نشریه علمی «علوم و فناوری‌های پدافند نوین»
سال یازدهم، شماره ۴، زمستان ۱۳۹۹؛ ص ۴۳۷-۴۲۹

«علمی- پژوهش»

کاهش هشدارهای سامانه‌های تشخیص نفوذ به کمک تعمیم ویژگی‌های حملات در حوزه داده‌کاوی چندبعدی

مهدی ملکی[۱*]، محمد لطفی[۲]

۱- استادیار، ۲- مربی، دانشگاه آیت ا... بروجردی

(دریافت: ۹۸/۱۲/۱۵، پذیرش: ۹۹/۰۵/۲۹)

چکیده

امروزه حجم حملات پیشرفته سایبری در حال افزایش است، لذا استفاده از سامانه‌های تشخیص نفوذ در شبکه‌ها امری اجتناب‌ناپذیر است. یکی از مشکلات عمده در استفاده این سامانه‌ها حجم زیاد هشدارهای تولیدشده سطح پایین است. در این مقاله یکی از روش‌های حوزه داده‌کاوی به نام استنتاج ویژگی محور، استفاده شده است. اساس این روش تعمیم داده‌های سطح پایین به مفاهیم سطح بالاست. با توسعه این راهبرد در حوزه حملات سایبری، حجم هشدارهای حسگرهای تشخیص نفوذ کاهش داده شده است. این کاهش نه‌تنها باعث اختلال در شناسایی حملات نمی‌شود بلکه با تمرکز بیشتر در ویژگی‌های مشترک حملات باعث افزایش دقت در تشخیص حملات خواهد شد. همچنین یکی از پایه‌های اساسی این روش، سلسله‌مراتب تعمیم است که برای ویژگی‌های مؤثر در حملات طراحی شده است. از نکات بارز دیگر این مقاله، ارائه یک روش شهودی مناسب در انتخاب ویژگی‌ها برای تعمیم است. برای ارزیابی روش پیشنهادی از مجموعه داده جدید CICIDS2017 استفاده شده است که کاستی‌های مجموعه داده‌های قبل خود را مرتفع نموده است. نتایج بیانگر کاهش هشدارها با نرخ ۹۹ درصد در پایین‌ترین سطح تعمیم و میانگین ۲۵ ٪ در سطوح دیگر تعمیم است. در کنار ترافیک نرمال ۱۴ نوع حمله مختلف شناسایی شده است که حمله Dos Hulk با فراوانی ۸,۱۶٪ بیشترین فراوانی و حمله Heartbleed با فراوانی ۰/۰۰۰۴٪ کمترین فراوانی را دارا بوده‌اند. از دیگر قابلیت‌های ارائه‌شده در روش پیشنهادی، امکان عملیات پردازش تحلیلی برخط و داده‌کاوی چندبعدی در فضای حملات سایبری به کمک حرکت در سطوح مختلف تعمیم است.

کلیدواژه‌ها: سامانه تشخیص نفوذ، تعمیم ویژگی‌ها، داده‌کاوی چندبعدی، پردازش تحلیلی برخط، حملات چندمرحله‌ای

The Reduction of Intrusion Detection Systems Alerts by Generalizing Attack Features in Multidimensional Data Mining Domain

M. Maleki*, M. Lotfi
Ayatollah Boroujerdi University
(Received: 05/03/2020; Accepted: 10/08/2020)

Abstract

The volume of advanced cyber attacks is increasing today; hence the use of intrusion detection systems in networks is inevitable. One of the major problems by using these systems are considered as the high volume of low-level alarms produced. In the present paper, one of the data mining techniques called Attribute-Oriented Induction is utilized. The basis of this approach is to generalize low-level data to high-level concepts. By the development of this strategy in the field of cyber attacks, the volume of intrusion detection alarms has been decreased. This reduction not only disrupts the detection of attacks but by more focusing on the common features of the attacks, it will increase the accuracy of detection. Moreover, one of the basic foundations of this method is a generalized hierarchy designed for effective attack features. Another highlight of this investigation is to provide an intuitive approach to selecting features for generalization. The new CICIDS2017 data set was employed to evaluate the proposed method, which overcame the shortcomings of its previous data set. In conclusion, the results show a 99% decrease in alarms at the lowest generalization level and an average of 25% at the other generalization levels. In addition to the normal traffic, 14 different attack types were identified, with the Dos Hulk attack being the most frequent with 8.16% and the Heartbleed attack having the lowest frequency 0.0004%. Other capabilities were offered in the proposed method include the possibility of online analytical processing and multidimensional data mining in cyber attack space by moving at different levels of generalization..

Keywords: Intrusion Detection System, Feature Generalization, Multidimensional Data Mining, OLAP, Multistage Attacks

