نشریه علمی پدافند غیرعامل

سال دوازدهم، شماره ۲، تابستان ۱۴۰۰، (پیاپی ۴۶): صص ۱۷-۱

علمی- ترویجی

مروری تحلیل ترافیک شبکه گمنام‌ساز پارس با استفاده از یادگیری ماشین

حامد همایون[۱]، مهدی دهقانی[۲*]، حمید اکبری[۳]

تاریخ دریافت: ۱۳۹۹/۰۲/۱۸
تاریخ پذیرش: ۱۳۹۹/۱۱/۲۵

چکیده

یکی از روش‌های تأمین امنیت و گمنامی، استفاده از شبکه‌های گمنام‌ساز می‌باشد. گمنام‌ساز پارس یکی از گمنام‌سازهایی است که توسط متخصصین کشورمان تولید شده است. یکی از نقاط ضعف شبکه‌های گمنام‌ساز، تفکیک‌پذیری و شناسایی ترافیک آن‌ها در میان دیگر ترافیک‌های شبکه است. تشخیص ترافیک عبوری از یک شبکه، به منزله تشخیص ماهیت آن ترافیک است و اگر این ترافیک، ترافیک یک گمنام‌ساز باشد به این معنی است که در شبکه اطلاعات محرمانه در حال رد و بدل شدن است و این به معنی خدشه وارد شدن به گمنامی است. یکی از معیارهای ارزیابی شبکه گمنامی، غیرقابل تفکیک بودن و غیرقابل شناسایی بودن ترافیک شبکه گمنامی از ترافیک عادی است. رده‌بندی ترافیک، یک روش بسیار قوی در داده‌کاوی است که کاربردهای فراوانی دارد. از جمله این کاربردها می‌توان به مدیریت ترافیک با استفاده از شناسایی ترافیک عبوری از شبکه اشاره نمود. در این تحقیق با استفاده از روش‌های داده‌کاوی، در گام اول، میزان تفکیک‌پذیری گمنام‌ساز پارس با ترافیک گمنام‌سازهای مسیریاب پیازی، پروژه اینترنت نامرئی، جاندو و ترافیک HTTPS، و در گام دوم و در یک بررسی عمیق‌تر، میزان تفکیک‌پذیری ترافیک چهار سرویس متفاوت درون گمنام‌ساز پارس مورد بررسی قرار گرفت. نتایج این آزمایش‌ها در گام اول، رده‌بندی با دقت ۱۰۰٪ و در گام دوم، دقت بالای ۹۵٪ را (با استفاده از الگوریتم جنگل تصادفی) نشان می‌دهد. علاوه بر آن، با رتبه‌بندی ویژگی‌های استفاده شده در هر آزمایش، میزان تأثیرگذاری این ویژگی‌ها بر دقت کل و زمان ساخت مدل بررسی شده است.

کلیدواژه‌ها: گمنامی، شبکه گمنام‌ساز، داده‌کاوی، رده‌بندی، یادگیری ماشین، تحلیل ترافیک

۱- کارشناس ارشد دفاع سایبری دانشگاه جامع امام حسین(ع)
۲- استادیار دانشگاه جامع امام حسین(ع)، (drdehghani@ihu.ac.ir)- نویسنده مسئول
۳- استادیار دانشگاه جامع امام حسین(ع)

۱- مقدمه

حریم خصوصی[1] یکی از حقوق بسیار مهم افراد در تمامی جوامع به حساب می‌آید. بسیاری از کشورهای دنیا برای آنکه حفظ حریم خصوصی افراد را دارای ضابطه نمایند، قوانینی را در این راستا وضع نموده‌اند و چون این قوانین ذیل حقوق بشر تعریف می‌گردد، معمولاً قوانین سخت‌گیرانه‌ای نیز به شمار می‌آید. گمنامی یکی از ارکان حریم خصوصی به ویژه در محیط اینترنت به شمار می‌آید که در اولین سال‌های به وجود آمدن اینترنت مطرح شد [۱] و رعایت آن توسط دولت‌ها و همچنین سرویس‌های خدمات‌رسانی امری ضروری است و شرکت‌ها و دولت‌ها متعهد به انجام آن هستند.

زیرساخت‌ها، ابزارها، سامانه‌ها و وب‌سایت‌های بسیاری در جهت ارائه خدمات گمنامی و حریم خصوصی به کاربران اینترنت، طراحی شده‌اند. هر کدام از این خدمات روش ویژه‌ای برای ایجاد گمنامی و جلوگیری از شناسایی کاربر و داده‌های عبوری دارند. به‌عنوان مثال برخی از این سرویس‌ها مانند سرویس‌های پیازی[2] صرفاً به‌عنوان یک بازپخش عمل کرده و ترافیک کاربر را پس از عبور از چند ایستگاه متفاوت به مقصد می‌رسانند. این سرویس‌ها بدون ثبت وقایع و دانستن فعالیت کاربر میزان گمنامی بالایی را برای کاربران خود فراهم می‌آورند. شبکه مسیریاب پیازی[3]، جاندو[4]، پروژه اینترنت نامرئی[5] و فری‌نت[6] چهار زیرساخت معروف و پیشرفته در زمینه گمنام‌سازی هستند.

تشخیص ترافیک عبوری از یک شبکه به منزله تشخیص ماهیت آن ترافیک است و اگر این ترافیک، ترافیک یک شبکه گمنام‌ساز[7] باشد به این معنی است که در شبکه اطلاعات محرمانه در حال رد و بدل شدن است. بسیاری از مدیران شبکه و حتی مسئولین دولتی نیاز دارند تا بدانند دقیقاً چه اطلاعاتی در حال رد و بدل شدن است [۲]. گرچه مطالعات زیادی در مورد شبکه‌های گمنام‌ساز صورت گرفته است، اما در حوزه تحلیل ترافیک شبکه‌های گمنام‌ساز کارهای زیادی صورت نگرفته است. تشخیص ترافیک شبکه‌های گمنام‌ساز موضوعی است که اخیراً مورد توجه محققان قرار گرفته است. وظیفه اصلی شبکه‌های گمنام‌ساز پنهان کردن هویت فرستنده و گیرنده پیام و رمزگذاری پیام در طول مسیر است. سوال اینجاست که آیا خود ترافیک شبکه‌های گمنام‌ساز در محیط اینترنت قابل تفکیک از دیگر ترافیک‌ها هست یا خیر؟ در طی سال‌های اخیر آزمایش‌هایی صورت گرفته است که نتایج آن بیانگر این است که ترافیک برخی از شبکه‌های گمنام‌ساز قابل تفکیک و رده‌بندی[8] توسط ابزارها و روش‌های داده‌کاوی است. این میزان تفکیک و شناسایی برای هر شبکه گمنام‌ساز بر اساس طراحی و پیچیدگی‌های آن شبکه گمنام‌ساز متفاوت است. همچنین این تحقیقات بیان می‌دارد که یک شخص، خارج از یک شبکه گمنام‌ساز تا چه عمقی از اطلاعات موجود در جریان ترافیک یک شبکه گمنام‌ساز را کشف می‌کند.

رده‌بندی، یک روش بسیار قوی در داده‌کاوی است که کاربردهای فراوانی دارد. از جمله این کاربردها می‌توان به استفاده آن در برقراری امنیت، مدیریت ترافیک و منابع شبکه و کاربران [۵-۳]، مهندسی ترافیک و استفاده در تحقیقات اشاره نمود.

ترافیک شبکه‌های مسیریابی پیازی، جاندو، پروژه اینترنت نامرئی و فری‌نت مواردی هستند که در طی آزمون‌های مختلفی تحلیل و مورد رده‌بندی قرار گرفته‌اند، اما تحلیل ترافیک شبکه‌های گمنام‌ساز بومی مسئله‌ای است که تاکنون به آن پرداخته نشده است. طراحی شبکه‌های گمنام‌ساز، علمی نوپا در ایران است و طراحی و توسعه یک شبکه گمنام‌ساز بومی، بسیار نادر و در مقیاس‌های کوچک و آزمایشگاهی در حال اجراست. در میان شبکه‌های گمنام‌ساز بومی، شبکه گمنام‌ساز پارس برای این منظور انتخاب شده است. علت این انتخاب پیشرفته بودن این شبکه‌های گمنام‌ساز و قابل رقابت بودن آن‌ها با نمونه‌های جهانی است. در این تحقیق پس از معرفی کارهای مرتبط در بخش دوم، روش تحقیق را در بخش سوم به‌صورت کامل توضیح می‌دهیم و در چهارمین بخش با تشریح تمام مراحل آزمون به مقایسه نتایج می‌پردازیم و نتایج آن را با دیگر تحقیقات مورد مقایسه قرار می‌دهیم. در بخش پنجم و پایانی به نتیجه‌گیری و ارائه کارهای بعدی خواهیم پرداخت.

۲- کارهای مرتبط

بر اساس دانش و یافته‌های محقق، تاکنون گمنام‌ساز بومی در محیط اینترنت وجود نداشته و بر همین اساس تابه‌حال تحقیق داخلی درزمینه رده‌بندی جریان ترافیک یک گمنام‌ساز بومی صورت نپذیرفته است. اولین تلاش‌ها برای تحلیل ترافیک شبکه‌های گمنام‌ساز، از طریق ابزارهای شبیه‌سازی و یا در شبکه‌های خصوصی مجازی[9]، بر روی مسیریاب پیازی انجام شد.

در سال ۲۰۰۹، هرمان به یکی از مشکلات این حوزه پرداخته است [۶ و ۷]. آن تحقیق در زمینه فنّاوری‌های مختلف بهبود و

[1] Privacy
[2] Onion Services
[3] The Onion Router (TOR)
[4] Jondo (JonDonym)
[5] Invisible Internet Project (I2P)
[6] Freenet
[7] Anonymity Network
[8] Classification
[9] Virtual Private Network (VPN)

افزایش گمنامی در ابزارهای گمنام‌ساز تک لایه (OpenSLL و OpenVPN) و چندلایه (مسیریاب پیازی و جاندو) است. در آن تحقیق با نمونه‌برداری از ۷۷۵ وب‌سایت و حدود ۳۰۰ هزار نسخه‌برداری[1] ثبت شده، با پیشنهاد و استفاده از یک رده‌بند بیز چندجمله‌ای[2] که بر پایه اندازه بسته‌های عبوری از شبکه کار می‌کند، توانسته است در ابزارهای تک لایه با دقت حدود ۹۷٪ به موفقیت دست پیدا کند. همچنین نتایج آن تحقیق نشان می‌دهد که این روش برای ابزارهای چند لایه، چندان موفقیت‌آمیز نبوده است و دقت خروجی این روش برای مسیریاب پیازی و جاندو به ترتیب ۳٪ و ۲۰٪ بوده است.

در سال ۲۰۱۱، پانچنکو برای بهبود نقاط ضعف روش هرمان، روشی را معرفی کرده است که با استفاده از ویژگی‌های ترافیک مانند حجم، جهت، زمان و درصد بسته‌های ورودی و همچنین بهره‌گیری از رده‌بند ماشین بردار پشتیبان[3] نتایج قابل قبولی گرفته است [۸]. دادگان[4] آن تحقیق شامل ۱۵۵۰۰ نمونه از ۷۷۵ وب‌سایت تحقیق هرمان [۶] به همراه بیش از ۱ میلیون صفحه وب جدید می‌باشد. خروجی آن تحقیق نشان می‌دهد نرخ شناسایی، برای مسیریاب پیازی از ۳٪ تحقیق هرمان به ۵۵٪ و برای جاندو از ۲۰٪ تحقیق هرمان به ۸۰٪ افزایش یافته است. علاوه بر آن، در آن تحقیق، حدود ۴۰۰۰ نشانی وب از بین ۱ میلیون وب‌سایت برتر در الکسا[5] و حدود ۱۰۰۰ نشانی وب دیگر در دادگان استفاده شد که در این مورد خاص، نرخ شناسایی حدود ۷۳٪ به‌دست‌آمده است.

