Journal of Electronical & Cyber Defence
Vol. 9, No. 3, 2021, Serial No. 35

Presentation of a Comprehensive Semi-Supervised Model for Collaborative Intrusion Detection Based on Network Behavior Profiling Using the Concept of Deep Learning and Fuzzy Correlation of Alerts

M. Ahmadzadeh[1], J. Vahidi[2]*, B. Minaei Bidgoli[3], A. R. Pourebrahimi[4]

*School of Mathematics, Iran University of Science and Technology, Tehran, Iran

(Received: 11/03/2021, Accepted: 08/06/2020)

ABSTRACT

Today, intrusion detection systems are extremely important in securing computers and computer networks. Correlated systems are next to intrusion detection systems by analyzing and combining the alarms received from them, appropriate reports for review and producing security measures. One of the problems face intrusion detection systems is generating a large volume of false alarms, so one of the most important issues in correlated systems is to check the alerts received by the intrusion detection system to distinguish true-positive alarms from false-positive alarms. The main focus of this research is on the applied optimization of classification methods to reduce the cost of organizations and security expert time in alert checking. The proposed Incrimental Intrusion Detetection Model using Correlator (IIDMC) is tested on a valid test dataset and the results show the efficiency of the proposed model and consequently its high accuracy.

Keywords: Intrusion Etection, Fuzzy Correlator, Incremental Online Learning, Active Learnin

* Corresponding Author Email: jvahidi@iust.ac.ir

نشریه علمی "پدافند الکترونیکی و سایبری"
سال نهم، شماره ۳، پاییز ۱۴۰۰، ص ۱۸۶- ۱۶۵

علمی- پژوهشی

ارائه مدل جامع نیمه نظارتی تشخیص نفوذ مشارکتی مبتنی بر نمایه‌سازی رفتار شبکه با استفاده از مفهوم یادگیری عمیق و همبسته‌سازی فازی هشدارها

محمد احمدزاده[۱]، جواد وحیدی[۲*]، بهروز مینایی بیدگلی[۳]، علیرضا پورابراهیمی[٤]

۱- دانشجوی دکترای رشته مدیریت فناوری اطلاعات، گروه مدیریت فناوری اطلاعات، دانشکده مدیریت و اقتصاد، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران، ۲- استادیار دانشکده ریاضی، دانشگاه علم و صنعت ایران، تهران، ایران، ۳- دانشیار دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران ٤- استادیار دانشکده مدیریت و حسابداری، واحد کرج، دانشگاه آزاد اسلامی، کرج، ایران

(دریافت: ۱۳۹۹/۱۲/۲۱، پذیرش: ۱۴۰۰/۰۳/۱۸)

چکیده

امروزه سیستم‌های تشخیص نفوذ اهمیت فوق‌العاده‌ای در تامین امنیت رایانه‌ها و شبکه‌های رایانه‌ای بر عهده دارند سیستم‌های همبسته‌ساز در کنار سیستم‌های تشخیص نفوذ قرار گرفته و با تحلیل و ترکیب هشدارهای دریافتی ازآن‌ها گزارش‌های مناسب برای بررسی و انجام اقدامات امنیتی تولید می‌نمایند یکی از مشکلاتی که سیستم‌های تشخیص نفوذ با آن روبرو هستند، تولید حجم زیادی از هشدارهای غلط است، بنابراین یکی از مهمترین مسائل در سیستم‌های همبسته‌ساز، وارسی هشدارهای دریافت شده از سیستم تشخیص نفوذ به‌منظور تشخیص هشدارهای مثبت کاذب از هشدارهای مثبت صحیح است در این مقاله یک مدل جامع و کاربردی ارائه شده است که شامل یک سیستم تشخیص نفوذ ترکیبی برای وارسی جریان ترافیک به‌صورت برخط و یک سیستم همبسته‌ساز مبتنی بر یادگیری افزایشی برای وارسی هشدارها با کمک یادگیری فعال است. تمرکز اصلی این پژوهش بر روی بهینه‌سازی کاربردی روش‌های دسته‌بندی به‌منظور کاهش هزینه سازمان‌ها و زمان متخصص امنیت برای در وارسی هشدارها هست. روش ارائه شده روی چند مجموعه داده تست معتبر آزمایش شده و نتایج حاصل بیانگر کارآمدی مدل پیشنهادی با دقت بالای ۹۹ درصد و با نرخ مثبت کاذب بسیار پایین است.

کلید واژه‌ها: سیستم تشخیص نفوذ مشارکتی، همبسته‌ساز، یادگیری افزایشی، یادگیری فعال، یادگیری برخط

۱- مقدمه

امروزه امنیت فناوری اطلاعات موضوعی مهم است و تلاش زیادی برای تحقیق در مورد نفوذ و کشف نفوذ انجام شده است. نفوذ را می‌توان هر نوع فعالیت غیرمجازی که باعث آسیب به یک سیستم اطلاعاتی شود، تعریف کرد. درواقع هرگونه حمله‌ای که محرمانه بودن اطلاعات، یکپارچگی یا در دسترس بودن اطلاعات را تهدید نماید، یک نفوذ محسوب می‌شود. به‌عنوان مثال، فعالیتهایی که باعث می‌شوند سرویس‌های رایانه‌ای نسبت به کاربران معمولی، پاسخگو نباشند، یک نفوذ محسوب می‌شود. IDS یک سیستم نرم افزاری یا سخت افزاری است که برای حفظ امنیت سیستم، اقدامات مخرب را بر روی سیستم‌های رایانه‌ای شناسایی می‌کند [۱]. هدف از IDS شناسایی انواع مختلف ترافیک شبکه مخرب و استفاده خرابکارانه از رایانه است که توسط یک دیوار آتش سنتی قابل شناسایی نیست.