*Corresponding Author E-mail: mahdi.maleki@abru.ac.ir

۱. مقدمه

امروزه به‌دلیل وابستگی شدید به سامانه‌های ارتباطی و کامپیوتری اثرات حملات سایبری بسیار زیاد است. به همین دلیل استفاده از سامانه‌های دفاعی چون ضد بدافزارها، دیواره‌های آتش و سامانه‌های تشخیص نفوذ[1] غیرقابل‌انکار است. حجم فزاینده ترافیک شبکه‌های کامپیوتری استفاده از چندین مؤلفه دفاعی را در قطعات مختلف شبکه ایجاب نموده است [۱]. افزایش سامانه‌های تشخیص نفوذ مسائل جدیدی را به دنبال دارند که از مهم‌ترین آن‌ها عبارت‌اند از : "مدیریت حجم هشدارهای تولیدشده و کاهش آن به صورتی‌که باعث مخفی شدن حملات نشوند و دوم کشف ارتباط و همبستگی بین هشدارها[2] به‌منظور کشف سناریوی حملات به‌ویژه در نوع حملات پیشرفته چندمرحله‌ای و حملات مخفیانه"[3]. سامانه‌های تشخیص نفوذ همکار[4] برای حل این مسائل ارائه‌شده‌اند [۲ و ۳]. در این سامانه‌ها مؤلفه‌های دفاعی در حوزه‌های اجرایی متفاوت با مشارکت هم هشدارهای لازم را تولید کرده و پس از پردازش و کشف ارتباطات بین هشدارها حملات هماهنگ سایبری مشخص می‌شوند. در کشف وابستگی‌ها این روش‌ها به دودسته تک‌بعدی [۴ و ۵] و چندبعدی [۶ و ۷] تقسیم می‌شوند در روش تک‌بعدی کشف وابستگی‌ها با محور قرار دادن یک ویژگی انجام می‌شود (مانند کشف همه هشدارهایی که از یک آدرس مبدأ تشکیل‌شده‌اند) این روش‌ها ساده می‌باشند اما قدرت تشخیص مناسبی را ارائه نمی‌دهند در مقابل روش‌های چندبعدی با تمرکز بر روی چندین ویژگی از قدرت تشخیص مناسبی برخوردارند ولی در عوض حجم محاسبات بخصوص در ترافیک‌های بالای شبکه بسیار سنگین است. برای مقابله با محاسبات سنگین روش‌های مختلف سعی در تبدیل هشدارهای خام به قالبی دارند که در عین سادگی حالت کلی حملات را حفظ کنند و هشدارهای کاذب را تقلیل دهند. در یک روش از گراف‌هایی با محوریت شبکه‌های نقشینه[5] برای کاهش هشدارها استفاده‌شده است [۸]. در روش دیگر از بین تمام ویژگی‌ها چندین ویژگی(شامل آدرس مبدأ، درگاهی مبدأ، درگاهی مقصد و پروتکل) بنا بر تحلیل و نقش آن‌ها در شناسایی حملات مختلف انتخاب‌شده و گراف(گراف مشبکه[6]) بر اساس ریشه آدرس مبدأ شکل‌گرفته و تعداد الگوهای مختلف در گره-های گراف محاسبه می‌شود سپس با توجه به مقادیر آستانه بخشی از هشدارها حذف می‌گردند [۹]. در رویکردی دیگر از الگوریتم کشف الگوهای تکراری[7] برای پیدا کردن هشدارها با تعداد تکرار مشخص در ویژگی‌ها استفاده شده است. مشکل عمده روش‌های فوق درکاهش هشدارها عدم توجه به ارتباطات منطقی و مفهومی بین ویژگی‌ها است که باعث عدم شفافیت لازم در خروجی این سامانه‌ها و قدرت تصمیم‌گیری کم در تشخیص حملات خواهد شد. یکی از روش‌های تعمیم داده‌ها[8] در داده‌کاوی عبارت‌اند از: "تبدیل داده‌های سطح پایین به مفاهیم سطح بالاتر و حذف ویژگی‌های نامرتبط" . این عمل که به آن شرح مفهوم[9] هم گفته می‌شود، قدرت توصیف و تمایز[10] کلاس‌های مختلف داده را فراهم می‌آورد. این روش مهم در حوزه تعمیم داده‌ها روش استنتاج ویژگی محور[11] است [۱۰-۱۳]. اساس این روش، تعمیم ویژگی‌های با کمک سلسله‌مراتب موجود در بین آن‌ها است. پایگاه‌های داده رابطه‌ای با کمک پرس‌وجوها (توسط زبان‌هایی مانند SQL) داده‌های لازم برای داده‌کاوی را در روش استنتاج ویژگی محور فراهم می‌آورند [۱۴ و ۱۵]. پرس‌وجوهایی با ساختار SQL با هدف داده‌کاوی بر مبنای استنتاج ویژگی محور[12] ایجاد شده است [۱۶]. از استنتاج ویژگی محور در استخراج قوانین [۱۷]، رده‌بندی [۱۸]، خوشه‌بندی [۱۹] و در کشف الگوهای تکراری [۲۰] استفاده شده است. در این مقاله هدف کاهش هشدارهای سطح پایین با استفاده از سلسله‌مراتب موجود در بین ویژگی‌های حملات بوده است. حرکت در سطوح مختلف تعمیم در فضای حملات سایبری دستاورد ویژه این مقاله در کسب دانش در این حوزه هست.

سازمان‌دهی مقاله در ادامه به شرح زیر است:

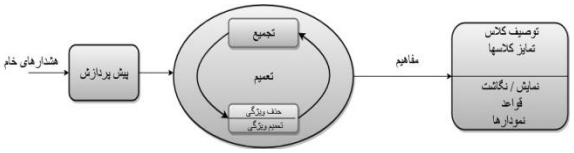
در بخش دو الگوریتم استنتاج ویژگی محور توضیح داده شده است. در ادامه روش پیشنهادی مبتنی بر الگوریتم فوق در حوزه حملات سایبری توسعه داده شده و متناسب با شبکه استفاده‌شده در مجموعه داده پیاده‌سازی شده است. در بخش نتایج، خروجی روش روی مجموعه داده ارائه‌شده نمایش یافته است. همچنین نتایج و قوانین تولیدشده مورد ارزیابی قرارگرفته است.

۲. روش تحقیق

روابط مفهومی بین سلسله‌مراتب موجود بین ویژگی‌ها این امکان را می‌دهد که با استخراج مناسب آن‌ها توصیف بهتر و درنتیجه درک کامل‌تری از هشدارها صورت پذیرد و به‌تبع آن تهدیدهای رخداده کشف شود. لذا از الگوریتم‌هایی در داده‌کاوی استفاده شده است که بتوانند با در نظر گرفتن روابط بین ویژگی‌ها هشدارها را تعمیم داده و حجم هشدارها را با حفظ روابط بین آن‌ها کاهش دهند. در این روش ابتدا داده‌های مرتبط با حوزه خاص توسط پرس‌وجوی مجموعه داده جمع‌آوری می‌شود سپس رکوردهای ناقص در مرحله پیش‌پردازش اصلاح می‌گردند. مرحله

[1] Intrusion Detection Systems (IDS)
[2] Alert Correlation
[3] Stealthy Attack
[4] Collaborative Intrusion Detection Systems (CIDS)
[5] Network Motif
[6] Lattice Graph
[7] Frequent Itemset Mining
[8] Data Generalization
[9] Concept Description
[10] Characterization & Discrimination
[11] Attribute-Oriented Induction
[12] Data Mining Query Language (DMQL)

تعمیم شامل حذف ویژگی[1] یا تعمیم ویژگی[2] است. سپس عمل تجمیع بر روی رکوردهای تکراری انجام شده و خروجی مفاهیم تولید می‌شود (شکل ۱). مفاهیم تولیدشده با نمایش‌ها و نگاشت‌های مختلفی چون نمودارها، قواعد و غیره قابل‌ارائه می‌باشند. در قسمت بعدی الگوریتم اصلی روش ارائه‌شده، توضیح داده‌شده است. همچنین روش پیشنهادی و پیاده‌سازی مؤلفه‌های اساسی آن بیان شده است.