بارکر در سال ۲۰۱۱، در یک تحقیق، با استفاده از یک شبکه خصوصی مجازی، سعی کرده است محیطی آزمایشگاهی فراهم کند تا بتواند ترافیک HTTPS و مسیریاب پیازی را از یکدیگر تفکیک نماید. خروجی آن مقایسه، بر اساس این رده‌بندی صورت پذیرفت: الف) ترافیک HTTPS معمولی، ب) ترافیک HTTP در حال عبور از شبکه مسیریاب پیازی خصوصی، ج) ترافیک HTTPS در حال عبور از شبکه مسیریاب پیازی خصوصی. در آن تحقیق از سه الگوریتم رده‌بندی استفاده شده است و بر اساس نتایج آن تحقیق، ترافیک HTTP و HTTPS که از طریق مسیریاب پیازی در حال عبور است با نرخ بالای ۹۳٪ (با ۳/۷٪ نرخ مثبت کاذب) قابل شناسایی هستند [۹].

در سال ۲۰۱۲، الصباح ، با استفاده از یادگیری ماشین، آزمون‌هایی جهت شناسایی نوع برنامه‌هایی (مرور اینترنت[6]، جریان داده[7] و اشتراک‌گذاری فایل) کرده است که کاربران مسیریاب پیازی از آن‌ها استفاده کرده‌اند. الگوریتم‌های استفاده‌شده در آن تحقیق بیز ساده[8]، شبکه بیز[9] و درخت تصمیم[10] است. آن تحقیق در دو سطح برخط و غیربرخط به آزمون گذاشته شد که نتایج هر دو سطح به‌صورت روشن بیانگر این مسئله بود که نوع برنامه‌های کاربردی مورد استفاده در شبکه مسیریاب پیازی قابل شناسایی می‌باشد (دقت ۹۷/۸٪ برای آزمون غیربرخط و دقت ۹۱٪ برای آزمون برخط) [۱۰ و ۱۱].

المبید با استفاده از الگوریتم‌های رده‌بندی نظارت‌شده[11]، سعی در شناسایی کردن و تفکیک ترافیک شبکه مسیریاب پیازی (شامل ترافیک ضبط‌شده از مشاهده ۵ وب‌سایت برتر الکسا) و ترافیک HTTPS (شامل ترافیک ضبط‌شده از مشاهده ۱۰۰ وب‌سایت برتر الکسا) از یکدیگر داشت. در آن تحقیق که در سال ۲۰۱۵ انجام شده است از ۴۰ ویژگی جریان ترافیک برای استفاده در داده‌کاوی بهره برده است. در آن تحقیق از الگوریتم‌های بیز ساده، شبکه بیز، C4.5، جنگل تصادفی[12] و بردار پشتیبان استفاده شده است و نتایج آن تحقیق نشان‌دهنده نرخ بالای مثبت واقعی و نرخ پایین مثبت کاذب برای تمام سناریوهای طراحی شده برای آزمون و تمام رده‌بندها بوده است (به ترتیب ۹۹٪≤ و ۱٪≤ برای مثبت واقعی و مثبت کاذب) [۱۲].

اسپرینگال در سال ۲۰۱۵، دو آزمون جدید برای شناسایی ترافیک ارائه کرده است. آن دو آزمون مربوط به ترافیک‌های HTTP و SSH در گره[13] خروجی مسیریاب پیازی است. محقق در آن مقاله با تفکیک ارتباطات عادی (HTTP و SSH) از ارتباطاتی که بر اساس مسیریاب پیازی شکل گرفته‌اند، نوعی فیلترینگ هوشمند ارتباط محور به‌جای فیلترینگ آی‌پی محور پیشنهاد داده است. در این آزمون از ویژگی‌های تأخیر و زمان رفت و برگشت[14] برای تشخیص نوع ترافیک استفاده شده است. نتایج آن تحقیق به خوبی نشان می‌دهد که ترافیک شبکه مسیریاب پیازی به خوبی قابل تفکیک از ترافیک عادی است. نرخ تشخیص آن آزمون‌ها برای مقایسه ترافیک مسیریاب پیازی با ترافیک SSH و HTTP (تفکیک ترافیک مسیریاب پیازی و غیر از آن) به ترتیب ۹۸/۹۹٪ و ۱۰۰٪ بود [۱۳].

خالد شهبار در سال ۲۰۱۴، با به‌کارگیری الگوریتم‌های بیز ساده، شبکه بیز، جنگل تصادفی و C4.5، و با به‌کارگیری

[1] Dump
[2] Multinomial Naive Bayes
[3] Support Vector Machine
[4] Data-set
[5] Alexa.com
[6] Browsing

[7] Data Stream
[8] Naive Bayes
[9] Bayesian Network
[10] Decision Tree
[11] Supervised
[12] Random Forest
[13] Node
[14] Round Trip Time (RTT)

ویژگی‌های متفاوت یعنی استخراج ویژگی‌های جریان ترافیک (۹۲ ویژگی)، سعی در شناسایی فعالیت‌های کاربر شده است. نتایج تحقیق بیانگر دقت بالای شناسایی (نزدیک به ۱۰۰٪) بود [۱۴-۱۶].

ایشان در سال ۲۰۱۵ به این سؤال که آیا می‌شود ترافیک روش انتقال پوششی[1] در مسیریاب پیازی را با استفاده از ویژگی‌های جریان ترافیک شناسایی کرد یا خیر، پاسخ می‌گوید [۱۵-۱۷]. شهبار به‌وسیله یک رده‌بند C4.5 ثابت می‌کند که مبهم‌سازی که روش انتقال پوششی ایجاد می‌کند، می‌تواند شکل محتوایی جریان ترافیک را متفاوت از حالت عادی مسیریاب پیازی نشان دهد، که در این صورت این نوع مبهم‌سازی بر روی مسیریاب پیازی نیز توسط یک رده‌بند، قابل تشخیص و تفکیک از دیگر ترافیک‌ها خواهد بود [۱۵، ۱۶ و ۱۸].

همچنین این محقق در همان سال به بررسی تأثیر اعمال اشتراک پهنای باند[2] در پروژه اینترنت نامرئی می‌پردازد. در آن تحقیق نویسنده با بررسی ترافیک ایجادشده به‌وسیله استفاده از اشتراک پهنای باند، میزان تفکیک ترافیک این شبکه را با دیگر ترافیک‌ها می‌سنجد. محقق با تولید دادگان دو ترافیک با اشتراک پهنای باند ۰٪ و ۸۰٪ و مقایسه آن‌ها به نتیجه‌گیری می‌پردازد. نتیجه آن تحقیق این بود که با افزایش اشتراک پهنای باند میزان تفکیک این ترافیک‌ها مشکل‌تر و پیچیده‌تر خواهد شد [۱۵، ۱۶ و ۱۹].

آنتونیو در سال ۲۰۱۸ با استفاده از یک دادگان معرفی‌شده در تحقیق شهبار و تحلیل آن با استفاده از روش‌های داده‌کاوی و رده‌بندی ترافیک، به نتایج قابل توجهی رسیده است [۲۰]. هدف آن تحقیق دریافتن این مسئله است که یک مشاهده‌گر از بیرون تا چه درجه‌ای می‌تواند یک ابزار گمنام‌ساز را تشخیص دهد و این تشخیص تا چه حدی دقیق است. در آن تحقیق از الگوریتم‌های بیز ساده، شبکه بیز، بیز چندجمله‌ای، C4.5 و جنگل تصادفی استفاده شده است و برای دقیق‌تر بودن تحقیق، آزمون‌ها در سه لایه انجام پذیرفته است. تعداد ویژگی‌های استفاده شده در داده‌کاوی در آن تحقیق ۷۴ عدد می‌باشد. نتایج نشان می‌دهد که با استفاده از رده‌بندها می‌توان با دقت بالای ۹۹٪ کلیت ترافیک شبکه‌های گمنام‌ساز را از ترافیک معمولی تشخیص داد و در لایه دوم می‌توان بالای ۹۰٪ نوع شبکه گمنام‌ساز را نیز تشخیص داد. همچنین در لایه سوم آن تحقیق، نشان داده می‌شود که با دقت بالای ۶۵٪ می‌توان نوع نرم‌افزار استفاده شده برای استفاده از شبکه گمنام‌ساز را نیز تشخیص دهیم.

در جدیدترین این تحقیقات، مقاله لی در تاریخ نوامبر ۲۰۱۸ بر روی اینترنت قرار گرفته است. گرچه تنها مقدمه آن مقاله در دسترس عموم قرار گرفته، اما می‌توان حدس زد که در آن مقاله نیز با استفاده از تولید دادگان مربوط به شبکه گمنام‌ساز فری‌نت، به رده‌بندی این ترافیک پرداخته است. در چکیده آن مقاله گفته شده است که یک مشاهده‌گر از بیرون می‌تواند به راحتی گره‌ها استفاده کننده از این گمنام‌ساز را شناسایی نماید بدون آن‌که نیاز باشد به این شبکه بپیوندد. همچنین در آن تحقیق، کارایی برخی الگوریتم‌های استفاده شده نیز مورد بررسی قرار گرفته است. نتیجه آن تحقیق نشان‌دهنده دقت بالای ۹۴٪ برای الگوریتم درخت تصمیم است [۲۱].

آنچه تاکنون می‌دانیم این است که هیچ شبکه گمنام‌ساز بومی تاکنون در دسترس عموم قرار نداشته و به تناسب، هیچ‌گاه از منظر شناسایی ترافیک آن‌ها مورد ارزیابی قرار نگرفته است. آنچه در این تحقیق انجام پذیرفته است، ارزیابی یک گمنام‌ساز بومی از منظر تفکیک‌پذیری از دیگر گمنام‌سازها و ترافیک HTTPS است و با ارزیابی و آزمون آن شبکه، به این سؤال پاسخ دادیم که آیا ترافیک شبکه گمنام‌ساز پارس نیز قابل تفکیک و رده‌بندی هست یا خیر و این رده‌بندی با چه دقتی انجام شده است. همچنین هیچ‌کدام از تحقیق‌های ذکر شده الگوریتم‌های استفاده شده را از منظر زمان ساخت مدل بررسی نکرده‌اند و تنها از معیارهای آماری برای مقایسه الگوریتم‌ها استفاده نموده‌اند. در این تحقیق الگوریتم‌های استفاده شده علاوه بر مقایسه‌های آماری از منظر زمان ساخت مدل نیز مقایسه شده‌اند و از این منظر نیز مورد تحلیل قرار گرفته‌اند.

