولین بسته تصدیق (به صورت دو طرفه)[28]
RTT samples	تعداد کل نمونه‌های RTT یافت شده در کل بسته‌ها (رو به جلو)
Med_Bytes	میانه طول بایت بسته‌ها (رو به جلو)
Var_Bytes	واریانس طول بایت بسته‌ها (رو به عقب)
IP_ratio	نسبت بین بزرگ‌ترین اندازه بسته به کوچکترین اندازه بسته(دو طرفه)
Goodput	حاصل تقسیم تعداد کل بایت‌های فریم به مدت زمان جریان(دو طرفه)

مجموعه ویژگی سوم CIC2018 نام دارد [21] که در آن چهار ویژگی موثر برای تشخیص بات‌نت‌ها توسط الگوریتم جنگل تصادفی انتخاب شده است. نام هر ویژگی به همراه توضیحات مرتبط با آن برای این مجموعه ویژگی نیز در جدول (۲) مشاهده می‌شود.

جدول (۲) : مجموعه ویژگی CIC2018

نام ویژگی	توضیحات

1 Maximum Likelihood
2 Support vector machine (SVM)
3 Second order stochastic gradient
4 Averaged stochastic gradient
5 Loss function
6 Maximum Likelihood
7 Gaussian mixture model (GMM)
8 Bivariate Gaussian Mixture Model
9 Correlation-based filtering
10 Forwarding
11 Backwarding
12 Bidirectional

Flow duration	طول مدت زمان جریان بر حسب میکرو ثانیه
Total Len F.Packets	کل طول بسته‌های رو به جلو
SubFlow F.Bytes	تعداد بایت در زیر جریان رو به جلو
B.Packers/s	تعداد بسته‌های رو به عقب در ثانیه

مجموعه ویژگی چهارم که Li2009 نامیده می‌شود [22] دارای ۱۲ ویژگی موثر مستقل از مکان در شبکه و پایدار در زمان هستند که توسط معیار عدم قطعیت متقارن[1] انتخاب شده‌اند؛ مشخصات این مجموعه داده در جدول (۳) آورده شده است.

جدول (۳) : مجموعه ویژگی Li2009

نام ویژگی	توضیحات
Push_pkt_serv	تعداد کل بسته‌ها با مجموعه بیت PSH در سرآیند TCP (رو به عقب)
Int_win_bytes	تعداد کل بایت‌های ارسال شده در پنجره اولیه (رو به عقب)
Int_win_bytes	تعداد کل بایت‌های ارسال شده در پنجره اولیه (رو به جلو)
Avg_seg_size	میانگین اندازه سگمنت (داده بر حسب بایت به تعداد کل بسته ها) رو به عقب
IP_bytes_med	میانه کل بایت بسته‌ها (رو به جلو)
Act_data_pkt	تعداد کل بسته‌ها با حداقل یک بایت از محموله داده‌ی TCP (رو به جلو)
Data_bytes_var	واریانس طول بایت بسته‌ها (رو به عقب)
Min_seg_size	کمترین اندازه سگمنت مشاهده شده (رو به جلو)
RTT samples	تعداد کل نمونه‌های RTT یافت شده در کل بسته‌ها (رو به جلو)
Push_pkt_clnt	تعداد کل بسته‌ها با مجموعه بیت PSH در سرآیند TCP (رو به جلو)
Serv_port	پورت سرور
Clnt_port	پورت مشتری

در نهایت مجموعه ویژگی پنجم که ISCX2014 نامیده می‌شود [23]، شامل چهار ویژگی موثر است که بر اساس چهار معیار مبتنی بر زمان، مبتنی بر بسته، مبتنی بر جریان و مبتنی بر زمان طی یک سری آزمایشات متوالی انتخاب شده اند. مشخصات این مجموعه ویژگی نیز در جدول (۴) مشاهده می‌شود.