دستیابی به حفاظت بالا در برابر اقداماتی که در دسترس بودن، یکپارچگی، یا محرمانه بودن سیستم‌های رایانه‌ای را به خطر می‌اندازد، امری حیاتی می‌باشد. سیستم‌های IDS را می‌توان به دو گروه طبقه‌بندی کرد: سیستم تشخیص نفوذ مبتنی بر امضاها (SIDS) و سیستم تشخیص نفوذ مبتنی بر ناهنجاری (AIDS). در جدول (۱) طبقه بندی انواع روش‌ها و سیستم‌های تشخیص نفوذ نشان داده شده است.

جدول (۱): مقایسه انواع فناوری IDS، با استفاده از نمونه‌هایی از ادبیات تحقیق. "P" توانایی تشخیص حملات از پیش تعریف شده را نشان می‌دهد و "Z" توانایی تشخیص حملات صفر روزه را نشان می‌دهد.

منبع تشخیص			HIDS	NIDS
روش‌های تشخیص	SIDS		P	P
	AIDS	مبتنی بر آمار	Z	Z
		مبتنی بر دانش	Z	Z
		مبتنی بر یادگیر ماشین	Z	Z
	SIDS+ AIDS		P+Z	

پس از نصب و راه اندازی یک IDS مدیریت هشدارهای تولیدشده توسط آن یکی از فرآیندهای حساس در تشخیص و پاسخگویی به نفوذ میباشد. مدیریت هشدارها با اعمالی از قبیل اولویت‌بندی هشدارها، شناسایی هشدارهای کاذب، رده‌بندی هشدارهای مرتبط و استخراج سناریوی ارتباطات منطقی بین هشدارهای منفصل، به کاربری بهتر IDS کمک می‌کند. چنین

* رایانامه نویسنده مسئول: jvahidi@iust.ac.ir

سیستم مدیریت هشدارها را معمولاً سیستم همبسته‌سازی هشدارها می‌نامند.

همبسته‌سازی دارای دو هدف اصلی است: کم کردن تعداد هشدارهایی که مدیر سیستم دریافت می‌کند و افزایش میزان مرتبط بودن و سطح تجرید گزارشات تولیدی [۲] علیرغم مشخص بودن اهداف فرآیند، به دلیل پیچیدگی و چندمولفه‌ای بودن آن، در کارهای مختلف به جنبه‌های متفاوتی از این فرآیند پرداخته شده است. فرآیند همبسته‌سازی را می‌توان در قالب‌ساده پیش‌پردازش، پردازش و پس‌پردازش هشدارها تشریح نمود [۳] که در هر یک بخشی از انتظارات موجود از سیستم همبسته‌ساز تامین می‌شود. در بخش پیش‌پردازش، نرمال‌سازی هشدارها و تقلیل دادهای صورت می‌گیرد. تقلیل دادهای از طریق تجمیع هشدارها، فیلترکردن هشدارها و کاهش هشدارهای کاذب صورت می‌پذیرد. در بخش پردازش انواع روش‌های قابل تصور برای کشف ارتباطات بین هشدارها و رده‌بندی آنها مورد استفاده قرار می‌گیرد. پردازش انجام‌شده در این بخش اصلی‌ترین عملیاتی است که منجر به تولید سناریوهای حملات و ایجاد سطح تجریدی بالاتر نسبت به حمله در حال وقوع می‌شود [۴]. سرانجام در مرحله پس‌پردازش اعمال تکمیلی برای کمک به مدیر سیستم در جهت استفاده بهتر از اطلاعات تولید شده صورت می‌پذیرد. کارهایی از قبیل اولویت‌بندی هشدارها و تشخیص مقصود اصلی‌ترین اعمالی است که در این مرحله انجام می‌شوند. با توجه به اینکه کلیه اعمال انجام شده در این سه مرحله به نوعی منجر به همبسته‌سازی هشدارها و دستیابی به اهداف آن که همان کم کردن تعداد هشدارها و افزایش سطح تجرید آنهاست می‌انجامد. در بخش بعدی به بررسی اجمالی کارهای انجام شده پرداخته است. در بخش ۳ مدل پیشنهادی ارائه شده است همچنین در بخش ۴ به بررسی نتایج آزمایشگاهی و در انتها نیز به بحث و نتیجه‌گیری پیرامون تحقیق پرداخته شده است.

۲- روش تحقیق (شامل: کارهای مرتبط، سیر تحول موضوع و سئوال مورد تحقیق و راهکار پاسخ به آن)

۲-۱- سیستم های تشخیص نفوذ

یکی از مهمترین روش‌های تشخیص نفوذ می‌توان به روش تشخیص ناهنجاری در شبکه شامل تشخیص ناهنجاری‌های براساس طبقه‌بندی آماری، تشخیص بر اساس شبکه عصبی، تشخیص بر اساس داده‌کاوی و غیره اشاره کرد. در طول این سال‌ها محققان و طراحان بسیاری از این تکنیک‌ها در طراحی سیستم‌های تشخیص نفوذ استفاده کرده‌اند. اما یک یا چند مشکل در سیستم‌های تشخیص نفوذ در حال حاضر وجود دارد. سیستم‌های تشخیص نفوذ برای تشخیص ناهنجاری باید ابتدا ویژگی فعالیت‌های عادی و فعالیت‌های غیر طبیعی را یاد بگیرد و پس از آن سیستم تشخیص نفوذ یا IDS، ترافیکی که از فعالیت‌های عادی منحرف می‌شود را تشخیص می‌دهد. تشخیص ناهنجاری تلاش می‌کند تا تعیین کند که آیا می‌تواند از انحراف از الگوهای عادی را به‌عنوان نفوذ استفاده نماید [۵]. در [۶] یک استراتژی موثر معرفی شد که این استراتژی، استراتژی‌های داده‌کاوی و سیستم خبره را برای طراحی IDS ترکیب می‌کند. این روش به نظر می‌رسد امیدوار کننده باشد اما هنوز هم با مشکلات ساختاری و عملکردی مواجه است و نیاز به بهبود بیشتری دارد.