۳- روش تحقیق

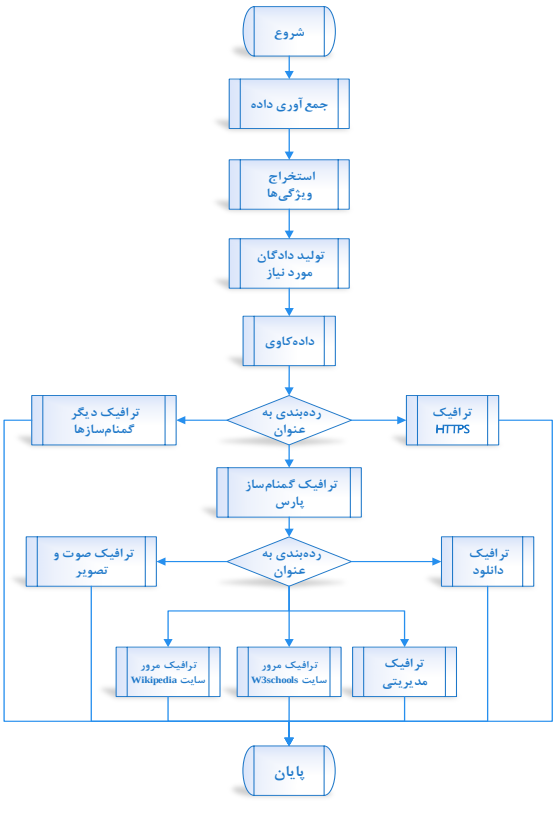
در این تحقیق ترافیک گمنام‌ساز پارس با ترافیک گمنام‌سازهای معروفی همچون شبکه پیازی، جاندو و پروژه اینترنت نامرئی و همچنین ترافیک عادی رمز شده HTTPS مقایسه شده است. برای انجام این آزمون باید از روش‌های داده‌کاوی به‌خصوص روش رده‌بندی استفاده کرد. الگوریتم‌های استفاده شده در رده‌بندی شامل ۵ الگوریتم بیز ساده، شبکه بیز، ماشین بردار پشتیبان، C4.5 و جنگل تصادفی می‌باشد. برای انجام آزمون رده‌بندی نیازمند دادگان مناسب می‌باشیم. با توجه به شکل (۱) کلیت انجام این تحقیق در دو مرحله انجام شده است. در مرحله اول، هدف تولید دادگان مناسب برای اجرای فرایند رده‌بندی است. برای انجام آزمون رده‌بندی نیازمند سه نوع دادگان می‌باشیم. نوع اول دادگان، دادگان مناسب برای گمنام‌ساز پارس است[3]. دادگان نوع دوم، دادگانی است که در تحقیق‌های قبلی مورد استفاده قرار گرفته است و با نام Anon17 مشهور است [۲۲]. دادگان نوع سوم، دادگان ترافیک HTTPS است. از آنجاییکه دادگان نوع اول

[1] Pluggable Transport (PT)
[2] Bandwidth Sharing

[3] دادگان گمنام‌ساز پارس دارای طبقه بندی است و با رعایت ضوابط دانشگاه جامع امام حسین(علیه السلام)، از بخش اسناد محرمانه کتابخانه باقرالعلوم و نیز از گروه علمی سایبری قابل دریافت و استفاده است.

و سوم مناسب برای این تحقیق موجود نیست، باید توسط محقق تولید شوند. ایـن دو نـوع دادگـان در مرحلـه اول تحقیـق در دو آزمایشگاه مجزا تولید شده است.

اما مرحله دوم این تحقیق، اجرای آزمون‌های رده‌بندی و تحلیل نتایج آن است که به‌وسیله ابزار داده‌کاوی Weka انجام شده است. اجرای آزمون رده‌بندی با هدف محاسبه تفکیک‌پذیری ترافیک‌ها در دو گام انجام می‌گیرد. گام اول، مقایسه ترافیک شبکه گمنام‌ساز پارس با گمنام‌سازهای معروف و ترافیک HTTPS است. در این آزمون فرض می‌شود که مهاجم، دادگانی از گمنام‌ساز پارس در اختیار دارد و می‌تواند با آن نرم‌افزار خود را آموزش بدهد. این مقایسه بدین منظور انجام می‌پذیرد که مشخص شود آیا ترافیک گمنام‌ساز پارس از دیگر گمنام‌سازها و ترافیک HTTPS قابل تفکیک و شناسایی است یا خیر؟ و این تفکیک‌پذیری به چه میزان و دقتی صورت می‌پذیرد. هدف از مقایسه ترافیک گمنام‌ساز پارس با ترافیک HTTPS، بررسی میزان عادی بودن ترافیک گمنام‌ساز پارس در محیط اینترنت است. از آنجایی که بسیاری از ترافیک‌های عبوری از یک شبکه، ترافیک مشاهده صفحات وب توسط کاربران است و بسیاری از صفحات وب از پروتکل HTTPS استفاده می‌کنند، این مقایسه می‌تواند میزان مشاهده‌پذیری ترافیک شبکه گمنام‌ساز پارس را در میان ترافیک عادی مرور اینترنت، بسنجد.



شکل (۱): روندنمای فرایند اجرای تحقیق

در گام دوم آزمون تفکیک‌پذیری، صرفاً به ترافیک گمنام‌ساز پارس پرداخته شده است. در این گام، میزان شناسایی و تفکیک‌پذیری ترافیک ابزارها و فعالیت‌های متفاوت در این گمنام‌ساز، بررسی شد. سه نوع متفاوت از ابزارهای کاربردی شامل استفاده از مرورگر برای مرور صفحات وب، دریافت فیلم و صوت به‌صورت جریان صوتی تصویری و اشتراک فایل می‌باشد. علاوه بر آن ترافیک مدیریتی خود گمنام‌ساز نیز به‌عنوان یک ترافیک مجزا در این آزمون بررسی شده است. در گام دوم این آزمون این سؤال نیز مطرح می‌شود که آیا ترافیک مرور یک پایگاه اینترنتی خاص با یک پایگاه دیگر متفاوت است یا خیر و آیا این ترافیک با ترافیک مرور عادی پایگاه‌های اینترنتی تفکیک‌پذیر است؟ برای این منظور دادگان ترافیک دو پایگاه اینترنتی W3schools و Wikipedia به‌صورت مجزا تولید می‌شوند.

برای مقایسه نتایج آزمون تفکیک‌پذیری از معیارهای صحت[1]، دقت[2]، بازخوانی[3]، مقیاس F[4] استفاده شده است که فرمول‌های آن‌ها به ترتیب (۱ تا ۴) آورده شده است. مهم‌ترین معیار سنجش، معیار دقت کل می‌باشد که این معیار به‌صورت خودکار توسط نرم‌افزار Weka محاسبه می‌گردد. همچنین برای یافتن میزان تأثیرگذاری ویژگی‌های استفاده شده در این تحقیق، پس از رتبه‌بندی ویژگی‌ها، تمامی آزمون‌ها با ویژگی‌های انتخابی، چندین و چند بار تکرار شد. نوع انتخاب ویژگی‌ها به‌صورت دسته‌های ده‌تایی (دسته آخر، ۱۲ ویژگی با پایین‌ترین رتبه تأثیرگذاری را شامل می‌شود)، سه ویژگی اول و سه ویژگی آخر است و هر آزمون با هر دسته از این ویژگی‌ها تکرار شد. همچنین تمامی آزمون‌ها تنها با استفاده از ویژگی برتر هر آزمون تکرار می‌شوند و نتایج آن‌ها ثبت می‌گردند.

$$\text{صحت} = \frac{\text{مثبت‌واقعی}+\text{منفی‌واقعی}}{\text{مثبت‌کاذب}+\text{منفی‌کاذب}+\text{مثبت‌واقعی}+\text{منفی‌واقعی}} \quad (۱)$$

$$\text{دقت} = \frac{\text{مثبت‌واقعی}}{\text{مثبت کاذب}+\text{مثبت واقعی}} \quad (۲)$$

$$\text{بازخوانی} = \frac{\text{مثبت واقعی}}{\text{منفی کاذب}+\text{مثبت واقعی}} \quad (۳)$$

$$\text{مقیاس F} = \frac{2*\text{بازخوانی}*\text{دقت}}{\text{بازخوانی}+\text{دقت}} \quad (۴)$$

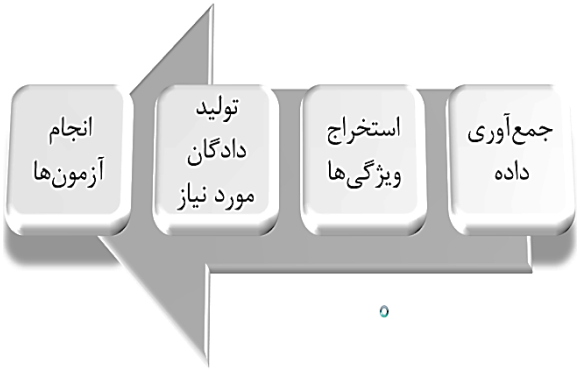
[1] Accuracy
[2] Precision
[3] Recall
[4] F-Measure

یکی دیگر از اهداف این تحقیق بررسی امکان جلوگیری از رده‌بندی و تفکیک‌پذیری ترافیک گمنام‌ساز پارس است. این موضوع برای بررسی این موضوع است که آیا می‌توان کاری کرد که ترافیک گمنام‌ساز در مقایسه با ترافیک گمنام‌ساز پارس را نتوان از دیگر ترافیک‌ها تشخیص داد یا خیر؟ این امر با شناسایی مهم‌ترین و تأثیرگذارترین ویژگی‌ها انجام گرفت و تلاش شد که با تغییر ویژگی‌های مهم میزان تأثیرگذاری آن‌ها در دقت نهایی محاسبه گردد.

برای مقایسه الگوریتم‌ها استفاده شده در این تحقیق از معیار دقت کل و همچنین زمان ساخت مدل بهره می‌بریم [۲۳]. معیار زمان ساخت مدل الگوریتم‌ها برای هر آزمون به‌صورت جداگانه محاسبه شد و در تحلیل نهایی از آن بهره‌برداری شد. این معیارها به محقق کمک خواهد کرد تا تحلیل بهتری از الگوی رفتاری هر الگوریتم ارائه دهد.

۴- اجرای آزمون و نتایج

برای این تحقیق به سه دسته دادگان احتیاج است. دسته اول دادگان مناسب برای گمنام‌ساز پارس است. دسته دوم، دادگان مناسب از دیگر گمنام‌سازهای معروف است و دسته سوم، یک دادگان مناسب از ترافیک HTTPS است. واضح است که دادگان دسته اول باید تولید گردد. دادگانی که برای دسته دوم انتخاب شده، دادگانی است که در تحقیق‌های شهبار مورد استفاده قرار گرفته است. این دادگان به نام Anon17 معروف می‌باشد و در این تحقیق دیگر دادگان به‌گونه‌ای تولید می‌گردند که بتوان آن را با این دادگان مقایسه کرد. از آنجایی که هیچ دادگان مناسبی برای دسته سوم به‌منظور مقایسه با دو دسته دیگر وجود نداشت، در این تحقیق دادگان مناسب برای دسته سوم نیز تولید شد. روش استفاده شده برای رسیدن به نتایج مورد نظر در این تحقیق در چهار اقدام کلی پیش‌بینی شده است که در شکل (۲) به‌صورت خلاصه آورده شده است.



شکل (۲): روش‌شناسی انجام این تحقیق

۴-۱- جمع‌آوری داده

گام اول برای رسیدن به نتیجه در این تحقیق جمع‌آوری داده می‌باشد. جمع‌آوری داده به معنای ذخیره داده‌های خام عبوری از شبکه بر اساس سناریوهای از پیش تعیین شده است. در این تحقیق با دو دسته جمع‌آوری داده مواجه هستیم. دسته اول جمع‌آوری داده از شبکه گمنام‌ساز پارس است که برای بررسی مشاهده‌پذیری آن، انجام می‌شود و این دادگان در یک آزمایشگاه اختصاصی و بدون اتصال به اینترنت انجام گرفت. به همین دلیل آماده‌سازی سه کارساز وب، فایل و جریان صوتی تصویری به‌صورت غیربرخط لازم است. برای آماده‌سازی کارساز وب، محتویات دو پایگاه اینترنتی Wikipedia.org و W3schools.com ذخیره و به‌صورت غیربرخط راه‌اندازی شد. برای این کار حدود ۲۰ هزار صفحه از این دو پایگاه جمع‌آوری و ذخیره شد. برای کارساز فایل حدود ۳۰۰ فایل با پسوندهای مختلف مجموعاً به حجم سه گیگابایت آماده و کارساز فایل با آن راه‌اندازی شد. در نهایت برای آخرین کارساز حدود ۱۲۰ فایل صوتی و تصویری، به حجم ۲ گیگابایت آماده و کارساز جریان صوتی و تصویری با آن راه‌اندازی شد.

۴-۱-۱- جمع‌آوری داده گمنام‌ساز پارس

به علت اختصاصی بودن گمنام‌ساز موردنظر و نبود منابع آن به‌صورت متن‌باز، باید جمع‌آوری دادگان در محل شرکت تولیدکننده آن صورت می‌پذیرفت. این گمنام‌ساز یک نوع ترافیک دیگر نیز تولید می‌کرد که ترافیک اختصاصی برای مدیریت رله‌های خود می‌باشد. این ترافیک به‌صورت عادی و همواره در حال تولید است و می‌توان از آن یک نوع دیگر دادگان نیز تولید کرد و نام آن را ترافیک مدیریتی نامید.