جدول (۴) : مجموعه ویژگی ISCX2014

نام ویژگی	توضیحات
Flow duration	طول مدت زمان جریان بر حسب میکرو ثانیه
IOPR	نسبت بین تعداد بسته‌های دریافتی بر تعداد بسته‌های ارسالی
APL	میانگین طول بسته‌ها
BS	میانگین بیت بر ثانیه بسته‌ها

۴- آزمایشات

در این بخش ابتدا مدل ارزیابی و مجموعه داده شرح داده شده و سپس نتایج ارائه می‌شود. به این صورت که هر کدام از مجموعه ویژگی‌ها تحت شرایط یکسان توسط فرآیند یادگیری فعال نیمه‌نظارتی مورد مقایسه قرار می‌گیرند.

۴-۱- مدل ارزیابی و مجموعه داده

همانطور که در بخش‌های قبلی نیز عنوان شد، یکی از نیازمندی‌های سیستم‌های تشخیص بات‌نت، تنوع گونه‌های جدید و توسعه‌پذیری آن می‌باشد. مجموعه داده مورد استفاده در این مقاله، بات‌نت ISCX2014 است که توسط آزمایشگاه CIC دانشگاه UNB کانادا جمع‌آوری شده است [23]. تنوع گونه های بات‌نت در این مجموعه داده بالا بوده و انواع جدیدی از آن در مجموعه آزمون وجود دارد. از این مجموعه داده، کل ترافیک با پورت مقصد ۸۰ استفاده شده است. تعداد داده‌های مجموعه آموزش و آزمون به ترتیب برابر ۴۰۹۵۷ و ۱۸۱۷۰ می‌باشد که مجموعه داده آموزش دارای ۷ نوع بانت‌نت به حجم ۵/۳ گیگابایت و مجموعه آموزش دارای ۱۶ نوع به حجم ۸/۵ گیگابایت خواهد بود.

برای ارزیابی سیستم، ابتدا ما ۱۰ درصد از مجموعه آموزش(داده‌های برچسب دار) را برای یادگیری استفاده کردیم و بقیه نمونه‌ها بدون برچسب باقی ماندند. سپس طی فرآیند یادگیری فعال نیمه‌نظارتی نسبت به انتخاب نمونه‌های موثر برای برچسب‌گذاری اقدام کرده و در نهایت سیستم را با بات‌های جدید مجموعه آزمون مورد آزمایش قرار دادیم و نتایج را بدست آوردیم. تعداد تکرار مراحل به منظور برچسب‌گذاری توسط خبره ۵۰ مرتبه در نظر گرفته شد. دقت سیستم با استفاده از معیار صحت رده‌بندی[2] بر اساس ماتریس درهم‌ریختگی ایجاد می‌شود (جزییات آن در جدول(۵) و نحوه محاسبه آن در رابطه‌ی (۴) آورده شده)، به همراه معیارهای معروف نرخ تشخیص(بات‌نت) و امتیاز F گزارش شده است.

جدول (۵) : ماتریس درهم‌ریختگی

		کلاس تخصیص یافته توسط مدل	
		مثبت	منفی
کلاس واقعی	مثبت	مثبت حقیقی (TP)	منفی کاذب (FN)
	منفی	مثبت کاذب (FP)	منفی حقیقی (TN)