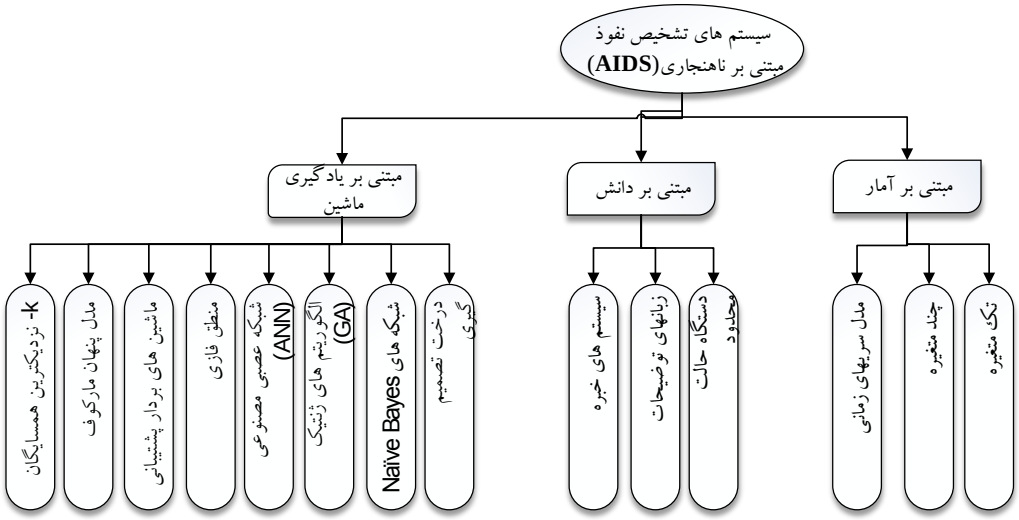
در تحقیقی با عنوان استفاده از الگوریتم داده‌کاوی برای توسعه مدلی برای سیستم تشخیص نفوذ که توسط دوک[1] و همکاران ارائه شده است، نشان می دهد که آیا زمانیکه تعداد صحیحی از خوشه‌ها در الگوریتم K-Means بکار برده شود، می‌تواند به نرخ اثربخشی بالایی برسد [۷].

رویکرد روال جستجو تطبیقی تصادفی حریصانه با رده‌بندی بازپخت تصادفی[2] که توسط [۸] ارائه شد به عنوان یکی از اثربخش ترین رویکردها با درصد بالاتری از دقت معرفی شده است.

در مقاله‌ای که توسط [۹] ارائه شد به مرور برخی رویکردهای مبتنی بر داده‌کاوی مانندSVM, KSVM, ELM, KELM برای تشخیص نفوذ پرداخته و در آخر به این نتیجه رسیده است که ترکیب نتایج بیش از یک الگوریتم داده‌کاوی بایکدیگر می‌تواند معایب یکدیگر را از بین ببرند.

در مقاله [۱۰] تلاش شده است با استفاده از الگوریتم‌های متفاوت داده‌کاوی در تشخیص نفوذ در شبکه شامل الگوریتم K-Means و رگرسیون خطی، قواعدی را برای رده‌بندی فعالیت‌های شبکه تولید کند و در نهایت یک مطالعه تطبیقی از عملکرد این الگوریتم‌ها بر روی مجموعه داده‌های NSL-KDD انجام داده است. در مقاله (اشفق و همکاران ۲۰۱۷). تحت عنوان رویکرد یادگیری نیمه نظارتی فازی برای تشخیص نفوذ، نشان داده شده است که به دلیل عدم وجود برچسب برای داده‌های نفوذ و نیاز به تلاش زیاد خبرگان برای برچسب‌گذاری داده‌ها و با توجه به بدون برچسب بودن مسائل دنیای واقعی به ارائه مدل نیمه نظارتی فازی می‌پردازد که توسط حجم زیادی از نمونه‌های بدون برچسب به الگوریتم یادگیری با نظارت برای بهبود عملکرد رده‌بندی برای سیستم‌های تشخیص نفوذ کمک می‌کند و در انتها نیز به مقایسه مدل خود با الگوریتم‌های رده‌بندی پایه بیز ساده، SVM و جنگل‌های تصادفی[3] توسط مجموعه داده NSL-KDD پرداخته است.

[1] Duque
[2] GAR-Forest
[3] Random Forests



شکل(۱): طبقه‌بندی از روش‌های تشخیص نفوذ مبتنی بر ناهنجاری (AIDS) در یک نگاه

در [۱۱] نویسندگان یک روش یادگیری افزایشی نظارتی بر اساس آبشارسازی دستبند سرویس (SC) و یادگیرنده افزایشی درخت (ITI) پیشنهاد کرده‌اند. دو محدودیت اساسی روش ITI آن است که اولاً ITI نیازمند تعداد کافی نمونه‌های آموزشی برای پشتیبانی انواع مختلف رفتار طبیعی است. ثانیاً این روش قادر نیست نمونه‌های جدیدی که دسته جدیدی دارند را پشتیبانی نماید. در طی فرایند یادگیری افزایشی، ITI نمونه‌های جدید دارای دسته جدید را نمی‌تواند در یادگیری درخت دودویی و به روزرسانی وزنها دخیل نماید. برای رفع این مشکل، روش آبشارسازی دستبند سرویس و ITI پیشنهاد شده است.