۴-۱-۲- جمع‌آوری داده HTTPS

دسته دوم جمع‌آوری داده از ترافیک HTTPS است که برای مقایسه با ترافیک گمنام‌ساز پارس تولید شد. لازم به ذکر است که دادگان متعددی در محیط اینترنت حول محور ترافیک HTTPS و یا SSL تولید شده است، ولی از آنجایی که این دادگان قابل استفاده در این آزمون نبوده‌اند، بنابراین دادگان مورد نظر برای این آزمون تولید شد. برای تولید دادگان HTTPS، اطلاعات عبوری از مرور دو پایگاه اینترنتی Wikipedia.org و W3schools.com ذخیره و در آزمون‌های بعدی از آن استفاده می‌شود. برای این موضوع همان صفحاتی که به‌صورت غیربرخط در دسته اول دادگان مورد استفاده قرار گرفته بود، در دسته دوم دادگان به‌صورت برخط مشاهده و داده‌های آن ذخیره شد. برای این دادگان آزمایشگاهی لازم است که به محیط اینترنت متصل باشد.

در هنگام جمع‌آوری داده‌های خام بر روی شبکه گمنام‌ساز پارس داده‌های ذخیره‌شده در هر دو آزمایشگاه، در همان نرم‌افزار Wireshark تصفیه و ترافیک‌های غیر مرتبط از میان آن‌ها حذف می‌شوند. دادگان خام تولیدشده در این مرحله، پس از پالایش اولیه به‌صورت فایل‌هایی با پسوند pcapng ذخیره شده‌اند. در مرحله بعدی، جریان‌های داده موجود در این فایل‌ها با استفاده از نرم‌افزار Tranalyzer [۲۴] استخراج شدند. هر چند تعداد جریان‌های داده هر فایل به حجم آن مربوط نمی‌باشد ولی حجم فایل‌های pcapng تولیدشده در این مرحله در جداول (۱ و ۲) آورده شده است.

جدول (۱): حجم داده‌های اولیه تولیدشده در آزمایشگاه غیربرخط

نوع دادگان	حجم فایل pcapng
مدیریتی	۳۰ مگابایت
مرور وب‌سایت Wikipedia	۶/۵ گیگابایت
مرور وب‌سایت W3schools	۱/۲ گیگابایت
جریان صوت و تصویر	۹/۳ گیگابایت
دریافت فایل	۳ گیگابایت

جدول (۲): حجم داده‌های اولیه تولیدشده در آزمایشگاه برخط

نوع دادگان	حجم فایل pcapng
مرور وب‌سایت Wikipedia	۱/۵ گیگابایت
مرور وب‌سایت W3schools	۳۰۰ مگابایت

۴-۲- استخراج ویژگی‌ها

از آنجایی که یکی از سناریوهای مد نظر این تحقیق مقایسه تفکیک‌پذیری دادگان گمنام‌ساز پارس با دادگان آماده چهار گمنام‌ساز دیگر است، باید این دادگان مشابه یکدیگر تولید شوند. روند تولید دادگان آماده چهار گمنام‌ساز مطرح‌شده در تحقیق آنتونیو بیانگر این موضوع است که تولید جریان ترافیک با استفاده از ابزار Tranalyzer انجام پذیرفته است. این ابزار یکی از قدرتمندترین ابزارهای شنود و ضبط ترافیک و همچنین استخراج ویژگی‌های جریان داده به‌صورت برخط است. این ابزار تحت سیستم عامل لینوکس کار می‌کند و علاوه بر توانایی استخراج اطلاعات به‌صورت برخط، توانایی استخراج جریان داده از فایل pcapng به‌صورت غیربرخط را نیز دارا می‌باشد. این ابزار دارای یک هسته مرکزی برای استخراج اطلاعات می‌باشد و برای استخراج اطلاعات بیشتر دارای افزونه‌هایی است که به‌صورت رایگان قابل استفاده می‌باشند. برای استخراج ویژگی‌های جریان داده از فایل‌های خام pcapng باید فایل‌ها را به‌صورت دستور ورودی[1] به ابزار Tranalyzer داد و خروجی مناسب را دریافت کرد.

ابزار Tranalyzer می‌تواند بیش از ۱۲۰ ویژگی از جریان داده را استخراج نماید. با بررسی‌هایی که در تحقیق آنتونیو انجام شده است، تعداد ۷۴ ویژگی خاص از این ویژگی‌ها برای یادگیری ماشین و انگشت‌نگاری انتخاب شده است. ویژگی‌هایی که از فهرست حذف شده‌اند شامل برخی ویژگی‌هایی بودند که تأثیری بر روند رده‌بندی ترافیک نداشته‌اند، مانند ویژگی‌های پروتکل ICMP و VLAN و یا برخی ویژگی‌هایی مربوط به حریم خصوصی مانند آی‌پی مبدأ و مقصد و یا درگاه‌های مبدأ و مقصد و یا زمان انجام آزمون. برخی دیگر از ویژگی‌هایی که حذف شده‌اند به خاطر تکراری بودن آن‌ها در دادگان نهایی است.

پس از استخراج تمامی جریان‌های داده از فایل‌های ترافیک ذخیره شده، تعداد جریان‌های استخراج شده به همراه ویژگی‌های آن‌ها از هر فایل pcap در جدول (۳) آورده شده است.

جدول (۳): تعداد نمونه‌های اولیه دادگان پیش از پالایش

نوع دادگان	تعداد
مدیریتی	۱۸۰۰
مرور وب‌سایت Wikipedia با گمنام‌ساز پارس	۸۰۰۰۰
مرور وب‌سایت W3schools با گمنام‌ساز پارس	۹۱۰۰۰
جریان صوت و تصویر با گمنام‌ساز پارس	۴۰۰۰۰
دریافت فایل با گمنام‌ساز پارس	۲۴۰۰
مرور وب‌سایت Wikipedia با پروتکل HTTPS	۱۵۵۰۰
مرور وب‌سایت W3schools با پروتکل HTTPS	۸۳۰۰

۴-۳- تولید دادگان مورد نیاز

پس از استخراج ویژگی‌های مذکور از جریان داده باید این فایل‌های خروجی را برای استفاده در برنامه Weka آماده‌سازی نمود. برای این منظور تغییراتی بر روی فایل خروجی به‌دست‌آمده نیاز است تا قابل‌استفاده در برنامه داده‌کاوی شود. برای این کار می‌توان خروجی برنامه Tranalyzer را با پسوند CSV ذخیره کرد و به‌عنوان ورودی به برنامه Weka داد. اما برای رسیدن به دادگان نهایی هنوز نیازمند برخی تغییرات روی این دادگان می‌باشد. از آنجایی که برنامه Weka برای پردازش داده‌ها، تنها داده‌های عددی و اسمی[2] را درک می‌کند، و از طرف دیگر چون برخی از مقادیر ویژگی‌ها نه عددی بودند و نه اسمی باید تغییرات زیر بر روی دادگان اولیه و پالایش نشده انجام پذیرد.

[1] Tranalyzer -r File.Pcapng -w Outputfolder/output.txt
[2] Nominal

۴-۳-۱- حذف برخی ویژگی‌ها

همان‌طور که قبلاً هم عنوان شد برخی ویژگی‌ها مانند TimeFirst و TimeLast به دلیل اینکه جزئی از ویژگی‌های جریان داده نمی‌باشند و تأثیری بر روال داده‌کاوی نخواهند داشت از دادگان اولیه حذف شده‌اند. همچنین ویژگی‌های dsMeanPL، dsMaxPL و dsMeanPL به دلیل اینکه دقیقاً مقادیر مشابه سه ویژگی دیگر بودند و عملاً تأثیری در روند رده‌بندی ندارند حذف شدند. این سه ویژگی به ترتیب مشابه minPktSz، maxPktSz و avePktSz می‌باشند. ویژگی‌های ipMinTTL و ipMaxTTL نیز حذف شدند.دلیل این کار این است که این دو ویژگی جزو ویژگی‌های ذاتی شبکه است و مقایسه این ویژگی در محیط آزمایشگاه غیربرخط با محیط واقعی اینترنت صحیح نمی‌باشد. در نهایت ۷۲ ویژگی به‌عنوان ویژگی‌های نهایی برای این آزمون انتخاب شدند که لیست این ویژگی‌ها به همراه توضیح مختصر هر ویژگی، در جدول (۹) (پیوست ۱) آمده است.

۴-۳-۲- تبدیل مقادیر رشته‌ای به اسمی

ابزار داده‌کاوی weka با دو نوع داده کار می‌کند، اسمی و عددی. بدیهی است مقادیری که جزو این دو دسته نباشد باید به این دو نوع تبدیل شوند تا بتوان آن‌را مورد تجزیه و تحلیل قرار داد. نرم‌افزار weka معمولاً این نوع داده‌ها را به‌عنوان رشته می‌شناسد و قابلیتی در خود نرم‌افزار وجود دارد که بتوان رشته‌ها را به داده اسمی تبدیل نمود. به‌عنوان مثال، نرم‌افزار weka اعداد در مبنای ۱۶ را به‌عنوان رشته می‌شناسد و برای تبدیل آن به مقادیر اسمی باید از فیلتر StringToNominal استفاده کرد[1].

۴-۳-۳- برچسب‌گذاری نهایی

در مرحله بعد باید هر دادگان برچسب مخصوص خود را داشته باشد تا در مرحله آموزش و آزمون از دیگر دادگان قابل تفکیک باشد. برای این کار ویژگی دیگری به نام TrafficType به هر دادگان اضافه و نوع ترافیک در آن لحاظ شده است.

۴-۳-۴- نمونه‌گیری از دادگان[2]

یکی از مشکلاتی که معمولاً در رده‌بندی و داده‌کاوی پیش می‌آید، عدم توازن دادگان‌های موجود با هم است. این عدم توازن به دلایل مختلف به وجود می‌آید. گرچه عدم توازن در مسائل داده‌کاوی مشکل بسیار جدی تلقی نمی‌شود، ولی برای نتیجه‌گیری بهتر و دقیق‌تر می‌توان از روش‌هایی برای توازن دادگان آزمون استفاده کرد. یکی از این روش‌ها نمونه‌گیری از دادگان بزرگ است. مثلاً در دادگانی که تولید شده است، تعداد نمونه‌های ترافیک مرور وب و ترافیک جریان صوت و تصویر به‌مراتب بیشتر از دیگر دادگان هستند. برای توازن بهتر می‌توان از این دادگان حدود ۵ الی ۱۰ درصد نمونه‌برداری کرد و از آن در آزمون استفاده نمود. همچنین می‌توان با حذف نمونه‌هایی که جریان داده آن‌ها دارای بسته‌هایی با اندازه صفر هستند، تعداد این نمونه‌ها را کم کرد.

این تغییرات باید بر روی دادگان Anon17 نیز اعمال گردد تا آماده استفاده در برنامه داده‌کاوی شود. پس از اعمال این تغییرات، بسته به نوع سناریوی مورد استفاده باید با ترکیب دادگان‌های تولید شده، به دادگان نهایی برای هر سناریو رسید. تعداد نمونه‌های نهایی دادگان تهیه‌شده برای استفاده در داده‌کاوی پس از پالایش و نمونه‌گیری در جدول (۴) آورده شده است.