شکل ۱. فرایند کلی روش ارائه‌شده بر مبنای استنتاج ویژگی محور

۲-۱. استنتاج ویژگی محور

این روش که یکی از روش‌های داده‌کاوی محسوب می‌شود مبتنی بر تعمیم داده‌ها با نظر گرفتن سلسله‌مراتب موجود بین ویژگی‌ها است. تعمیم داده از دو قسمت اصلی حذف ویژگی‌ها و تعمیم ویژگی‌ها تشکیل شده است:

حذف ویژگی‌ها: ستون‌هایی که در آن‌ها تعداد حالت‌های مختلف ویژگی‌ها از مقدار آستانه‌ای بیشتر می‌باشند در صورت وقوع یکی از شرایط زیر حذف می‌شوند:

۱. سلسله‌مراتب مفهومی[3] برای آن ستون موجود نباشد
۲. سلسله‌مراتب مفهومی برای این ستون در ستون‌های دیگر موجود باشد.

تعمیم ویژگی‌ها: اگر تعداد حالت‌های مختلف ویژگی‌ها یک ستون از یک مقدار آستانه بیشتر باشد در صورت وجود سلسله‌مراتب مفهومی برای آن ستون، ویژگی‌ها آن ستون تعمیم پیدا می‌کنند.

انتخاب مقادیر آستانه به ویژگی‌ها و حوزه کاربرد وابسته است. این مقادیر توسط فرد خبره و یا کاربر بسته به اینکه کدام ویژگی‌ها بیشتر و کدام کمتر نیاز به تعمیم دارند انتخاب می‌شود. این انتخاب، کنترل تعمیم ویژگی[4] نام دارد. این مقدار طوری تعیین می‌شود که ویژگی‌ها نه بیشتر از حد لازم تعمیم یابند[5] و نه ویژگی‌ها در سطوح پایین قرار بگیرند[6] در هر دو حالت قوانین تولید شده اطلاعات لازم و مفید را تولید نخواهند کرد. روش‌های مختلفی برای تعیین مقادیر آستانه وجود دارد که دو روش مهم عبارت‌اند از:

۱) کنترل آستانه تعمیم ویژگی[7]
۲) کنترل آستانه روابط تعمیم‌یافته[8]

در اولی مقادیر آستانه برای همه ویژگی‌ها و یا تک‌تک ویژگی‌ها تعریف می‌شوند، که در داده‌کاوی این مقادیر معمولاً بین ۲ تا ۸ هستند و مقدار خاص آن توسط خبره یا کاربر انتخاب می‌شود. در روش دوم مقدار آستانه برای تعداد قوانین (خروجی‌های تعمیم‌یافته) انتخاب می‌شود و معمولاً در سامانه‌های داده‌کاوی عددی بین ۱۰ تا ۳۰ هست. به‌عنوان‌مثال اگر بعد از تعمیم تعداد روابط خروجی از مقدار آستانه بیشتر بود تعمیم دوباره روی ویژگی‌ها انجام می‌شود. در الگوریتم ارائه‌شده روش اول استفاده شده است و با توجه به تعداد سطوح سلسله‌مراتب موجود در ویژگی‌ها این مقدار برای ویژگی‌ها آدرس (IP Source, IP Dest) در محدوده ۸ تا ۳ تغییر کرده تا خروجی‌های مختلف تعمیم حاصل شود. برای ویژگی‌ها دیگر به علت سطوح کمتر عدد ۲ انتخاب شده است.

در شکل (۲) سلسله‌مراتب مفهوم برای ویژگی موقعیت به تصویر درآمده است همین‌طور که از شکل مشخص است در تعمیم ویژگی موقعیت، جهت حرکت به‌صورت زیر است:

خیابان › شهر › استان › کشور › همه

شکل ۲. سلسله‌مراتب مفهومی برای ویژگی موقعیت مکانی

الگوریتم استنتاج ویژگی محور از قدم‌های زیر تشکیل شده است:

۱. ویژگی‌ها مرتبط به حوزه کاری از پایگاه داده استخراج می‌شود
۱-۱. در این مرحله مقادیر داده‌های معیوب اصلاح می‌شوند (پیش‌پردازش)
۲. مرحله مقدماتی تعمیم انجام می‌شود:
۲-۱. تعداد حالت‌های مختلف هر ویژگی در ستون‌ها محاسبه می‌شود
۲-۲. عمل حذف یا تعمیم ویژگی انجام می‌گردد و در صورت تعمیم مقدار هر ویژگی با مقدار تعمیم‌یافته جایگزین می‌گردد

[1] Attribute Removal
[2] Attribute Generalization
[3] Concept Hierarchy
[4] Attribute Generalization Control
[5] Overgeneralization
[6] Undergeneralization
[7] Attribute Generalization Threshold Control
[8] Generalized Relation Threshold Control

۳. عمل تعمیم در کل جدول پایگاه داده انجام می‌شود

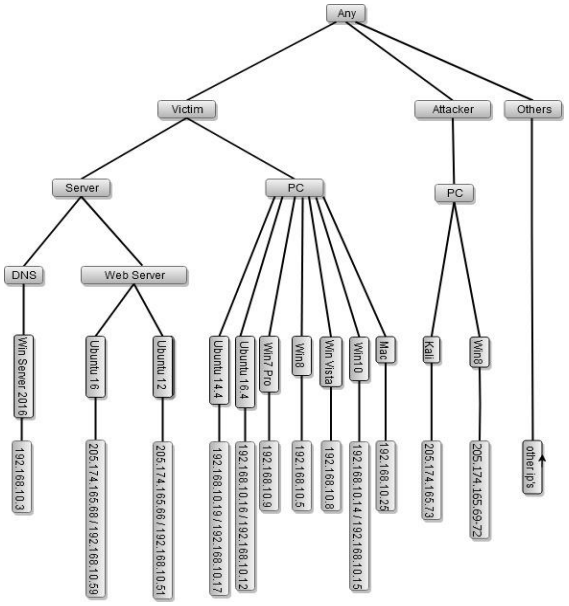
۳-۱. این کار با ادغام رکوردهای تکراری و محاسبه تعداد رکوردهای تکراری انجام می‌شود.

۴. اگر تعداد رکوردهای جدول به میزان موردنظر نرسیده(اطلاعات به میزان موردنظر تعمیم نیافته باشد) مراحل ۲ به بعد تکرار می‌گردند.

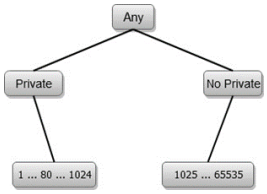
در قسمت بعدی الگوریتم اصلی مطابق با شبکه ارائه‌شده برای تشخیص حمله سایبری توسعه داده شده و همچنین سلسله‌مراتب مفهومی برای ویژگی‌ها موردنظر طراحی خواهد شد همچنین یک روش شهودی برای انتخاب ویژگی‌ها به‌منظور تعمیم تهیه شده است.