در [۱۲] نویسندگان روش RS-ISVM را پیشنهاد می‌کنند، الگوریتم SVM افزایشی بهبود یافته برای تشخیص نفوذ می‌باشد. برای کاهش اختلال ناشی از اختلاف زیاد مقادیر ویژگی‌ها، نویسندگان این مقاله، تابع هسته‌ای تغییر یافته U-RBF را پیشنهاد نموده‌اند که میانگین و میانگین مربعات، اختلاف مقدار ویژگی‌ها را در تابع هسته RBF تعبیه می‌کند. در مقاله مذکور با توجه به مشکل نوسان که معمولاً در روش‌های سنتی افزایشی SVM اتفاق می‌افتد یک راهبرد مجموعه رزرو پیشنهاد شده است. توسط این راهکار نمونه‌هایی که احتمال دارد بردارهای پشتیبان باشند، ذخیره می‌شوند علاوه بر این برای کوتاه کردن زمان آموزش، روش دوایر متحدالمرکز برای انتخاب نمونه‌های تشکیل دهنده مجموعه رزرو شده، پیشنهاد می‌شود.

در مقاله [۱۳] یک روش غیرافزایشی تشخیص نفوذ ارائه شده است، این روش موسوم به روش لایه‌ای تشخیص نفوذ نام دارد که در هر لایه یک نوع حمله خاص پوشش داده می‌شود بدین صورت که ترافیک عبوری در هر لایه بر اساس بردار زیر مجموعه‌ای از ویژگی‌های مرتبط با آن لایه مورد بررسی قرار می‌گیرد و در صورتی که ویژگی‌های مرتبط با آن نوع از حمله خاص را دارا باشد در همان نقطه فیلتر می‌شود. عملکرد این روش مناسب است اما دارای ضعف‌هایی می‌باشد مثلاً اگر ترافیکی در لایه اول DOS تشخیص داده شد دیگر در لایه دوم که مربوط به فیلتر نوع حمله PROB می‌باشد بررسی نمی‌شود و یا به خاطر سری بودن لایه‌ها سرعت سیستم تشخیص نفوذ با لایه‌های زیاد، می‌تواند کند شود.

در شکل (۱) انواع تکنیک‌های و رویکردهای تشخیص ناهنجاری به‌کار برده شده در ادبیات تحقیق در یک نگاه نشان داده شده است.

۲-۲- سیستم همبسته‌ساز هشدارها

در [۱۴] چارچوبی برای همبستگی هشدارها به صورت بلادرنگ ارائه شده است که از تکنیک جدیدی برای تجمیع هشدارها استفاده می‌نماید و الگوهای جدیدی از هشدارها استخراج می‌نماید.

فیلتر پس پردازش هشدارها بر مبنای پردازش و تواتر بالای هشدارها توسط [۱۵] ارائه شد که بر دو فرضیه مهم استوار است. اول آنکه توزیع تعداد هشدارهای همسایه تغییر زیادی از مثبت کاذب به مثبت صحیح دارد. فرضیه دوم آن است که در صورتیکه تواتر تکرار یک هشدار از میانگین تواتر هشدارهای با امضای مشابه بیشتر باشد، احتمال اینکه آن هشدار مثبت صحیح باشد زیاد است.

در [۱۶] از شبکه عصبی چند لایه پرسپترون و همچنین ماشین بردار پشتیبان برای محاسبه احتمال همبستگی دو هشدار استفاده کرده‌اند. آن‌ها احتمال وقوع هشدارهای پی در پی را در ماتریسی ذخیره و همبستگی‌های جدید را در آن بروز می‌نمایند و

توسط شبکه‌های عصبی و ماشین بردار پشتیبان با استفاده از این احتمالات، احتمال همبسته شدن دو هشدار جدید را محاسبه و مجدداً ماتریس احتمال همبستگی بروز می‌شود. در ابتدا بر اساس شباهت ویژگی‌های هشدارها فرایند طبقه‌بندی صورت می‌گیرد ولی به مرور زمان و با محاسبه شدت همبستگی بین انواع هشدارها و بر اساس همزمانی آماری به وسیله شبکه عصبی و ماشین بردار پشتیبان ماتریس همبستگی هشدار ساخته می‌شوند.

نینگ[1] و دیگران در [۱۷-۱۸] چارچوبی ارائه نموده که هشدارها توسط واحد پردازش هشدار به فرا هشدارهایی کد می‌شوند که بیانگر یک سناریو می‌باشد. بدین صورت که برای وقوع یک هشدار چه هشدارهای قبل و بعد از آن می‌تواند رخ دهد. فرا هشدارها بر اساس انواع فراهشدارهایی که از قبل در پایگاه دانش سیستم ذخیره شده‌اند تولید می‌شود. هر نوع فرا هشدار اطلاعات مربوط به یک نوع حمله را کد می‌نماید. اگر پیامد یک حمله پیشین با پیش نیاز حمله فعلی مطابقت کند آن دو حمله مربوط به یک سناریو بوده و با یکدیگر همبسته می‌شوند این کار با استفاده از یک پرس و جوی SQL بر اساس بازه زمانی روی پیش نیاز و پیامدهای یک هشدار صورت می‌گیرد تعدادی از مشکلات روش پیشنهادی شامل بزرگ شدن سریع گراف‌های تولیدی، تعداد زیاد قوانین پیش‌نیاز و پیامد و سختگیری در توالی هشدارها می‌باشد.