جدول (۴): تعداد نمونه‌های نهایی دادگان بعد از پالایش

نوع دادگان	تعداد
مدیریتی گمنام‌ساز پارس	۱۸۰۰
مرور Wikipedia با گمنام‌ساز پارس (۱۰٪ نمونه‌گیری)	۵۲۰۰
مرور W3schools با گمنام‌ساز پارس (۱۰٪ نمونه‌گیری)	۵۴۰۰
جریان صوت و تصویر با گمنام‌ساز پارس	۳۱۰۰
دریافت فایل با گمنام‌ساز پارس	۲۴۰۰
مرور وب‌سایت Wikipedia با پروتکل HTTPS	۱۵۵۰۰
مرور وب‌سایت W3schools با پروتکل HTTPS	۸۳۰۰
دادگان مسیریاب پیازی	۵۲۰۰
مسیریاب پیازی با انتقال پوششی (۵٪ نمونه‌گیری)	۱۶۳۰۰
پروژه اینترنت نامرئی (۵٪ نمونه‌گیری)	۱۹۳۰۰
جاندو	۵۴۰۰

۴-۴- انجام آزمون

همان‌طور که ذکر شد رده‌بندی در دو گام انجام شده است. در گام اول ترافیک گمنام‌ساز پارس با ترافیک چهار گمنام‌ساز دیگر و ترافیک HTTPS مقایسه می‌گردد. این مقایسه دو به دو می‌باشد. یعنی ترافیک گمنام‌ساز پارس با تک تک دیگر ترافیک‌ها به‌صورت جداگانه و با پنج الگوریتم متفاوت مورد آزمایش قرار می‌گیرد. این الگوریتم‌ها عبارت‌اند از بیز ساده(NB)، شبکه بیز(BN)، جنگل تصادفی(RF)، C4.5 و ماشین بردار پشتیبان(SVM). برای تشابه آزمون‌ها، تمامی آن‌ها با روش یادگیری ماشین یکسان انجام شده است. اما در گام دوم مقایسه، تنها ترافیک دادگان مورد نظر مورد ارزیابی قرار می‌گیرد. در این ارزیابی سعی شده است تا انواع ترافیک تولیدشده توسط گمنام‌ساز پارس مورد تفکیک قرار گیرند.

برای یادگیری ماشین روش‌های مختلفی وجود دارد. در این

[1] Weka.filters.unsupervised.attribute.StringToNominal
[2] Weka.filters.unsupervised.instance.Resample

تحقیق روش اعتبار سنجی متقابل[1] با تعداد ۱۰ زیرمجموعه[2] برای این منظور انتخاب شده است. البته روش‌های دیگری در برنامه Weka برای یادگیری ماشین با دادگان آموزشی پیش‌بینی شده است که این روش، روش پیش‌فرض نرم‌افزار Weka می‌باشد. خروجی این آزمون‌ها در ادامه آورده شده است.

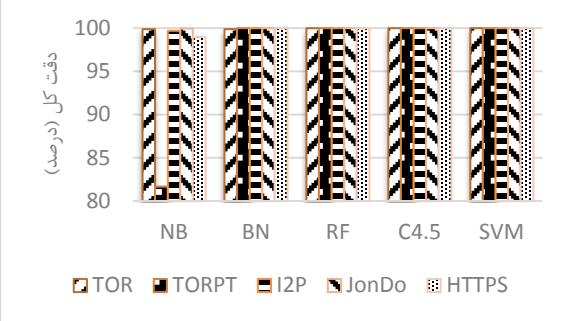
۴-۴-۱- نتایج گام اول

در گام اول رده‌بندی ترافیک گمنام‌ساز پارس با دیگر ترافیک‌ها به‌صورت دو به دو مقایسه شد. این آزمون‌ها با پنج الگوریتم متفاوت انجام شده است و خلاصه‌ای از نتایج این آزمون‌ها در جدول (۵) و شکل‌های (۳ و ۴) آورده شده است. در شکل (۳ و ۵) دسته ستون قابل مشاهده است که هر دسته نشان‌دهنده ۵ مقایسه انجام شده با یک الگوریتم است. این مقایسه‌ها، از چپ به راست بین دادگان گمنام‌ساز پارس با دادگان شبکه پیازی، انتقال پوششی، پروژه اینترنت نامرئی، جاندو و HTTPS انجام شده است. بهترین خروجی‌ها برای الگوریتم‌های جنگل تصادفی، ماشین بردار پشتیبان و C4.5 با دقت کل ۱۰۰٪ و مقیاس F ۱۰۰٪ بوده است. در چنین حالتی می‌توان گفت که ترافیک گمنام‌ساز پارس به طورکامل و بدون هیچ اشتباهی از دیگر ترافیک‌ها قابل تفکیک می‌باشد. در جدول (۵) درصد دقت کل و مقیاس F و همچنین زمان ساخت مدل برای ۵ مقایسه ذکر شده و با ۵ الگوریتم مختلف آورده شده است.

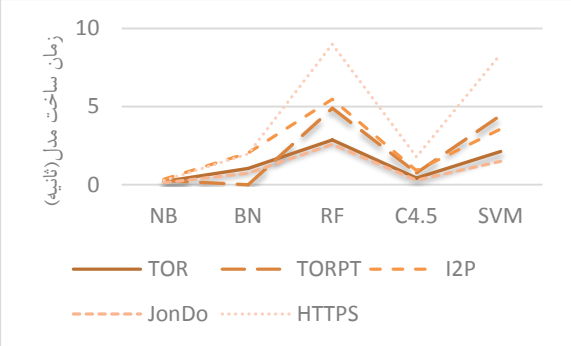
جدول (۵): نتایج گام اول آزمون رده‌بندی به درصد

	معیار	NB	BN	RF	C4.5	SVM
TOR	T-A[3]	۹۹/۸۶	۹۹/۴۹	۱۰۰	۱۰۰	۱۰۰
	F-M[4]	۹۹/۹۰	۹۹/۵۰	۱۰۰	۱۰۰	۱۰۰
	T[5]	۰/۲۱	۱/۰۴	۲/۸۷	۰/۴۴	۲/۱۴
TORPT	T-A	۸۱/۶۹	۱۰۰	۱۰۰	۹۹/۹۹	۹۹/۹۹
	F-M	۸۱/۲۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰
	T	۰/۲۷	۲	۴/۸۹	۰/۷۵	۴/۴۴
I2P	T-A	۹۹/۷۱	۹۹/۹۹	۱۰۰	۹۹/۹۹	۹۹/۹۹
	F-M	۹۹/۷۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰
	T	۰/۳۵	۲/۰۳	۵/۴۶	۰/۹۱	۳/۵۷
JonDo	T-A	۹۹/۹۵	۱۰۰	۱۰۰	۱۰۰	۱۰۰
	F-M	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰
	T	۰/۱۳	۰/۷۳	۲/۵۸	۰/۲۹	۱/۵
HTTPS	T-A	۹۸/۹۹	۹۹/۹۹	۱۰۰	۹۹/۹۹	۹۹/۹۷
	F-M	۹۹	۱۰۰	۱۰۰	۱۰۰	۱۰۰
	T	۰/۲۶	۱/۹۶	۸/۹۹	۱/۷۹	۸/۳۶

همچنین برای مقایسه بین الگوریتم‌ها علاوه بر درصد خروجی، زمان ساخت مدل نیز ثبت می‌گردد. قابل ذکر است که زمان ساخت مدل برای هر الگوریتم با پیچیدگی آن الگوریتم نسبت مستقیم دارد و این زمان سنجه خوبی برای انتخاب بهترین الگوریتم خواهد بود. شکل (۴) شمایی از نتایج به‌دست‌آمده از زمان ساخت مدل برای این آزمون‌ها را نشان می‌دهد. در این شکل (۵) نمودار متفاوت، نشان‌دهنده ۵ مقایسه ذکرشده با ۵ الگوریتم متفاوت است.



شکل (۳): دقت کل در گام اول آزمون رده‌بندی



شکل (۴): زمان ساخت مدل در گام اول آزمون رده‌بندی

طبق نتایج به‌دست‌آمده واضح و مشخص است که ترافیک گمنام‌ساز پارس به‌طور کامل از ترافیک چهار گمنام‌ساز دیگر و ترافیک HTTPS قابل تفکیک است. این نشان‌دهنده این است که اگر در شبکه‌ای که فقط ترافیک‌های مورد نظر در این آزمون وجود داشته باشد، یک مشاهده‌گر خارجی در هر جایی از شبکه می‌تواند با استفاده از فنون داده‌کاوی به راحتی ترافیک گمنام‌ساز پارس را از دیگر ترافیک‌ها تشخیص دهد. البته هدف از این آزمون تنها بررسی انگشت‌نگاری شدن ترافیک گمنام‌ساز پارس است و تفکیک‌پذیری دیگر ترافیک‌های استفاده شده در این آزمون از یکدیگر، موضوعی است که قبلاً در تحقیق‌های پیشین بررسی و با دقت‌های بالایی تفکیک‌پذیری آن‌ها اثبات شده است.

این یک نقطه ضعف برای گمنام‌ساز پارس به شمار می‌آید. ازآنجایی‌که یکی از موارد بسیار مهم امنیتی یک گمنام‌ساز در شبکه، پنهان ماندن گره‌های آن گمنام‌ساز است، می‌توان گفت که با این روش به راحتی گره‌های این گمنام‌ساز شناسایی خواهند شد.

[1] Cross-Validation
[2] Fold
[3] Total Accuracy
[4] F-Measure
[5] Time

از یک منظر دیگر، این تفکیک‌پذیری کامل می‌تواند یک نقطه مثبت برای این گمنام‌ساز باشد. اگر فرض کنیم که شرکت‌های امنیتی به دنبال جلوگیری از فعالیت گمنام‌سازها باشند و با استفاده از داده‌کاوی بخواهند ترافیک این گمنام‌سازهای معروف را شناسایی کنند، گمنام‌ساز پارس می‌تواند در عین ایجاد گمنامی مناسب، از دید این شرکت‌ها به دور باشد.

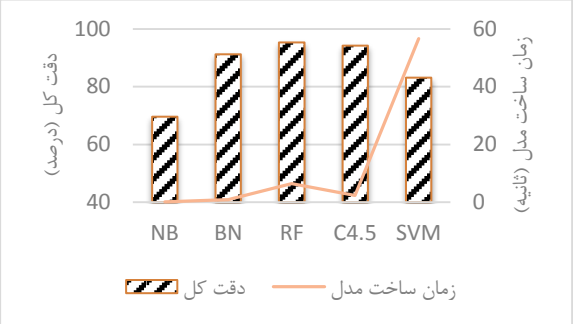
۴-۴-۲- نتایج گام دوم

اما در گام دوم رده‌بندی، بر روی ترافیک دادگان مورد نظر تمرکز می‌شود. پنج نوع ترافیک متفاوت برای گمنام‌ساز پارس تولید شده است. ترافیک مدیریتی، ترافیک بارگیری و اشتراک فایل، ترافیک جریان صوت و تصویر، ترافیک مشاهده وب‌سایت ویکی‌پدیا و ترافیک مشاهده وب‌سایت W3schools. در این آزمون سعی می‌شود تا یک لایه عمیق‌تر به ترافیک گمنام‌ساز پارس نگاه شود و این موضوع سنجیده شود که یک مشاهده‌گر از بیرون تا چه حد می‌تواند درک کند تا فردی که از گمنام‌ساز پارس استفاده می‌کند، در حال چه کاری است. این موضوع، با طرح سه موضوع پرکاربرد و متفاوت انجام می‌پذیرد. ترافیک مشاهده وب‌سایت، ترافیک استفاده از جریان صوت و تصویر و در نهایت ترافیک انتقال فایل و بارگیری. در این آزمون حتی این موضوع پیش‌بینی شده است که مرور وب‌سایت‌های متفاوت چه تأثیری در مشاهده‌پذیری ترافیک گمنام‌ساز پارس دارد و آیا مشاهده هر وب‌سایت قابل انگشت‌نگاری می‌باشد یا خیر؟ برای این منظور، ترافیک مرور وب‌سایت به دو بخش تقسیم شده است و دو وب‌سایت معروف و نسبتاً پرکاربرد برای این آزمون مد نظر قرار گرفته شده است. نتایج آزمون انجام گرفته در جدول (۶) و شکل (۵) آورده شده است. در جدول (۶)، درصد دقت کل و مقیاسF و همچنین زمان ساخت مدل به تفکیک نوع الگوریتم ثبت شده است و شمایی از این اطلاعات را می‌توان در شکل شماره ۵ مشاهده کرد.