$$Accuracy = \frac{TN + TP}{TN + FN + TP + FP} \quad (۴)$$

۴-۲- نتایج و مقایسه

نتایج حاصل از روند آموزش و مقایسه راهکار پیشنهادی با روش یادگیری با ناظر در جدول (۶) آورده شده است. همان‌گونه که مشاهده می‌شود مجموعه ویژگی Qiu2017 دقت رده‌بندی، نرخ تشخیص بات‌نت و معیار F بالاتری را نسبت به سایر مجموعه ویژگی‌ها به خود اختصاص داده است. در واقع از آنجایی که این مجموعه ویژگی تنها از طریق اندازه و جهت ارسال بسته‌ها نسبت به ساخت بردارهای ویژگی اقدام کرده است، نمایانگر مطلوبیت این راهکار برای ارائه‌ی مجموعه ویژگی‌ها در رده‌بندی جریان‌های ترافیکی می‌باشد. همچنین در مقایسه با روش یادگیری با ناظر رده‌بندهای لجستیک و ماشین بردار پشتیبان خطی، روش پیشنهادی توانست با تعداد محدودی از نمونه‌های برچسب خورده و کاهش هزینه‌های برچسب گذاری در کاربردهای واقعی، نتایجی بهتری نسبت به آن داشته باشد. از سوی دیگر در بین پژوهش‌های مشابه (مجموعه داده یکسان) این روش توانست با نرخ تشخیص درست ۸۹ درصد نسبت به مقاله

[1] Symmetrical Uncertainty

[2] Correctly classified instances (Accuracy)

[3] True positive rate (Sensitivity)

[23] با نرخ تشخیص درست ۷۵ درصد، وضعیت بهتری را ارائه دهد. اما از جمله مشکلات این سیستم می‌توان به افزایش نرخ مثبت نادرست به دلیل افزوده شدن باتنت‌های جدید به مجموعه آموزشی، در طول زمان اشاره کرد که انتظار می‌رود با ترکیب ویژگی‌های وزن‌دار و فرآیندهای پیش‌پردازشی دیگر، این موضوع را بهبود بخشید.

جدول (۶) : نتایج پیاده‌سازی و مقایسه مجموعه ویژگی‌ها

نام مجموعه ویژگی	دقت رده‌بندی فعال نیمه نظارتی (سیستم پیشنهادی)	معیار امتیاز F	نرخ تشخیص (باتنت)	دقت رده بند لجستیک با ناظر	دقت رده‌بند svm خطی با ناظر
Milcom2015	۸۲,۸۲	۰,۸۰	۰,۸۲	۸۱,۱۵	۸۱,۳۶
CIC2018	۸۱,۲۴	۰,۷۲	۰,۸۱	۸۱,۲۰	۸۰,۲
Li2009	۶۶,۴	۰,۶۶	۰,۶۶	۶۵,۰۱	۶۶,۱
ISCX2014	۸۱,۲۶	۰,۷۳	۰,۸۱	۸۱,۲۳	۸۱,۱
Qiu2017	۸۹,۸۵	۰,۸۵	۰,۸۹	۸۳,۰۱	۸۴,۱

۵- نتیجه

یک سیستم تشخیص باتنت قابل اعتماد باید در زمان مناسب و با کارایی بالا، نسبت به شناسایی جریان‌های مشکوک و حملات اقدام نماید. در این مقاله ما یک سیستم تشخیص باتنت بر مبنای یادگیری فعال نیمه‌نظارتی ارائه کردیم که قادر است نمونه‌های مناسب را برای یادگیری انتخاب نماید. مدل پیشنهادی با پنج مجموعه ویژگی مختلف بر روی یک مجموعه داده‌ی جامع مورد ارزیابی و مقایسه قرار گرفت. نتایج حاکی از کارایی روش پیشنهادی در تشخیص باتنت ها می‌باشد. از جمله کارهای آینده تغییر استراتژی‌های نمونه‌برداری و سناریو برای یادگیری فعال و رتبه‌بندی ویژگی‌های استخراج شده خواهد بود.

مراجع

[1] Dhote, Yogesh, Shikha Agrawal, and Anjana Jayant Deen. "*A survey on feature selection techniques for internet traffic classification*." Computational Intelligence and Communication Networks (CICN), 2015 International Conference on. IEEE, 2015.

[2] Shafiq, Muhammad, et al. "*Network traffic classification techniques and comparative analysis using machine learning algorithms*." Computer and Communications (ICCC), 2016 2nd IEEE International Conference on. IEEE, 2016.