المموری و ژانگ [2] در [۱۹-۲۰] با ارائه مفهوم زیر حملات سعی کرده مشکلات روش پیشنهادی نینگ را حل نماید، بدین گونه که همبسته‌سازی بین زیر حملات صورت می‌گیرد تا گراف‌های تولیدی مختصرتر شوند. وی به جای تمرکز بر روی هر هشدار بر روی نوع هشدار تولیدی متمرکز می‌شود تا سناریوی حمله را شفاف تر نشان دهد. سیستم پیشنهادی وی شامل چهار مولفه اولویت‌بندی، بسته‌بندی، تجمیع و تولید گراف می‌باشد و وظیفه مولفه اول این است که اولویت نسبی هشدارها را بر اساس مشخصات و سیاست‌های شبکه تعیین کند. مولفه رده‌بندی هشدار را در گروه‌هایی با سطح معنایی بالاتری رده‌بندی می‌کند. گروه‌هایی از قبیل جمع‌آوری اطلاعات میزبان، جمع‌آوری اطلاعات سرویس، منع سرویس، فعالیت تروجان و غیره. نتیجه این رده‌بندی در اختیار واحد تجمیع قرار می‌گیرد تا بر اساس یک پنجره زمانی هشدارهای مشابه را به یکدیگر پیوند دهد. نتیجه تولیدشده توسط واحد تجمیع، توسط واحد تولید گراف در قالب گراف‌های مختصرتری نسبت به گراف نینگ ارائه می‌شود. به نظر می‌رسد مهمترین دستاورد کار وی نسبت به نینگ ارائه گراف‌های مختصرتری می‌باشد که منجر به افزایش سطح انتزاع سیستم می‌شود.

[1] Ning
[2] Al-Mamory and Zhang

۲-۳- راهکارها و چالش‌ها

یک سیستم تشخیص نفوذ بر داده‌های زمان واقعی شبکه و میزبان‌ها به‌منظور تشخیص رفتارهای بدخواهانه نظارت می‌کند و زمانیکه یک فعالیت نفوذ تشخیص داده شود یک هشدار تولید می‌نماید. دو روش برای طبقه‌بندی سیستم‌های تشخیص نفوذ وجود دارد.

انواع داده‌های تحلیل‌شده: تشخیص نفوذ مبتنی بر شبکه(NIDS)، سیستم تشخیص نفوذ مبتنی بر میزبان(HIDS) و سیستم تشخیص نفوذ مشارکتی(CIDS) [۲۱].

انواع حملات: سیستم تشخیص سوء استفاده و سیستم تشخیص ناهنجاری [۲۱].

در بخش بعدی به بررسی چالش‌ها و محدودیت‌های هر یک و راهکار پیشنهادی برای مقابله با این چالش‌ها پرداخته می شود.

۲-۳-۱- طبقه‌بندی بر اساس انواع داده‌های تحلیل‌شده، چالش‌ها و راهکارها

سیستم تشخیص نفوذ مبتنی بر شبکه NIDS: NIDSها ارتباطات یک شبکه را مورد بررسی قرار می‌دهند تا نفوذ را بیابند. به عنوان نمونه از روش‌های داده‌کاوی به‌منظور تشخیص ناهنجاری در داده‌های ترافیک شبکه استفاده می‌کنند.

• **محدودیت‌های NIDS**: NIDS اغلب به شکل سخت افزاری بین اینترنت و شبکه داخلی یک سازمان استقرار می‌یابد. فواید این شکل از استقرار این است که NIDS می‌تواند نفوذ به شبکه را فوراً تشخیص دهد. گرچه این موضوع علاوه بر اینکه برای NIDSهای سنتی به دلیل پردازش بسته‌های رمزگذاری شده و کندی سرعت جریان شبکه، مشکل ایجاد نماید، کشف حملات داخلی برای این نوع ازNIDSها نیز مشکل خواهد بود.

سیستم‌های تشخیص مبتنی بر میزبان HIDS: HIDS فعالیت بین میزبان‌ها را برای کشف رفتارهای احراز هویت نشده، نظارت می‌نماید. HIDSها می‌توانند انواع مختلفی از روش‌های داده‌کاوی همچون شبکه‌های عصبی مصنوعی را روی داده‌های ممیزی میزبان یعنی لاگ فایل‌ها برای کشف حملات اجرا نماید.

• **چالش HIDS**: سرعت اجرای HIDS پس از تجمیع زمان یادگیری و تست مشخص می‌شود. با توجه به اینکه HIDS نیاز به مدیریت حجم عظیمی از داده‌ها دارد سرعت دارای محدودیت است. علاوه بر این نگهداری و به‌روزرسانی تعداد زیادی از HIDSهای سنتی نصب شده روی هر میزبان یا میزبان مجازی در یک شبکه در مقایسه با تعداد ابزارهای NIDS خسته‌کننده است. علاوه بر این HIDSهای سنتی قدرت زیادی را در مقابله با

تهدیدات در حال پیشرفت مداوم نشان ندادند. بنابراین نیاز به HIDSهایی که در آینده بتوانند به‌صورت تعاملی با دیگر مکانیزم‌های امنیتی کار کنند، ضروری است.

سیستم‌های تشخیص نفوذ مشارکتی CIDS: CIDS در مورد ترکیب مشارکتی NIDS و HIDS و دیگر مکانیزم‌های امنیتی یک شبکه به‌منظور کارآمدی و اثرگذاری در تشخیص حملات سایبری است. CIDSها را می‌توان به سه گروه متمرکز، غیر متمرکز و توزیع شده تقسیم نمود.