نتایج به‌دست‌آمده برای این آزمون بیانگر تفکیک‌پذیری بالای انواع فعالیت‌های انجام شده در داخل گمنام‌ساز پارس می‌باشد. دقت کل بالای ۹۵٪ برای الگوریتم جنگل تصادفی و دقت بالای ۹۰٪ برای دو الگوریتم شبکه بیز و C4.5 نشان از این دارد که رفتار یک کاربر در درون شبکه گمنام‌ساز پارس کاملاً قابل شناسایی است. این موضوع را می‌توان از تفکیک بسیار بالای دو ترافیک وب متفاوت، برای دو وب‌سایت Wikipedia و W3schools کاملاً متوجه شد.

جدول (۶): نتایج گام دوم آزمون رده‌بندی به درصد

معیار	NB	BN	RF	C4.5	SVM
T-A	۶۲/۶۹	۹۱/۱۹	۹۵/۲۸	۹۴/۱۹	۸۳/۰۶
F-M	۶۸/۵۰	۹۱/۱۰	۹۵/۳۰	۹۴/۲۰	۸۲/۷۰
T	۰/۱۵	۰/۸۶	۶/۴۵	۲/۳۲	۵۶/۶۹



شکل (۵): دقت کل و زمان ساخت مدل در گام دوم آزمون رده‌بندی

دقت کل بالای محاسبه‌شده در گام دوم آزمون‌ها این احتمال را قوی می‌کند که استفاده از هر نرم‌افزار کاربردی در شبکه گمنام‌ساز پارس به‌راحتی توسط مدیران آن شبکه و یا هر فرد خارج از این شبکه (در محیط اینترنت)، قابل شناسایی است همچنین تفکیک‌پذیری بالای دو ترافیک مرور وب‌سایت، این احتمال را قوی‌تر می‌کند که مرور هر وب‌سایتی توسط این گمنام‌ساز قابل تفکیک‌پذیری می‌باشد. این مسئله به این معنی است دیگر یک کاربر، گمنامی لازم را برای مرور وب‌سایت نخواهد داشت و افرادی (گروهایی) که یک وب‌سایت خاص را مرور می‌کنند به‌راحتی می‌توانند شناسایی شوند.

شناسایی نوع رفتار یک گره در درون گمنام‌ساز و حتی شناسایی مرور یک وب‌سایت خاص توسط یک گره می‌تواند باعث از بین رفتن هویت آن گره و بعضاً ممکن است باعث اعمال محدودیت برای آن گره توسط مدیر شبکه باشد. این موضوع برای گمنام‌ساز شبکه پیازی نیز وجود دارد و موضوع جدیدی نمی‌باشد و طی تحقیق‌های پیشین، تفکیک‌پذیری بالای ۹۹٪ را برای مرور پنج وب‌سایت با استفاده از شبکه پیازی به اثبات رسانیده است. نتایج به‌دست‌آمده در گام دوم آزمون برای مقایسه تفکیک‌پذیری انواع ترافیک استفاده از گمنام‌ساز پارس، به همراه زمان ساخت مدل برای هر الگوریتم در نمودار آمده است.

۴-۴-۳- بررسی میزان تأثیرگذاری ویژگی‌ها

در ادامه بررسی‌ها می‌توان میزان تأثیرگذاری هر ویژگی در فرآیند داده‌کاوی و تفکیک‌پذیری و نتیجه پایانی سنجید. برای این کار باید با استفاده از ابزار Weka تمام ویژگی‌های این آزمون را به ترتیب میزان اثر، رتبه‌بندی کرد. ابزار Weka این قابلیت را دارد

تا با استفاده از چند روش میزان تأثیرگذاری هر ویژگی در نتیجه نهایی را بسنجد و با بقیه ویژگی‌ها مقایسه نماید[۱]. در این مقاله از روش Ranker استفاده شده است. قابل ذکر است که شماره ویژگی‌ها بر اساس جدول (۹) می‌باشد.

نتایج این رتبه‌بندی برای آزمون مقایسه گمنام‌ساز پارس با مسیریاب پیازی نشان داد که ویژگی‌های شماره ۲۸، ۳۱ و ۲۹ بیشترین تأثیر و ویژگی‌های ۳، ۲۰ و ۲۱ کمترین تأثیر را در نتایج خروجی خواهد داشت. این رتبه‌بندی برای تمام آزمون‌ها در گام اول و دوم انجام شد و خلاصه نتایج آن در جدول (۷) آورده شده است.

جدول (۷): شماره ویژگی‌ها و رتبه‌های آن‌ها

در آزمون مقایسه با	رتبه ۱	رتبه ۲	رتبه ۳	رتبه ۷۰	رتبه ۷۱	رتبه ۷۲
TOR	۲۸	۳۱	۲۹	۳	۲۰	۲۱
TORPT	۲۹	۲۸	۳۱	۳	۲۱	۲۰
I2P	۲۹	۲۸	۳۱	۱	۲۰	۲۱
JonDo	۳۱	۲۸	۲۹	۳	۱۸	۲۱
HTTPS	۱۴	۴۳	۱۰	۳	۲۰	۱۷
گام دوم	۲	۱۴	۲۶	۱۷	۳	۸

برای بررسی بیشتر این موضوع، آزمون تفکیک‌پذیری را با کم کردن ویژگی‌ها از دادگان آزمون تکرار کردیم. این کم کردن به این صورت است که ویژگی‌های رتبه‌بندی شده به ترتیب رتبه، به هفت بخش تقسیم شده و آزمون هر بار فقط با ۱۰ ویژگی انجام شد. علاوه بر این تمامی آزمون‌ها نیز یک‌بار فقط با در نظر گرفتن اولین ویژگی برتر و بار دیگر با در نظر گرفتن سه ویژگی برتر نیز تکرار شده است تا نتیجه آن را در دقت کل و زمان ساخت مدل مشاهده نماییم. نتایج به‌دست‌آمده فقط با درصد دقت کل هر آزمون سنجیده شده است و این آزمون‌ها فقط با الگوریتم جنگل تصادفی انجام پذیرفته است. دلیل این انتخاب بالاترین دقت کل به‌دست‌آمده بین تمام آزمون‌های انجام شده با الگوریتم‌های دیگر است (در همه آزمون‌ها درصد دقت کل این الگوریتم ۱۰۰٪ بوده است). البته این انتخاب به معنای این نیست که این الگوریتم، بهترین الگوریتم از نظر بازدهی است. نتایج این آزمون در جدول (۸) آمده است.

این نتایج نشان می‌دهد که انتخاب ده ویژگی اول در بین بقیه گروه‌ها نتایج بهتری را حاصل می‌کند. ده ویژگی دوم و سوم نیز با اختلاف کمی در رتبه‌های بعدی می‌باشند. نکته جالب توجه این است که انجام آزمون با ده و سه ویژگی اول نتایج بسیار نزدیکی با آزمون‌هایی دارد که همه ویژگی‌ها در آن لحاظ شده است. این مسئله را می‌توان در شکل (۶) مشاهده نمود. در این شکل (۶) دسته ستون بیانگر ۵ آزمون گام اول به همراه آزمون گام دوم است که در هر دسته، نتایج کم کردن تعداد ویژگی‌ها آورده شده است.

جدول (۸): دقت کل با تعداد ویژگی‌های انتخابی به درصد

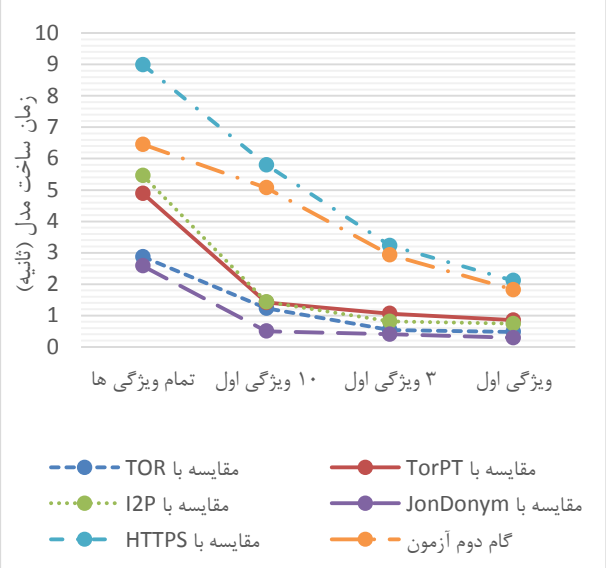
شماره ویژگی‌ها	TOR	TORPT	I2p	JonDo	HTTPS	گام دوم
۱	۹۹/۹۸	۹۹/۹۵	۹۹/۹۷	۱۰۰	۹۵/۸۶	۷۸/۵۲
۱ تا ۳	۹۹/۹۸	۹۹/۹۹	۱۰۰	۱۰۰	۹۹/۹۸	۹۳/۲۳
۱ تا ۱۰	۹۹/۹۸	۱۰۰	۱۰۰	۱۰۰	۹۹/۹۹	۹۴/۱۳
۱۱ تا ۲۰	۹۹/۹۴	۹۹/۹۸	۱۰۰	۹۹/۹۹	۹۹/۹۸	۹۲/۶۷
۲۱ تا ۳۰	۹۹/۹۲	۹۹/۸۹	۹۹/۹۶	۹۹/۹۷	۹۹/۹۹	۹۳/۲۶
۳۱ تا ۴۰	۹۹/۹۰	۹۹/۹۸	۹۹/۹۷	۹۹/۹۸	۹۹/۹۷	۸۳/۵۴
۴۱ تا ۵۰	۹۹/۹۹	۹۹/۹۳	۹۹/۹۸	۹۹/۹۴	۹۸/۶۸	۶۷/۱۸
۵۱ تا ۶۰	۹۹/۹۱	۹۹/۹۲	۹۹/۹۸	۹۹/۸۹	۸۶/۶۷	۷۸/۰۴
۶۱ تا ۷۲	۹۶/۶۹	۹۴/۰۱	۹۹/۸۸	۹۳/۳۳	۶۳/۳۵	۴۷/۵۵
۷۰ تا ۷۲	۷۷/۳۰	۵۲/۴۲	۵۲/۶۴	۷۳/۸۵	۵۷/۰۳	۳۰/۱۵

شکل (۶): دقت کل (درصد) با احتساب تعداد ویژگی‌های انتخابی

[۱] استفاده از روش Ranker در نرم‌افزار Weka

نتیجه جالب توجه دیگر انجام آزمون‌ها این‌بار فقط با اولین ویژگی و بار دیگر با سه ویژگی برتر در هر آزمون می‌باشد. موضوع مورد توجه علاوه بر نزدیکی بسیار زیاد دقت کل به‌دست‌آمده در استفاده از سه ویژگی برتر با آزمون با تمام ویژگی‌ها، زمان ساخت مدل آن‌هاست که در آزمون با سه ویژگی بسیار کاهش می‌یابد. این موضوع نشان‌دهنده این است که تفکیک ترافیک با استفاده از سه ویژگی برتر بسیار باصرفه‌تر از تفکیک ترافیک‌ها با استفاده از تمامی ویژگی‌هاست. خلاصه این نتایج در شکل (۷) مشاهده می‌شود.

نتیجه بسیار مهمی که از این آزمون می‌توان گرفت این است که تعداد ویژگی‌های منحصربه‌فرد و با تأثیرگذاری بالا در این دادگان بسیار زیاد است. زیرا با در نظر گرفتن ۱۰ ویژگی دوم، سوم، چهارم و پنجم نیز دقت کل محاسبه شده بسیار بالا و نزدیک به دقت کل با احتساب تمام ویژگی‌هاست. یکی از روش‌هایی که می‌توان ترافیک را از حالت تفکیک با دقت بالا خارج کرد، تغییر دادن هوشمندانه ویژگی‌هایی است که بیشترین تأثیرگذاری را در نتیجه تفکیک‌پذیری دارند (به‌عنوان مثال بتوانیم تعداد بایت‌های ارسالی را بیشتر کنیم). اما با توجه به این نتایج و تعداد بالای ویژگی‌های منحصربه‌فرد، تغییر الگوی ترافیک با استفاده از تغییر در ویژگی‌های جریان داده عملاً غیر ممکن است.