۲ -۲. روش پیشنهادی

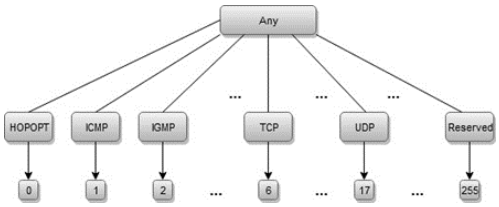
در این قسمت الگوریتم استنتاج ویژگی محور مطابق ضروریات تشخیص حملات سایبری تغییر داده شده است. مطابق مقاله‌ها و گزارش‌های اعلام‌شده در سایت‌های اعلام هشدار آسیب‌پذیری ویژگی‌ها ضروری برای تشخیص حملات عبارت‌اند از: "آدرس IP مبدأ، آدرس IP مقصد، درگاهی مبدأ، درگاهی مقصد و پروتکل" [۲۸-۲۱]. علاوه بر این ویژگی‌ها، ویژگی شمارش[1] و برچسب[2] نیز به داده‌ها اضافه شده است. ویژگی شمارش که با مقدار اولیه ۱ برای هر هشدار ایجاد می‌شود در فرآیند تعمیم افزایش یافته و به‌عنوان معیاری برای اندازه‌گیری اهمیت هشدار برای تصمیم‌گیرنده قابل‌استفاده است. ضمناً با کمک فیلد برچسب خروجی هشدارهای تعمیم‌یافته برحسب ترافیک عادی و ترافیک حملات(انواع حملات) قابل دسته‌بندی خواهند بود. بعد از استخراج اولیه داده‌ها برحسب ویژگی‌های بیان‌شده پیش‌پردازش اولیه‌ای روی‌داده‌ها انجام می‌شود، مثلاً رکوردهای ناقص حذف می‌شوند و یا بعضی از ویژگی‌ها (ویژگی شمارش) مقدار می‌گیرند. یکی از پایه‌های اصلی الگوریتم ارائه‌شده تولید ساختارهای لازم برای سلسله‌مراتب مفهوم برای ویژگی‌هایی است که چنین روابطی را دارا می‌باشند. طراحی چنین ساختارهایی که توسط خبرگان ایجاد می‌شوند بسیار به شبکه هدف مرتبط می‌باشند. این ساختارها به‌عنوان دانش پیش‌زمینه[3]، الگوریتم را در فرآیند تعمیم هدایت می‌کنند. در شکل‌های (۵-۳) ساختارهای درختی برای سلسله‌مراتب مفهوم برای آدرس‌های IP ، شماره درگاهی و پروتکل نمایش داده‌شده‌اند. این ساختارها مطابق شبکه هدف شکل (۶) طراحی شده‌اند که از شبکه هدف مجموعه داده CICIDS2017 اقتباس شده است [۲۹].



شکل ۳. سلسله‌مراتب مفهومی برای آدرس‌های IP



شکل ۴. سلسله‌مراتب مفهومی برای درگاهی



شکل ۵. سلسله‌مراتب مفهومی برای پروتکل

در ادامه الگوریتم پیشنهادی برای تعمیم هشدارهای سایبری که مبتنی بر روش استنتاج ویژگی محور است بیان شده است:

ورودی‌ها: مجموعه داده حملات D(r)، مقادیر آستانه (α) ، و سلسله‌مراتب مفهوم برای هر ستون G(c)

خروجی‌ها: GR(r)(Generalized Relation) روابط تعمیم‌یافته

الگوریتم:

۱) پیش‌پردازش D(r) برای حذف یا اصلاح ستون‌ها و فیلدهای لازم: W ← Preproccessing(D(r))

ستون‌های (IP+PORT+PROTOCOL) برای مرحله بعد انتخاب می‌گردند.

[1] Count
[2] Label
[3] Background Knowledge

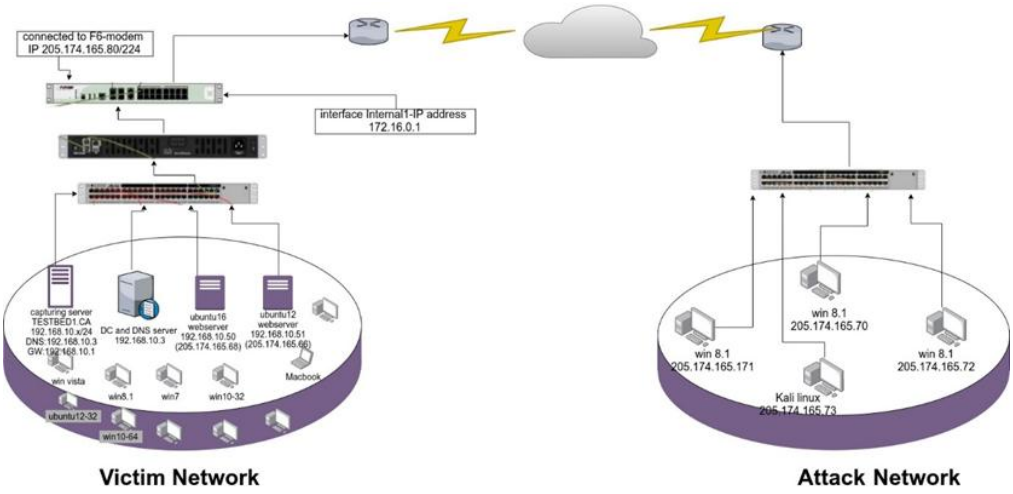
۲) اگر ویژگی قبلاً تعمیم نیافته باشد، برای تعمیم انتخاب می‌شود وگرنه انتخاب یک ویژگی برای تعمیم بر اساس بیشترین مقدار تابع شهودی H(c) است.