•محدودیت‌های CIDSهای سنتی: روش‌های CIDS سنتی اغلب تحلیل جامع و متمرکزی از ترافیک شبکه و لاگ‌های میزبان ارائه نمی‌دهد و همچنین CIDSهای سنتی به اندازه کافی برای مدیریت حجم عظیمی از جریان‌های داده بلادرنگ اثرگذاری ندارند.

۲-۳-۲- طبقه‌بندی بر اساس نوع حملات، چالش‌ها و راهکارها

سیستم کشف سوء استفاده. سیستم کشف سوء استفاده (سیستم کشف نفوذ مبتنی بر امضا) کتابخانه ای از امضاءهای حملات تایید شده را تعریف می‌نماید و هنگامی که ترافیک شبکه یا عملیات سیستم با هرگونه امضاء حمله در کتابخانه مطابقت داشته باشد هشداری ایجاد می‌شود که توسط مدیر سیستم از پیش تعریف شده است. این کتابخانه تلاش می‌کند تا هر گونه رفتار غیر عادی شبکه را به طور دقیق لیست کرده و حفظ نماید و با سایر رفتارهای شناخته شده و ناشناخته به صورت عادی رفتار می‌شود، بنابراین سیستم تشخیص سوء استفاده میتواند روش‌های حمله ای را که قبلاً مشخص شده‌اند را کشف نماید.

•محدودیت سیستم تشخیص سوء استفاده: سیستم تشخیص سوء استفاده به دلیل بالا بودن نرخ هشدار از دست رفته مورد انتقاد است. افزایش روز افزون حملات صفر روزه باعث شده این رویکرد منسوخ شود. نفوذگران می‌توانند به سادگی روش‌های حمله خود را مبهم نمایند تا از کتابخانه نفوذ از پیش تعریف شده عبور کنند.

سیستم کشف ناهنجاری. سیستم کشف ناهنجاری(ADS) نیازی به دانش حملات شناخته شده ندارد. [۲۲] یک مرور جامع از سیستم‌های تشخیص نفوذ ارائه کردند که در حوزه سیستم‌های تشخیص نفوذ، به عنوان سیستم تشخیص ناهنجاری مبتنی بر شبکه و سیستم تشخیص ناهنجاری مبتنی بر میزبان رده‌بندی می‌شوند. NADS ناهنجاری‌ها را از ترافیک عادی شبکه تفکیک می‌نماید. HADS معمولاً در سیستم‌هایی که رفتارهای نرمال متناوباً تغییر نمی‌کند استقرار می‌یابند. HADS به دنبال ساخت پایگاه داده نرمال یا مدل داده‌کاوی رفتار نرمال سیستم است تا از این پایگاه داده‌ها یا مدل‌ها بتواند به عنوان معیاری برای کشف ناهنجاری استفاده نماید. بنابراین سیستم تشخیص ناهنجاری می‌تواند حملات صفر روزه را کشف نماید.

•چالش‌های سیستم تشخیص ناهنجاری مبتنی بر میزبان: نرخ هشدار کاذب بالا یک چالش است. از آنجا که HADS فقط بانک‌های اطلاعاتی رفتار عادی یا مدل‌های استخراج داده از رفتارهای شناخته شده را نگهداری می‌نماید لذا رفتارهای نرمال جدید که مطابق با بانک‌های اطلاعاتی یا مدل‌ها نیستند ممکن است به غلط به‌عنوان نفوذ در نظر بگیرد.

ترکیب سیستم تشخیص ناهنجاری و تشخیص سوء استفاده. سیستم‌های تشخیص نفوذ اگرچه می‌تواند به این دو گروه طبقه‌بندی شود. ترکیب سیستم تشخیص سوء استفاده با یک سیستم تشخیص ناهنجاری رویکردی متداول برای توسعه یک زیرساخت تشخیص نفوذ جامع و برخط به‌منظور کشف حملات شناخته شده و ناشناخته است. به این شکل که در صورت مشاهده یک رفتار جدید پس از پیش پردازش می‌توان ابتدا آن را با قوانین از پیش تعریف شده از حملات شناخته شده مقایسه کرد اگر با هیچ حمله‌ای مطابقت نداشته باشد می‌توان آن را با پایگاه داده‌های رفتار عادی مقایسه نمود و یا با الگوریتم‌های دیگر تشخیص ناهنجاری مورد بررسی قرار داد.

این ردپاها می‌توانند برای تحقیقات آینده مورد استفاده کارکنان امنیت قرار گیرند ترکیب تشخیص سوء استفاده با تشخیص ناهنجاری به وسیله کارهای تحقیقاتی زیادی مانند [۲۳] مورد استفاده قرار گرفته‌اند، در بخش بعدی چارچوبی جامع و انعطاف‌پذیر به‌منظور پوشش تمامی ضعف‌های پیشین و استفاده از نقاط قوت رویکردهای متداول در این حوزه ارائه شده است.