شکل (۷): زمان ساخت مدل با احتساب تعداد ویژگی‌های انتخابی

۴-۴-۴- مقایسه الگوریتم‌ها

یکی از نکات قابل توجه نمودار زمان ساخت مدل برای تمامی آزمون‌ها است. در همه آزمون‌ها (به‌جز آزمون گام دوم که تنها زمان ساخت مدل برای الگوریتم ماشین بردار پشتیبان بسیار زیاد شده است)، الگوی نمودار زمان ساخت مدل‌ها یکسان است و این نشان‌دهنده این مسئله است که زمان محاسبات الگوریتم‌ها با دادگان متفاوت با نسبت یکسانی صورت می‌پذیرد و در واقع الگوی رفتاری هر الگوریتم برای دادگان متفاوت، مشابه است. این موضوع نشان‌دهنده آن است که می‌توان با تحلیل بر روی زمان ساخت مدل برای هر الگوریتم، نتایج خوبی را برای مقایسه الگوریتم‌ها انجام داد.

در تمامی آزمون‌ها در گام اول دقت کل به‌دست‌آمده برای الگوریتم جنگل تصادفی ۱۰۰٪ محاسبه شده است. حتی در گام دوم آزمون نیز این الگوریتم بالاترین دقت کل را داشته است. این موضوع نشان‌دهنده این است که این الگوریتم برای مصارفی که نیاز به دقت حداکثری در آن می‌باشد بسیار مناسب است. البته مسئله زمان محاسبات یکی از نقطه‌ضعف‌های این الگوریتم است. این مسئله به‌خوبی در پنج آزمون گام اول قابل مشاهده است. گرچه دقت بالای این الگوریتم بسیار نتایج را قابل‌اتکاتر می‌کند ولی زمان صرف شده برای به نتیجه رسیدن این الگوریتم باعث شده است.

در طرف مقابل الگوریتم بیز ساده در تمامی آزمون‌های انجام شده کمترین دقت کل محاسبه‌شده را دارا می‌باشد. ولی زمان صرف شده برای ساخت مدل برای این الگوریتم در تمامی آزمون‌ها کمترین زمان ثبت شده است. بنابراین این الگوریتم برای مصارفی که نیازمند رسیدن به یک نتیجه نسبتاً خوب با صرف کمترین زمان ممکن (مانند محاسبات زمان واقعی[1]) است، پیشنهاد می‌گردد.

اما بررسی الگوریتم C4.5 می‌تواند نتایج جالبی را به همراه داشته باشد. این الگوریتم با میانگین دقت کل ۹۹٪ برای تمامی آزمون‌ها یکی از بهترین الگوریتم‌ها از نظر دقت محاسبه است و می‌تواند یک الگوریتم با اطمینان بالا برای تفکیک کردن ترافیک باشد. تفاوت میانگین دقت کل این الگوریتم با الگوریتم جنگل تصادفی تنها ۰/۲٪ است. همچنین زمان ساخت مدل برای این الگوریتم نیز بسیار کمتر از الگوریتم جنگل تصادفی و بسیار نزدیک الگوریتم بیز ساده است. این الگوریتم با صرف اندک زمان بیشتری نسبت به بیز ساده، به دقت بسیار بالاتری دست پیدا می‌کند. این موضوع نشان‌دهنده این مطلب است که الگوریتم ظ می‌تواند به‌عنوان بهترین الگوریتم از منظر دقت و سرعت برای این آزمون‌ها باشد.

الگوریتم شبکه بیز نیز عملکردی بسیار شبیه الگـوریتم C4.5 دارد. البته زمان ساخت مدل بیشتر و دقت کمتر این الگوریتم آن

[1] Real-Time

را پایین‌تر از الگوریتم C4.5 قرار می‌دهد. و در نهایت الگوریتم ماشین بردار پشتیبان با صرف زمان ساخت مدل بسیار زیاد و همچنین میانگین دقت کل به‌دست‌آمده متوسط می‌تواند در رده آخر انتخاب یک الگوریتم مناسب برای انجام این آزمون باشد. این الگوریتم در آزمون‌های گام اول نتایج مناسب و شبیه به جنگل تصادفی و حتی بهتر از آن داشت، اما در آزمون گام دوم نتوانست در زمان مناسب رده‌بندی مناسبی داشته باشد.

۵- نتیجه‌گیری

در گام اول، مقایسه گمنام‌ساز پارس از منظر تفکیک‌پذیری با چهار گمنام‌ساز دیگر انجام شد و در گام دوم تفکیک‌پذیری ترافیک ابزارهای استفاده شده در داخل گمنام‌ساز پارس مورد بررسی قرار گرفتند. در گام اول با دقت ۱۰۰٪ تمام نمونه‌های ترافیک گمنام‌ساز پارس درست رده‌بندی شدند و این موضوع بیانگر مشاهده‌پذیری کامل این ترافیک در بین چهار ترافیک دیگر است. در گام دوم نحوه رفتار یک کاربر و نوع ترافیک تولیدی وی در داخل گمنام‌ساز پارس مورد بررسی قرار گرفت. بدین منظور چهار نوع رفتار در درون این گمنام‌ساز تعریف شد و ترافیک آن در آزمایشگاه تولید شد. ترافیک مدیریتی، ترافیک مرور وب، ترافیک بارگیری و ترافیک جریان صوت و تصویر رفتارهایی بودند که برای یک کاربر فرض شدند. نتایج آزمون دوم نیز نشان‌دهنده تفکیک‌پذیری بالای ۹۵٪ برای این آزمون بود. این مسئله نشان‌دهنده آن است که علاوه بر قابل شناسایی بودن ترافیک گمنام‌ساز پارس، نوع رفتار یک کاربر داخل این شبکه نیز توسط یک نظاره‌گر از بیرون قابل شناسایی است.

همچنین در یک آزمون دیگر میزان شباهت ترافیک گمنام‌ساز پارس را با دیگر ترافیک‌های آزمایش‌شده در این تحقیق بررسی شد. این آزمون با این فرض انجام شد که سامانه‌های نظارتی هنوز نتوانسته‌اند ترافیک این شبکه‌گمنام‌ساز را آشکارسازی و ثبت نمایند و بالطبع هنوز نتوانسته‌اند هیچ نمونه دادگانی را در اختیار داشته باشند و نتایج این آزمون تا زمانی که این گمنام‌ساز در دسترس عموم قرار بگیرد اعتبار دارد. آزمایش‌ها نشان می‌دهد که از مجموع حدود ۱۷۰۰۰ نمونه دادگان موجود برای گمنام‌ساز پارس، اکثر این نمونه‌ها به‌عنوان ترافیک HTTPS شناسایی شده‌اند. بیشترین میزان این شباهت برای الگوریتم C4.5 با حدود ۹۴٪ و کمترین این شباهت برای الگوریتم بیز ساده با ۶۷٪ ثبت شده است.

در ادامه به بررسی الگوریتم‌های استفاده شده در این آزمون پرداخته شده است. در بین پنج الگوریتم بیز ساده، شبکه بیز، جنگل تصادفی، C4.5 و ماشین بردار پشتیبان، الگوریتم جنگل تصادفی دارای بیشترین دقت و بیشترین زمان ساخت مدل و الگوریتم بیز ساده دارای کمترین دقت و همچنین کمترین زمان ساخت مدل می‌باشد. الگوریتم C4.5 نیز با دقتی نزدیک به دقت الگوریتم جنگل تصادفی و زمانی بسیار کم یک الگوریتم خوب برای استفاده در این ارزیابی‌ها باشد.

همچنین این آزمون‌ها با کم کردن تعداد ویژگی‌های جریان داده چند باره تکرار شد. در یک رتبه‌بندی، ویژگی‌های استفاده شده در هر آزمون به‌صورت جداگانه رتبه‌بندی شدند و این ویژگی‌ها در ۷ دسته ۱۰ تایی گروه‌بندی شده و آزمون‌ها با هر دسته تکرار شدند. نتایج نشان‌دهنده این موضوع بود که استفاده از ۱۰ ویژگی برتر هر آزمون نتایجی مشابه استفاده از تمام ویژگی‌ها دارد. علاوه بر این آزمون‌ها با سه ویژگی برتر هر آزمون نیز تکرار شدند و دقت کل محاسبه شده به همراه زمان ساخت مدل برای این سه ویژگی با مقدار کل محاسبه شدند. خروجی آن نشان‌دهنده این بود که حتی استفاده از سه ویژگی نیز می‌تواند ما را به نتایج دلخواه برساند با این تفاوت که زمان ساخت مدل برای آزمون‌ها با سه ویژگی بسیار پایین‌تر از زمان ساخت مدل با تمام ویژگی‌هاست.

کارهای بعدی

موارد ذیل را می‌توان برای جهت‌گیری کارهای تحقیقاتی بعدی ذکر کرد:

- مقایسه گمنام‌ساز پارس با ترافیک‌های بیشتر: به‌منظور بررسی بیشتر تفکیک‌پذیری این گمنام‌ساز می‌توان آن را با ترافیک‌های بیشتری مقایسه نمود. در این تحقیق، این مقایسه با چند گمنام‌ساز دیگر و ترافیک HTTPS انجام پذیرفت و مسلماً ترافیک‌های متعددی برای انجام آزمون وجود دارد. همچنین می‌توان ترافیک‌های بیشتری در گام دوم این آزمون تولید کرد تا میزان تفکیک‌پذیری انواع بیشتری از سرویس‌های تحت گمنام‌ساز مشخص شود.
- انجام این آزمون با دادگان برخط گمنام‌ساز پارس: دادگان گمنام‌ساز پارس در این آزمون در یک آزمایشگاه غیربرخط انجام پذیرفت. برای درک بهتر میزان تفکیک‌پذیری این گمنام‌ساز می‌توان تولید دادگان این گمنام‌ساز را در محیط برخط انجام داد.
- تحلیل ویژگی‌های مؤثر بر تفکیک‌پذیری جریان داده مورد نظر: می‌توان با تغییر و ایجاد اختلال در ویژگی‌های مؤثر بر میزان تفکیک‌پذیری، تغییرات در خروجی رده‌بند را مشاهده و آن را با نتایج این آزمون مقایسه نمود و ضعف تفکیک‌پذیری ترافیک شبکه گمنام‌ساز موردنظر را برطرف نمود.

۶- مراجع

[1] A. Pfitzmann and M. Hansen, “Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management–A Consolidated Proposal for Terminology,” Fachterminologie Datenschutz und Datensicherheit, pp. 111–144, 2008.

[2] V. Paxson, “Bro: a System for Detecting Network Intruders in Real-Time,” Computer Networks, pp. 2435–2463, 1999.

[3] “Bro intrusion Detection System-Bro Overview,” [Online]. Available: http://bro-ids.org. [Accessed 24 April 2019].

[4] “Snort-The de Facto Standard for Intrusion detection/prevention,” 14 August 2007. [Online]. Available: http://www.snort.org. [Accessed 18 April 2019].

[5] L. Stewart, G. Armitage, P. Branch, and S. Zander, “An Architecture For Automated Network Control of Qos over Consumer Broadband Links,” in Ieee International Region 10 Conference (Tencon 05), Melbourne, Australia, November 2005.

[6] D. Herrmann, R. Wendolsky, and H. Federrath, “Website Fingerprinting: Attacking Popular Privacy Enhancing Technologies with The Multinomial Naïve-Bayes Classifier,” in Acm Workshop on Cloud Computing Security (Ccsw), pp. 31–42, 2009.