۳) تعمیم ویژگی‌ها ستون انتخاب‌شده در مرحله ۲: $W \leftarrow Genralization(W(c))$

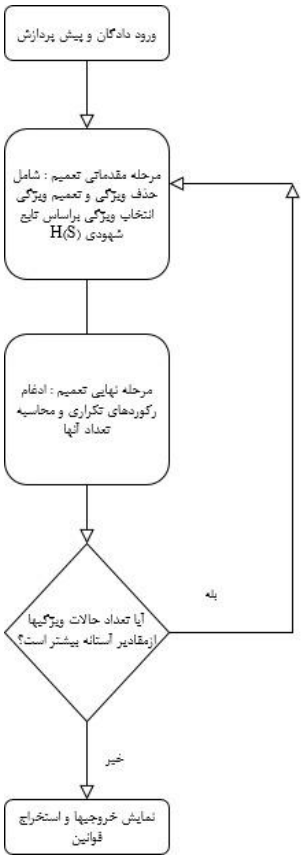
۴) تجمیع رکوردهای تکراری و مقداردهی در مجموعه داده جدید W : $W \leftarrow Aggregation(W)$

۵) اگر تعداد حالت‌های مختلف ویژگی‌ها و یا روابط تعمیم‌یافته از مقادیر آستانه α بیشتر است مراحل ۲ تا ۵ تکرار گردد در غیر این صورت $W \leftarrow GR(r)$

در روابط فوق r ردیف‌ها و c ستون‌ها است شکل (۷).



شکل ۶. ساختار شبکه مورد آزمایش، ویرایش شده از منبع [۲۹]



شکل ۷. فلوچارت پیشنهادی برای تعمیم هشدارهای سایبری مبتنی بر روش استنتاج ویژگی محور

۲-۲-۱. تابع شهودی H(c)

با بررسی و تحلیل هشدارهای سایبری و همچنین استخراج روابط موجود بین ویژگی‌ها اصلی هشدارها مانند آنچه در درخت‌های سلسله‌مراتب مفهومی نمایان شد متوجه می‌شویم سطوح سلسله‌مراتب در تمام ویژگی‌ها به‌طور یکسان تنظیم نشده است، مثلاً سلسله‌مراتب برای ستون آدرس‌های IP دارای بیشترین سلسله‌مراتب (سطح ۶) درحالی‌که سلسله‌مراتب برای ستون‌های درگاهی و پروتکل کمترین(سطح ۳) مقدار رادارند. لذا برای آنکه مقادیر تعمیم‌یافته شفافیت لازم را داشته باشند باید تعمیم را از ستون‌های IP شروع کنیم در غیر این صورت ستون‌هایی با سطح کمتر سریعاً به مقدار همه[1] تعمیم می‌یابند در حالی‌که ستون‌های با سطح بیشتر در مقادیر میانی قرار می‌گیرند و این در تابع شهودی حل شده است:

$$H(c)=MAX(Depth(H(c_i))) \quad (۱)$$

تابع Depth عمق درخت H(c) را می‌دهد و منظور از متغیر c_i ستون i ام یا ویژگی i ام است.

[1] Any

۳. نتایج و بحث

برای ارزیابی روش ارائه‌شده از مجموعه داده جدید CICIDS2017 استفاده شده است. مزایای این مجموعه داده رایگان نسبت به مجموعه داده‌های متداول DARPA، KDD'99، DEFCON، CAIDA و CDX حل مشکلاتی است که عبارت‌اند از[۳۰ تا ۳۴]: "عدم تناسب لازم بین کمیت حملات و کمیت ترافیک نرمال، دارا بودن بسته‌هایی با ویژگی‌های عمداً مخفی‌شده و نامشخص[1] که تحلیل حملات را با مشکل مواجه می‌کنند، محدودیت در انواع حملات بخصوص حملات جدید و عدم وجود مجموعه کاملی از ویژگی‌ها و ابر داده‌ها[2]. شبکه استفاده‌شده در این آزمایش مطابق شکل (۶) است که از دو قسمت اصلی شبکه قربانی و شبکه حمله تشکیل شده است. برای پیاده‌سازی طرح از زبان برنامه‌نویسی پایتون و کتابخانه pandas استفاده شده است[۳۵]. همچنین برای پیاده‌سازی درخت‌های سلسله‌مراتب از کتابخانه anytree بهره برده شده است [۳۶]. در شکل (۸) ساختار درختی پیاده‌سازی شده با ماژول anytree در پایتون نمایش داده شده است. این شکل درواقع پیاده‌سازی شکل (۳) به کمک ماژول فوق است. در جدول (۱) هشدارهای خروجی برنامه در سطوح مختلف تعمیم برای ۵ رکورد اول با بیشترین تکرار نمایش داده شده است. سطوح مختلف تعمیم[3] از GL0 بدون تعمیم تا GL5 بیشترین تعمیم انجام‌شده است. تعمیم بیشتر از این سطح اطلاعات جدیدی تولید نخواهد کرد. در خروجی GL0 اطلاعات با جزئیات زیاد موجود است حجم زیاد هشدارها (حدود ۲۸۳۰۰۰۰) و مقادیر عددی با تنوع زیاد تصمیم‌گیری سریع و مناسب را حتی برای متخصصین با مشکل مواجه می‌نماید درحالی‌که با تعمیم در سطوح بالاتر هم حجم هشدارها کمتر شده و هم تصویر مناسب‌تری از حملات با تعمیم مقادیر عددی به مقادیر حروفی بامعنی به‌دست آمده است. در جدول (۲) میزان کاهش هشدارها در سطوح مختلف نمایش داده شده است، همان‌طور که ملاحظه می‌شود GL1 بیشترین کاهش (۹۹٪) را داشته زیرا که هشدارهای تعمیم نیافته و سطح پائین که تعداد زیادی دارند(با اکثریت ترافیک نرمال) تعمیم‌یافته است در صورتی‌که در سطوح دیگر تعمیم، به‌دلیل آنکه عملیات روی هشدارهای تعمیم‌یافته با تعداد کمتر انجام شده است، نرخ کاهش کمتر است(۲۵٪). در جدول (۳) حملات کشف‌شده و نسبت ترافیک آن‌ها نمایش داده شده است.