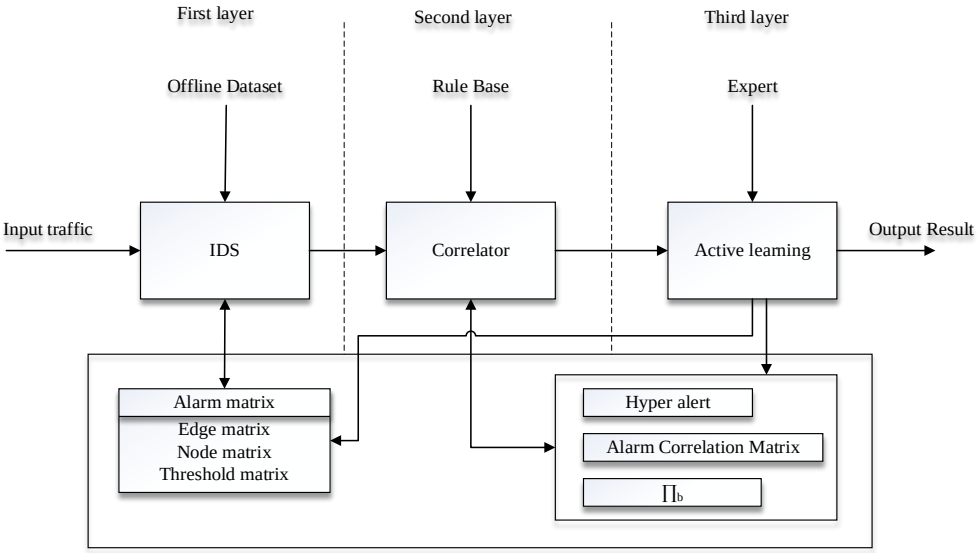
۳- مدل پیشنهادی

همانطور که اشاره شد بزرگترین ضعف سیستم‌های تشخیص نفوذ تولید هشدارهای مثبت کاذب است. متخصص امنیت باید نسبت به بررسی همبستگی هشدارهای تولید شده اقدام نماید تا بتواند هشدارهای واقعی را از مثبت کاذب تشخیص دهد.

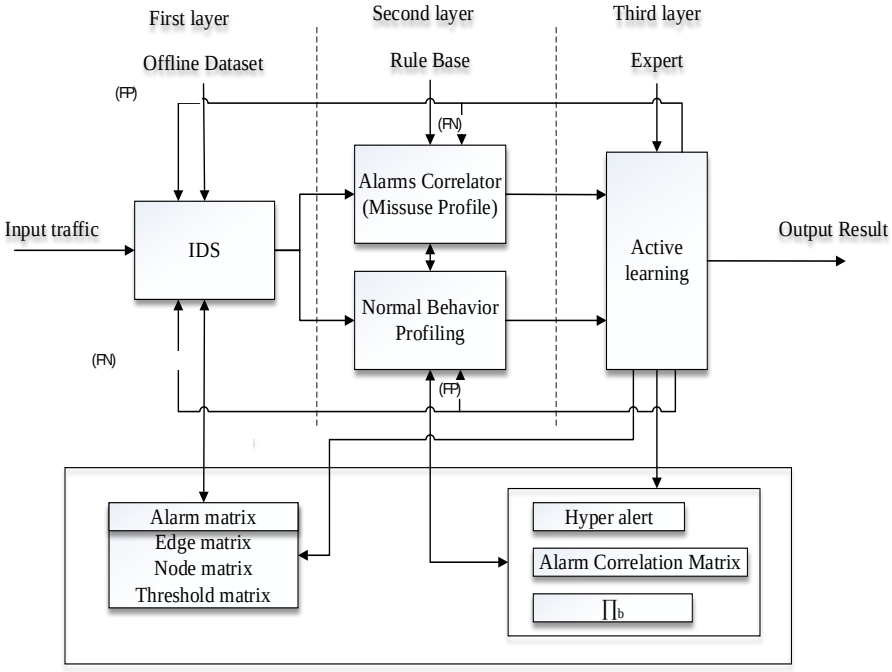
این موضوع هزینه سازمان را به شدت افزایش می‌دهد. این پژوهش، مدل جامع تشخیص نفوذی را ارائه می‌دهد که قادر است با ذخیره ی دانش کسب شده با استفاده از تکنیک‌های یادگیری فعال و یادگیری افزایشی به صورت برخط نسبت به کشف حملات، به کمک متخصصین امنیت بیاید.

مدل پیشنهادی از دو بخش اصلی شامل سیستم تولید هشدار و همبسته‌ساز تشکیل شده است که شمای کلی آن در شکل ۲ آمده است. همانطور که در شکل ۲ نشان داده شده است، بخش اول حاوی دو ماژول است. ماژول اول با استفاده از الگوریتمی مبتنی بر شبکه عصبی افزایشی خودسازمانده نسبت به یادگیری اقدام می‌نماید و ماژول دوم با استفاده از دانش کسب شده از مرحله قبل اقدام به تولید هشدار می‌نماید.