[7] D. Herrmann, “Online privacy: Attacks and Defenses,” it-Information Technology, vol. 57, no. 2, pp. 133-137, 2015.

[8] A. Panchenko, L. Niessen, A. Zinnen, and T. Engel, “Website fingerprinting in onion routing based anonymization networks,” ACM 10th annual Workshop on Privacy in the Electronic Society(WPES), pp. 103–114, 2011.

[9] J. Barker, P. Hannay And P. Szewczyk, “Using Traffic Analysis To Identify The Second Generation Onion Router,” in 9th Ieee/Ifip International Conference on Embedded and Ubiquitous Computing (Euc), pp. 72–78, 2011.

[10] M. AlSabah, K. S. Bauer, and I. Goldberg, “Enhancing Tor’s Performance Using Real-Time Traffic Classification,” in ACM Conference on Computer and Communications security (CCS), pp. 73–84, 2012.

[11] M. Alsabah and I. Goldberg, “Performance and Security Improvements for Tor: A Survey,” ACM Comput. Surv, vol. 49, no. 2, pp. 1-38, 2015.

[12] A. Almubayed, J. Atoum, and A. Hadi, “A Model for Detecting Tor Encrypted Traffic Using Supervised Machine Learning,” MECS, 2015.

[13] A. Springall, C. De Vito, and S.-H. S. Huang, “Per Connection Server-Side Identification of Connections Via Tor,” in IEEE 29th International Conference on Advanced Information Networking and Applications (AINA), pp. 727–734, 2015.

[14] K. Shahbar and N. Zincir-Heywood, “Benchmarking Two techniques for Tor Classification: Flow level and Circuit Level Classification,” in IEEE Symposium on Computational Intelligence in Cyber Security (CICS), pp. 1–8, 2014.

[15] K. Shahbar, Analysis of Multilayer-Encryption Anonymity Networks, Ph.D. Thesis, Dalhousie University Halifax, Nova Scotia, 2017.

[16] K. Shahbar and N. Zincir-Heywood, “Packet Momentum for Identificationof Anonymity Networks,” Journal of Cyber Security and Mobility, vol. 6, pp. 27-56, 2017.

[17] K. Shahbar and N. Zincir-Heywood, “Traffic flow Analysis of Tor Pluggable Transports,” in Ieee 11th International Conference on Network and Service Management(CNSM), pp. 178–181, 2015.

[18] K. Shahbar and N. Zincir-Heywood, “An analysis of Tor pluggable transports under adversarial conditions,” in Ieee Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2017.

[19] K. Shahbar And N. Zincir-Heywood, “Effects of Shared Bandwidth on Anonymity of The I2p Network Users,” Ieee Symposium on Security And Privacy, Workshop on Traffic Measurements For Cybersecurity (Wtmc), 2017.

[20] A. Montieri, D. Ciuonzo, G. Aceto, and A. Pescape, “Anonymity Services Tor, I2p, Jondonym Classifying In The Dark,” In Ieee Transactions on Dependable and Secure Computing, 2018.

[21] S. Lee, S. -H. Shin, and B. -H. Roh, “Classification of Freenet Traffic Flow Based on Machine Learning,” Journal of Communications, vol. 13, no. 11, pp. 654-660, 2018.

[22] K. Shahbar and N. Zincir-Heywood, “Anon17: Network Traffic Dataset of Anonymity Services,” Dalhousie University, Halifax, Canada, 2017.

[23] S. O. Akinola and O. J. Oyabugbe, “Accuracies and Training Times of Data Mining Classification Algorithms: An Empirical Comparative Study,” Journal of Software Engineering and Applications, pp. 470-477, 2015.

[24] S. Burschka and B. Dupasquier, “Tranalyzer: Versatile high performance network taffic analyzer,” IEEE Symposium Series ob Computatinal Intelligence (SSCI), pp. 1-8, 2016.

پیوست ۱:

در جدول (۹) عنوان ویژگی‌های نهایی استخراج شده همراه با شرح مختصری از هر ویژگی آورده شده است.

جدول (۹): عنوان ویژگی‌های نهایی استخراج شده همراه با شرح مختصر هر ویژگی

ردیف	مفهوم	عنوان ویژگی استخراج شده
۱	جهت جریان ترافیک	dir
۲	مدت زمان شروع جریان تا اتمام جریان	duration
۳	شماره پروتکل لایه ۴	l4Proto
۴	تعداد بسته‌های ارسالی	numPktsSnt
۵	تعداد بسته‌های دریافتی	numPktsRcvd
۶	تعداد بایت‌های ارسالی	numBytesSnt
۷	تعداد بایت‌های دریافتی	numBytesRcvd
۸	حداقل اندازه بسته‌های لایه ۳	minPktSz
۹	حداکثر اندازه بسته‌های لایه ۳	maxPktSz
۱۰	متوسط اندازه بسته‌های لایه ۳	avePktSize
۱۱	بسته‌های ارسالی در هر ثانیه	pktps
۱۲	بایت‌های ارسالی در هر ثانیه	bytps
۱۳	جریان بسته‌هاناهمگونی [1]	pktAsm
۱۴	ناهمگونی جریان بایت‌ها	bytAsm
۱۵	در سرآیند آی‌پی Identification حداقل مقدار بخش	ipMindIPID
۱۶	در سرآیند آی‌پی Identification حداکثر مقدار بخش	ipMaxdIPID
۱۷	تعداد تغییرات زمان زنده ماندن در سرآیند آی‌پی	ipTTLChg
۱۸	نوع خدمت آی‌پی در مبنای ۱۶	ipTOS
۱۹	مجموع پرچم‌های سرآیند آی‌پی در مبنای ۱۶	ipFlags
۲۰	سرآیند آی‌پی در مبنای ۱۶Options مجموع مقادیر بخش	ipOptCpCl_Num
۲۱	سرآیند آی‌پیOptions تعداد مقادیر در بخش	ipOptCnt
۲۲	TCP شماره توالی بسته	tcpPSeqCnt
۲۳	اختلاف بایت‌های توالی فرستاده شده	tcpSeqSntBytes
۲۴	تعداد خطاهای شماره توالی	tcpSeqFaultCnt
۲۵	TCPهای بسته ACK تعداد	tcpPAckCnt
۲۶	دریافتی بدون خطا ACK تعداد بایت‌های با	tcpFlwLssAckRcvdBytes
۲۷	های همراه با خطا ACK تعداد	tcpAckFaultCnt

[1] Asymmetry

۲۸	TCP[1] در سرآیند اولین اندازه پنجره مؤثر	tcpInitWinSz
۲۹	TCP میانگین اندازه پنجره مؤثر در سرآیند	tcpAveWinSz
۳۰	TCP حداقل اندازه پنجره مؤثر در سرآیند	tcpMinWinSz
۳۱	TCP حداکثر اندازه پنجره مؤثر در سرآیند	tcpMaxWinSz
۳۲	تعداد تغییرات روبه بالای اندازه پنجره مؤثر	tcpWinSzDwnCnt
۳۳	تعداد تغییرات روبه پایین اندازه پنجره مؤثر	tcpWinSzUpCnt
۳۴	تعداد تغییر جهت اندازه پنجره مؤثر	tcpWinSzChgDirCnt
۳۵	در مبنای TCP۱۶ مجموع پرچم‌های پروتکل	tcpFlags
۳۶	در مبنای TCP۱۶ سرآیند[2] مجموع پرچم‌های غیرمتعارف	tcpAnomaly
۳۷	در مبنای TCP۱۶ سرآیند Options مجموع مقادیر بخش	tcpOptions
۳۸	TCP[3] حداکثر طول یک قطعه	tcpMSS
۳۹	TCP اندازه پنجره	tcpWS
۴۰	TCP سرآیند Options تعداد مقادیر در بخش	tcpOptCnt
۴۱	زمان سفر	tcpSSASAATrip
۴۲	زمان سفر	tcpRTTSseqAA
۴۳	ACK حداقل زمان سفر	tcpRTTAckTripMin
۴۴	ACK حداکثر زمان سفر	tcpRTTAckTripMax
۴۵	ACK میانگین زمان سفر	tcpRTTAckTripAve
۴۶	TCP وضعیت خطای ارتباط	tcpStates
۴۷	تعداد ارتباطات از آی‌پی مبدأ به دیگر میزبان‌ها	connSip
۴۸	تعداد ارتباطات از آی‌پی مقصد به دیگر میزبان‌ها	connDip
۴۹	تعداد ارتباطات بین آی‌پی مبدأ و آی‌پی مقصد	connSipDip
۵۰	چارک پایین طول بسته‌ها	dsLowQuartilePl
۵۱	چارک میانی طول بسته‌ها	dsMedianPl
۵۲	چارک بالای طول بسته‌ها	dsUppQuartilePl
۵۳	فاصله بین چارک‌های طول بسته‌ها	dsIqdPl
۵۴	حالت طول بسته‌ها	dsModePl
۵۵	محدوده طول بسته‌ها	dsRangePl
۵۶	استاندارد[4] طول بسته‌هاانحراف	dsStdPl
۵۷	طول بسته‌هاانحراف استاندارد قوی[5]	dsRobStdPl

[1] Effective
[2] Anomaly
[3] Segment
[4] Deviation
[5] Robust

۵۸	طول بسته‌هاچولگی [1]	dsSkewPl
۵۹	طول بسته‌هامازاد [2]	dsExcPl
۶۰	حداقل زمان ورود [3]	dsMinIat
۶۱	حداکثر زمان ورود	dsMaxIat
۶۲	میانگین زمان ورود	dsMeanIat
۶۳	چارک پایین زمان‌های ورود	dsLowQuartileIat
۶۴	چارک میانی زمان‌های ورود	dsMedianIat
۶۵	چارک بالای زمان‌های ورود	dsUppQuartileIat
۶۶	فاصله بین چارک‌های زمان‌های ورود	dsIqdIat
۶۷	حالت زمان‌های ورود	dsModeIat
۶۸	محدوده زمان‌های ورود	dsRangeIat
۶۹	انحراف استاندارد زمان‌های ورود	dsStdIat
۷۰	انحراف استاندارد قوی زمان‌های ورود	dsRobStdIat
۷۱	چولگی زمان‌های ورود	dsSkewIat
۷۲	مازاد زمان‌های ورود	dsExcIat

[1] Skewness
[2] Excess
[3] Inter-arrival Time (iat)

Pars Anonymity Network Traffic Flow Analysis using Machine Learning

H. Homayun, M. Dehghani* , H. Akbari

Abstract

One of the common network security and anonymity methods, is the use of anonymity networks. Pars Anonymity Network is a domestic anonymizer network, developed by Iranian specialists. One of the main weaknesses of anonymous networks is their traffic differentiation and recognition among other network traffic. Uncovering the traffic passing through a network, means recognizing the nature of that traffic, and if this traffic is the traffic of an anonymity tool, it means that confidential information is being exchanged in the network, which puts anonymity in danger. One of the evaluation criteria of anonymity networks, is undifferentiability and indistinguishability of anonymous network traffic from normal traffic. Traffic classification - which has various applications - is one of the most powerful methods in data mining. Traffic management via detecting network traffic flow, is considered as one of these applications. In this research, by using data mining techniques, in the first step the detection rate of Pars Anonymity Network is assessed in comparison to the Onion Router, Invisible Internet Project, JonDo and HTTPS traffics, and in the next step, the classification rate of four different services in the desired anonymizer is studied in more detail. Results suggest that the classification accuracy rate of these experiments in the first step is 100% and in the next step -with the use of Random Forest algorithm- is 95%. Furthermore, by evaluating the used specifications in every experiment, the effectiveness of these specifications regarding the overall accuracy and the model construction time is assessed.

Key Words: *Anonymity, Anonymity Network, Data Mining, Classification, Machine Learning, Traffic Analysis*

* Imam Hossein Comprehensive University (drdehghani@ihu.ac.ir) - Writer-in-Charge