[1] Anonymized Packet
[2] Metadata
[3] Generalization Level (GLx)

```
Node('/Any')
├── Node('/Any/Victim')
│   ├── Node('/Any/Victim/Server')
│   │   ├── Node('/Any/Victim/Server/DNS Server')
│   │   │   └── Node('/Any/Victim/Server/DNS Server/Win Server 2016')
│   │   │       └── Node('/Any/Victim/Server/DNS Server/Win Server 2016/192.168.10
│   │   └── Node('/Any/Victim/Server/Web Server')
│   │       ├── Node('/Any/Victim/Server/Web Server/Ubunto 16')
│   │       │   ├── Node('/Any/Victim/Server/Web Server/Ubunto 16/192.168.10.50')
│   │       │   └── Node('/Any/Victim/Server/Web Server/Ubunto 16/205.174.165.68')
│   │       └── Node('/Any/Victim/Server/Web Server/Ubunto 12')
│   │           ├── Node('/Any/Victim/Server/Web Server/Ubunto 12/192.168.10.51')
│   │           └── Node('/Any/Victim/Server/Web Server/Ubunto 12/205.174.165.66')
│   └── Node('/Any/Victim/PC_V')
│       ├── Node('/Any/Victim/PC_V/MAC')
│       │   └── Node('/Any/Victim/PC_V/MAC/192.168.10.25')
│       ├── Node('/Any/Victim/PC_V/Win 10')
│       │   ├── Node('/Any/Victim/PC_V/Win 10/192.168.10.14')
│       │   └── Node('/Any/Victim/PC_V/Win 10/192.168.10.15')
│       ├── Node('/Any/Victim/PC_V/Win Vista')
│       │   └── Node('/Any/Victim/PC_V/Win Vista/192.168.10.8')
│       ├── Node('/Any/Victim/PC_V/Win 8_V')
│       │   └── Node('/Any/Victim/PC_V/Win 8_V/192.168.10.5')
│       ├── Node('/Any/Victim/PC_V/Win 7pro')
│       │   └── Node('/Any/Victim/PC_V/Win 7pro/192.168.10.9')
│       ├── Node('/Any/Victim/PC_V/Ubunto 16.4')
│       │   ├── Node('/Any/Victim/PC_V/Ubunto 16.4/192.168.10.16')
│       │   └── Node('/Any/Victim/PC_V/Ubunto 16.4/192.168.10.12')
│       └── Node('/Any/Victim/PC_V/Ubunto 14.4')
│           ├── Node('/Any/Victim/PC_V/Ubunto 14.4/192.168.10.19')
│           └── Node('/Any/Victim/PC_V/Ubunto 14.4/192.168.10.17')
└── Node('/Any/Attacker')
    └── Node('/Any/Attacker/PC_A')
        ├── Node('/Any/Attacker/PC_A/Kali')
        │   └── Node('/Any/Attacker/PC_A/Kali/205.174.165.73')
        └── Node('/Any/Attacker/PC_A/Win 8_A')
            ├── Node('/Any/Attacker/PC_A/Win 8_A/205.174.165.69')
            ├── Node('/Any/Attacker/PC_A/Win 8_A/205.174.165.70')
            └── Node('/Any/Attacker/PC_A/Win 8_A/205.174.165.71')
```

شکل ۸. ساختار درختی سلسله‌مراتب آدرس پیاده‌سازی شده با ماژول anytree

جدول ۱. پنج رکورد اول در سطوح مختلف تعمیم از GL0 تاGL5

GL0: Generalization Level #0						
Source IP	Source Port	Destination IP	Destination Port	Protocol	Label	count
192.168.10.17	123	198.206.133.14	123	17	BENIGN	1
192.168.10.51	59108	162.213.33.50	443	6	BENIGN	1
192.168.10.51	59108	162.213.33.50	443	6	BENIGN	1
192.168.10.15	54012	205.174.165.73	8080	6	Bot	1
192.168.10.25	123	17.253.14.125	123	17	BENIGN	1
GL1: Generalization Level #1 (IP+PORT+PROTOCOL)						
Source IP	Source Port	Destination IP	Destination Port	Protocol	Label	count
Win Server 2016	No Private	Others	Private	UDP	BENIGN	295361
Others	No Private	Ubunto 16	Private	TCP	DoS Hulk	231073
Win 10	No Private	Win Server 2016	Private	UDP	BENIGN	144503
Win 10	No Private	Others	Private	TCP	BENIGN	135772
Ubunto 16.4	No Private	Others	Private	TCP	BENIGN	134845
GL2: Generalization Level #2 (IP)						
Source IP	Source Port	Destination IP	Destination Port	Protocol	Label	count
PC_V	No Private	Any	Private	TCP	BENIGN	692429
PC_V	No Private	DNS Server	Private	UDP	BENIGN	633037
Any	Private	PC_V	No Private	TCP	BENIGN	343405
DNS Server	No Private	Any	Private	UDP	BENIGN	295361
Any	No Private	Web Server	Private	TCP	DoS Hulk	231073
GL3: Generalization Level #3 (IP)						
Source IP	Source Port	Destination IP	Destination Port	Protocol	Label	count
Victim	No Private	Any	Private	TCP	BENIGN	692429
Victim	No Private	Server	Private	UDP	BENIGN	650963
Server	No Private	Any	Private	TCP	BENIGN	346649
Any	Private	Victim	No Private	UDP	BENIGN	343405
Any	No Private	Server	Private	TCP	DoS Hulk	231073
GL4: Generalization Level #4 (IP)						
Source IP	Source Port	Destination IP	Destination Port	Protocol	Label	count
Any	No Private	Any	Private	TCP	BENIGN	702616
Any	No Private	Victim	Private	UDP	BENIGN	651252
Victim	No Private	Any	Private	TCP	BENIGN	346800
Any	Private	Any	No Private	UDP	BENIGN	344348
Any	No Private	Victim	Private	TCP	DoS Hulk	231073
GL5: Generalization Level #5 (IP)						
Source IP	Source Port	Destination IP	Destination Port	Protocol	Label	count
Any	No Private	Any	Private	UDP	BENIGN	1746444
Any	Private	Any	No Private	TCP	BENIGN	417942
Any	No Private	Any	Private	TCP	DoS Hulk	231073
Any	No Private	Any	No Private	TCP	PortScan	133891
Any	No Private	Any	Private	TCP	DDoS	128024

جدول ۲. نرخ کاهش هشدارها در سطوح مختلف تعمیم

سطح عملیات تعمیم	تعداد هشدارها	نرخ کاهش هشدارها
GL0 (IP+PORT+PROTOCOL)	۲۸۳۰۳۶۵	۰,۰%
GL 1 (IP)	۲۲۸	۹۹/۹۹%
GL 2 (IP)	۶۸	۷۰/۱۸%
GL 3 (IP)	۵۱	۲۵/۰۰%
GL 4 (IP)	۳۲	۳۷/۲۵%
GL 5 (IP)	۲۰	۳۷/۵۰%

جدول ۳. نسبت تعداد هشدارها در ترافیک نرمال و حملات

حملات	تعداد هشدارها	درصد
BENIGN	۲۲۷۳۰۹۷	۸۰/۳۰%
DoS Hulk	۲۳۱۰۷۳	۸/۱۶%
PortScan	۱۵۸۹۳۰	۵/۶۱%
DDoS	۱۲۸۰۲۷	۴/۵۲%
DoS GoldenEye	۱۰۲۹۳	۰/۳۶%
FTP-Patator	۷۹۳۸	۰/۲۸%
SSH-Patator	۵۸۹۷	۰/۲۱%
DoS slowloris	۵۷۹۶	۰/۲۰%
DoS Slowhttptest	۵۴۹۹	۰/۱۹%
Bot	۱۹۶۶	۰/۰۷%
Web Attack – Brute Force	۱۵۰۷	۰/۰۵%
Web Attack – XSS	۶۵۲	۰/۰۲%
Infiltration	۳۶	۰/۰۰%
Web Attack – Sql Injection	۲۱	۰/۰۰%
Heartbleed	۱۱	۰/۰۰%
مجموع	۲۸۳۰۳۶۵	