[3] Gu, Guofei, et al. "*BotMiner: Clustering Analysis of Network Traffic for Protocol-and Structure-Independent Botnet Detection*." USENIX security symposium. Vol. 5. No. 2. 2008.

[4] Yahyazadeh, Moosa, and Mahdi Abadi. "*BotGrab: A negative reputation system for botnet detection*." Computers & Electrical Engineering 41 (2015): 68-85.

[5] Silva, Sérgio SC, et al. "*Botnets: A survey*." Computer Networks 57.2 (2013): 378-403.

[6] Qiu, Zhicong, David J. Miller, and George Kesidis. "*Detecting clusters of anomalies on low-dimensional feature subsets with application to network traffic flow data*." Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on. IEEE, 2015.

[7] Tsiatsikas, Zisis, et al. "*Hidden in plain sight. SDP-Based covert channel for botnet communication*." International Conference on Trust and Privacy in Digital Business. Springer, Cham, 2015.

[8] Saad, Sherif, et al. "*Detecting P2P botnets through network behavior analysis and machine learning*." Privacy, Security and Trust (PST), 2011 Ninth Annual International Conference on. IEEE, 2011.

[9] Wang, Ping, et al. "*A systematic study on peer-to-peer botnets*." Computer Communications and Networks, 2009. ICCCN 2009. Proceedings of 18th International Conference on. IEEE, 2009.

[10] Celik, Z. Berkay, et al. "*Salting public traces with attack traffic to test flow classifiers*." CSET. 2011.

[11] Sheikhpour, Razieh, et al. "*A survey on semi-supervised feature selection methods*." Pattern Recognition 64 (2017): 141-158.

[12] Settles, Burr. "*Active learning*." Synthesis Lectures on Artificial Intelligence and Machine Learning 6.1 (2012): 1-114.

[13] Menahem, Eitan, et al. "*ACTIDS: an active strategy for detecting and localizing network attacks*." Proceedings of the 2013 ACM workshop on Artificial intelligence and security. ACM, 2013.

[14] Qiu, Zhicong, David J. Miller, and George Kesidis. "*A maximum entropy framework for semisupervised and active learning with unknown and label-scarce classes*." IEEE transactions on neural networks and learning systems 28.4 (2017): 917-933.

[15] Zhao, Yue, Ciwen Xu, and Yongcun Cao. "*Research on query-by-committee method of active learning and application*." International Conference on Advanced Data Mining and Applications. Springer, Berlin, Heidelberg, 2006.

[16] Duda, Richard O., Peter E. Hart, and David G. Stork. *Pattern classification*. John Wiley & Sons, 2012.

[17] Bottou, Léon. "*Large-scale machine learning with stochastic gradient descent*." Proceedings of COMPSTAT'2010. Physica-Verlag HD, 2010. 177-186.

[18] Qiu, Zhicong, David J. Miller, and George Kesidis. "*Flow based botnet detection through semi-supervised active learning.*" Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017.

[19] Celik, Z. Berkay, et al. "*Malware traffic detection using tamper resistant features.*" Military Communications Conference, MILCOM 2015-2015 IEEE. IEEE, 2015.

[20] Moore, Andrew, Denis Zuev, and Michael Crogan. *Discriminators for use in flow-based classification*. 2013.

[21] Sharafaldin, Iman, A. Habibi Lashkari, and Ali A. Ghorbani. "*Toward generating a new intrusion detection dataset and intrusion traffic characterization*." Proceedings of fourth international conference on information systems security and privacy, ICISSP. 2018.

[22] Li, Wei, et al. "*Efficient application identification and the temporal and spatial stability of classification schema*." Computer Networks 53.6 (2009): 790-809.

[23] Beigi, Elaheh Biglar, et al. "*Towards effective feature selection in machine learning-based botnet detection approaches*." Communications and Network Security (CNS), 2014 IEEE Conference on. IEEE, 2014.