کمیت‌های محاسبه‌شده از روابط (۲ و ۳) به‌دست‌آمده‌اند:

$$f_i = \frac{f}{n} \times 100 \qquad (۲)$$

در رابطه فوق f_i درصد فراوانی هشدارهای حملات نوع i و f تعداد هشدارهای حملات نوع i و n تعداد کل هشدارها است.

$$R_i = \frac{A_{i-1} - A_i}{A_{i-1}} \times 100 \qquad (۳)$$

در رابطه فوق A_{i-1} تعداد هشدارها در سطح i-1 تعمیم و A_i تعداد هشدارها در سطح i و R_i درصد کاهش هشدارها در تعمیم از GL_{i-1} به GL_i

چند استنتاج از خروجی‌های برنامه به‌صورت زیر است:

۱) در کنار هشدارهای نرمال[1] ۱۴ نوع حمله مختلف انجام شده است که با توجه به جدول (۳) حمله Dos Hulk با فراوانی ۸,۱۶ ٪ بیشترین فراوانی و حمله Heartbleed با فراوانی ۰/۰۰۰۴ ٪ کمترین فراوانی را در بین حملات دارا می‌باشند.

۲) از خروجی‌های GL0 مشخص است که حمله‌ای از نوع Bot از آدرس ۱۹۲,۱۶۸,۱۰,۱۵ از شماره درگاهی ۵۴۰۱۲ به آدرس ۲۰۵,۱۷۴,۱۶۵,۷۳ به شماره درگاهی ۸۰۸۰ با پروتکل شماره ۶ و تعداد یک انجام شده است. همان‌طور که مشاهده می‌شود داده‌های سطح پایین با تکرار کم امکان تصمیم‌گیری مناسب در مورد میزان اهمیت حملات را مهیا نمی‌سازند.

۳) با انجام سازوکار تعمیم و مراجعه به سطح بعدی سلسله مراتب GL1 مشخص می‌شود که حدود ۲۳۱۰۰۰ هشدار از نوع حمله DOS Hulk به مقصدی با سیستم‌عامل Ubunto16 با درگاهی مقصد خصوصی انجام پذیرفته است. تعداد زیاد هشدارها در این مورد و تعمیم هشدارهای سطح پایین در این مرحله مدیر شبکه را در تصمیم‌گیری درباره اهمیت این هشدارها و اقدامات لازم برای مقابله توانا خواهد ساخت.

۴) با مراجعه به سطوح بالاتر تعمیم (GL2-GL5) دید بالاتری از هشدارها و کشف روند موجود بین هشدارها به‌دست خواهد به‌عنوان‌مثال در آخرین مرحله تعمیم انجام‌شده سه نوع حمله DOS Hulk، Port Scan و DDos از فراوانی بیشتری در بین حملات برخوردار بوده‌اند با حرکت رو به عقب در تعمیم می‌توان اطلاعات با جزئیات بیشتری را کسب کرد مثلاً بیشتر حملات DDos به Web Server حمله کرده‌اند.

حرکت در سطوح مختلف تعمیم، امکان حرکت روبه‌جلو و عقب[2] در هشدارها را فراهم می‌آورد که از ابزارهای اصلی عملیات پردازش تحلیلی برخط[3] و داده‌کاوی چندبعدی[4] هست.

۴. نتیجه‌گیری

در این مقاله از یک روش مؤثر تعمیم داده‌ها از حوزه داده‌کاوی به نام استنتاج ویژگی محور استفاده‌شده است و اقدام به توسعه و پیاده‌سازی آن به‌منظور کشف حملات سایبری در شبکه هدف مطابق با مجموعه داده CICIDS2017 شده است. برای این منظور در ابتدا سطوح سلسله‌مراتب برای ویژگی‌ها مهم در شناسایی حملات طراحی‌شده‌اند، سپس یک روش شهودی برای انتخاب ویژگی برای عمل تعمیم ارائه شده است. نتایج بیانگر تعمیم مناسب(۹۹ ٪ در سطح ۱ و حداقل ۲۵٪ در سطوح دیگر تعمیم) هشدارهای تولیدشده در سامانه‌های تشخیص نفوذ به‌منظور کشف و تمایز با دقت حملات است. در کنار ترافیک نرمال ۱۴

[1] Benign

[2] Roll-Up & Drill-Down
[3] OLAP
[4] Multidimensional data mining

نوع حمله مختلف شناسایی شد که حملـه Dos Hulk بـا فراوانـی ۸/۱۶٪ بیشـترین و حملـه Heartbleed بـا فراوانـی ۰/۰۰۰۴٪ کمترین را دارا می‌باشند. همچنین با کمک خروجی‌های مختلف تعمیم امکان حرکت روبه‌جلو و عقب در سطوح مختلف هشـدارها کـه لازمـه تحلیـل مناسـب در داده‌کـاوی چندبعـدی در حـوزه تشخیص حملات سایبری است فراهم شده است. در پژوهش‌های آینده می‌تـوان بـا پیـاده‌سـازی معیارهـایی، کیفیـت هشـدارهای ارائـه‌شـده را در سـطوح مختلـف بـرای کـاربر محاسـبه کـرد و هشدارهای کاذب و واقعی را با مقادیر احتمالی محاسبه کرد. ایـن کار نیازمند بررسی دقیق‌تر ویژگـی‌هـای مـؤثر دیگـر در حمـلات سایبری دربسته‌های شبکه است. همچنین طراحی سازوکارهایی برای تشخیص حملات مخفیانه مفید اسـت. حملاتـی کـه باعـث تولید هشدارهای کم ولی اثرات تخریبی زیاد می‌شوند.

۵. مرجع‌ها

[1] Emmanouil, V.; Shankar, K.; Max, M.; Mathias, F. “Taxonomy and Survey of Collaborative Intrusion Detection”; ACM Computing Surveys (CSUR) 4, 2015.

[2] Min, C.; Kai, H.; Yu-Kwong, K.; Shanshan, S.; Yu, C. “Collaborative Internet Worm Containment”; IEEE Security & Privacy 3, 2005.

[3] Carlos, G., C.; Sascha, H.; Max, M.; Mathias, F. “Analyzing Flow-based Anomaly Intrusion Detection using Replicator Neural Networks”; In Annual Conference on Privacy, Security and Trust (PST), 2016.

[4] Onwubiko, C. “Situational Awareness in Computer Network Defense: Principles, Methods and Applications “; IGI Global, 2012.

[5] Estan, C; Savage, S; Varghese, G. “Automatically Iinferring Patterns of Resource Consumption in Network Traffic”; In: Proceedings of the conference on applications, technologies, architectures, and protocols for computer communications (SIGCOMM), 2003.

[6] Locasto, M.; Parekh, J.; Keromytis, A.; Stolfo, S. “Towards Collaborative Security and P2P Intrusion Detection”; In: Proceedings of the IEEE workshop on information assurance and security, 2005.

[7] Najafi, M.; Rafeh, R. “A New Light Weight Intrusion Detection Algorithm for Computer Networks”; Advanced Defence Sci. & Tech. 2016, 8, 191-200 (In Persian).

[8] Steffen, H.; Mathias, F. “GAC: Graph-Based Alert Correlation for the Detection of Distributed Multi-Step Attacks”; In ACM/SIGAPP Symposium On Applied Computing (SAC), 2018.

[9] Chenfeng, V. Z.; Christopher L.; Shanika, K. “Decentralized Multi-dimensional Alert Correlation for Collaborative Iintrusion Detection”, Volume 32, Issue 5, September 2009.

[10] Han, J.; Micheline, K.; Jian, P. “Data Mining Concepts and Techniques”; the Morgan Kaufmann Series in Data Management Systems, 2011.

[11] Beneditto, M. “Using Concept Hierarchies in Knowledge Discovery”. Lecture Notes in Computer Science, 2004.

[12] Han, J.; Fu, Y. “Exploration of the Power of Attribute-Oriented Induction in Data Mining”; in U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, eds. Advances in Knowledge Discovery and Data Mining, 399-421, 1995.

[13] Han, J.; Cai, Y.; Cercone, N. “Knowledge Discovery in Databases: an Attribute-oriented Approach”; In Proceedings of the 18th International Conference on Very Large Data Bases, 547-559, 1992.

[14] Meo, R.; Psaila, G.; Ceri, S. “An Extension to SQL for Mining Association Rules”; In Proceedings of Data Mining and Knowledge Discovery; 1998, 2,195-224.

[15] Muyeba, M.; Marnadapali, R. “A framework for Post-Rule Mining of Distributed Rules Bases”; In Proceeding of Intelligent Systems and Control; 2005.

[16] Elfeky, M.G.; Saad, A.; Fouad, S.A. “ODMQL: Object Data Mining Query Language”; In Proceedings of the International Symposium on Objects and Databases, 2000, 128-140.

[17] Cheung, D.W.; Hwang, H.; Fu, A.W. “Efficient Rule-Based Attribute-Oriented Induction for Data Mining”; Journal of Intelligent Information Systems, 2000, 15, 175–200.

[18] Cai, Y.; Cercone, N.; Han, J. “An attribute-oriented Approach for Learning Classification Rules from Relational Databases”; InProceedings. Sixth International Conference on Data Engineering IEEE, 1990, 281-288.

[19] WU, X.; XIE, L. “Attribute-oriented Induction and Conceptual Clustering”; Computer Engineering, Beijing, 2003, 92-99.

[20] Warnars, H. “Using Attribute Oriented Induction High level Emerging Pattern (AOI-HEP) to mine frequent patterns”; International Journal of Electrical and Computer Engineering (IJECE). 2016 Dec, 3037-46.

[21] Chenfeng, V. Z.; Christopher, L.; Shanika, K. “a Survey of Coordinated Attacks and Collaborative Intrusion Detection”; Computers & Security, 2010.

[22] Estan, C.; Savage, S.; Varghese, G.”Automatically Inferring Patterns of Resource Consumption in Network Traffic”; In: Proceedings of the conference on applications, technologies, architectures, and protocols for computer communications (SIGCOMM), 2003. 137–48.

[23] Haas, S.; Florian, W.; Mathias, F. "Efficient Attack Correlation and Identification of Attack Scenarios based on Network-Motifs." arXiv preprint arXiv: 1905.06685, 2019.

[24] ICS-CERT Advisories, Information about Current Security issues, Vulnerabilities, and Exploits. Available: https://www.us-cert.gov/ics/advisories, 2019.

[25] The National Institute of Standards and Technology (NIST) ,National Vulnerability Database, Available: https://nvd.nist.gov/vuln, 2019.

[26] Internet Storm Center, DShield.org. Available: http://www.dshield.orgi, 2019.

[27] Hu, Y.; Chiu, D.; Lui, J. “Adaptive Flow Aggregation—a New Solution for Robust Flow Monitoring under Security Attacks”; In: Proceedings of the 10th IEEE/IFIP network operations and management symposium (NOMS); 2006, 424–35.

[28] Taheri, R.; Parsaei, M.; Javidan, R. “Real-Time Intrusion Detection System Using a Combination of Discretization and Feature Selection”; Advanced Defence Sci. & Tech. 2017, 8, 251-263 (In Persian).

[29] Sharafaldin, I.; Habibi, A.; Lashkari; Ghorbani, A. “Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization”, 4th International Conference on

Information Systems Security and Privacy (ICISSP), Portugal, January 2018.

[30] Thomas, C.; Vishwas, S.; Balakrishnan N. "Usefulness of DARPA Aataset for Intrusion Detection System Evaluation"; Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security 2008. Vol. 6973. International Society for Optics and Photonics, 2008.

[31] Cup, K. D. D. "Intrusion Detection Data Set." The UCI KDD Archive Information and Computer Science University of California, Irvine. DOI= http://kdd. ics. uci.Edu/databases/kddcup99, 1999.

[32] Cowan, C. "Defcon Capture the Flag: Defending Vulnerable Code from Intense Attack"; Proceedings DARPA Information Survivability Conference and Exposition, IEEE, 2003, 1.

[33] CAIDA: Center for Applied Internet Data Analysis, Available: https://www.caida.org, 2019.

[34] CDX 2009 DataSet, Available: https://www.usma.edu/centers-and-research/cyber-research-center/data-sets, 2019.

[35] McKinney, W. "Pandas: a Foundational Python Library for Data Analysis and Statistics. "Python for High Performance and Scientific Computing 14.9, 2011.

[36] Anytree 2.7.3 Documentation, Available: https://anytree.readthedocs.io/en/latest/intro.html, 2019.