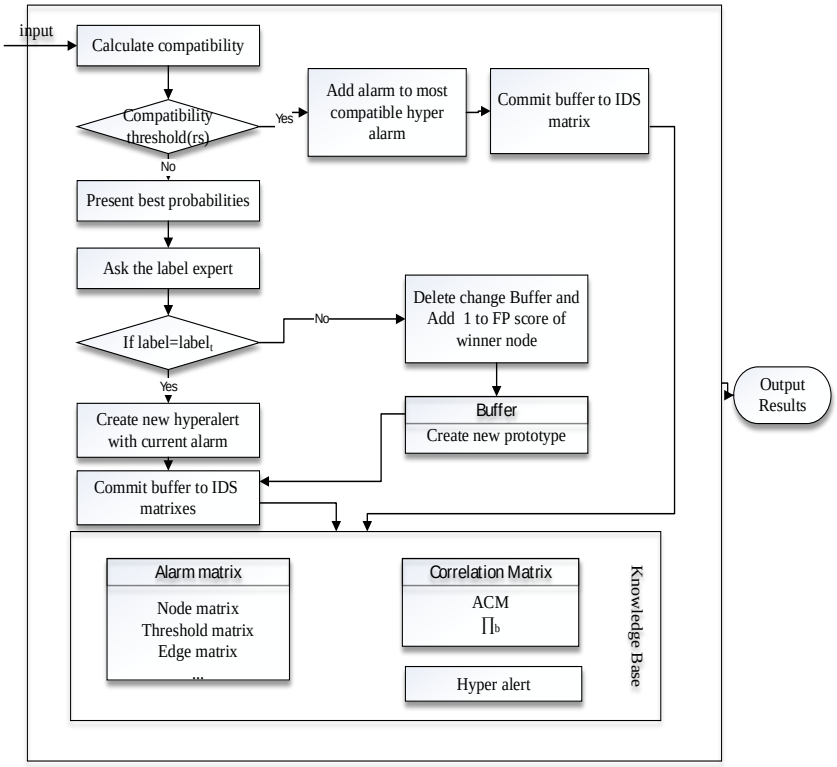
هشدارهای تولید شده به‌منظور کشف مثبت کاذب و خطاهای سیستم تشخیص نفوذ به بخش همبسته‌ساز منتقل می‌شود. سیستم همبسته‌ساز نسبت به بررسی برخی دیگر از ویژگی‌های هشدار تولید شده با کمک دانش ذخیره شده در پایگاه دانش اقدام نموده و میزان همبستگی آن را می‌سنجد. در صورتی‌که میزان همبستگی از حد آستانه بیشتر باشد، هشدار را حمله تشخیص می‌دهد. در غیر اینصورت با پرسش از خبره پاسخ مناسب به هشدار تولید شده می‌دهد و با کمک یادگیری فعال اقدام به ویرایش و ذخیره دانش جدید در پایگاه دانش می‌نماید.



شکل(۲): مدل پیشنهادی تشخیص نفوذ افزایشی



شکل(۳): مدل توسعه یافته تشخیص نفوذ افزایشی مبتنی بر همبسته‌سازی رفتارها بهمراه ویرایش مفهوم



شکل(۴): سیستم یادگیری فعال

در شکل (۳)، مدل توسعه یافته شکل (۲) نشان داده شده است این مدل با همان فرآیند قبلی اما با دو قابلیت که دقت مدل پیشنهادی را افزایش می‌دهد، ارائه شده است. در مدل توسعه یافته جدید در لایه دوم همبسته‌ساز از دو همبسته‌ساز برای همبستگی رفتار عادی و حملات که توسط سیستم تولید هشدار، نرمال و یا رفتار مخرب شناسایی شده‌اند، ذخیره می‌شوند. در این تحقیق مدل پیشنهادی تشخیص نفوذ افزایشی مبتنی بر همبسته‌ساز به اختصار IIDMFC[1] نامگذاری شده است. در ادامه جزئیات کارکرد ماژول‌های مدل پیشنهادی شرح داده شده است.

۳-۱- سیستم یادگیری فعال[2]

خروجی سیستم همبسته‌ساز احتمال همبستگی دو هشدار است. درصورتی‌که میزان احتمال همبستگی از میانگین آستانه همبستگی و آستانه انتخاب قانون کمتر باشد برای پالایش برچسب آن نمونه نماینده یا همان هشدار برچسب صحیح آن به‌صورت برخط از یک مرجع برچسب گذاری کاملاً دقیق مثلا متخصص انسانی پرسیده می‌شود که در ادبیات یادگیری ماشین و داده‌کاوی به آن یادگیری فعال گفته می‌شود. برای مثال اگر میزان همبستگی برای یک هشدار با برچسب برنده حاصل از سیستم تولید هشدار کمتر از حد آستانه باشد می‌توان هشدار مورد نظر را با برچسب‌های اول و حتی دوم و سوم برنده (حاصل از سیستم تولیدکننده هشدار) به همراه احتمال همبستگی و میزان انطباق قانون (حاصل از سیستم همبسته‌ساز) به سیستم یادگیری فعال ارائه نمود. حدآستانه در اکثر ادبیات تحقیق ۰٫۵ در نظر گرفته شده است البته این میزان می تواند بر حسب نوع مجموعه داده و و بر حسب دقت خروجی به‌صورت برخط یا برونخط تعیین و یا بروز شود در بخش بحث و نتیجه‌گیری به این موضوع پرداخته شده است.

این اطلاعات می‌تواند به مرجع برچسب‌گذاری مثلاً متخصص انسانی در راه رسیدن به برچسب دقیق کمک کند البته هیچ الزامی در استفاده از این اطلاعات وجود ندارد چرا که مزیت این سیستم این است که می‌تواند الگوهای حملاتی که تاکنون سیستم یاد نگرفته و در طی زمان یادگیری افزایشی نیز آنها را تشخیص نداده است را به پایگاه دانش خود اضافه نماید و به این روش هزینه متخصص انسانی را در آینده کاهش دهد.

اگر میانگین مجموع احتمال همبستگی با میزان انطباق قانون از حد آستانه مثلاً ۰٫۵ بیشتر باشد هشدار موردنظر حمله تشخیص داده خواهد شد در غیر اینصورت با استفاده از دانش خبره برچسب دقیق پرسیده می‌شود و دانش به دست آمده برای تکامل مفهوم و یا ویرایش مفهوم در پایگاه دانش اعمال می‌شود.

[1] Incrimental Intrusion Detetection Model using Fuzzy Correlator
[2] Active learning

کارکرد سیستم یادگیری فعال همانطور که در شکل (۴) نشان داده شده است به این گونه است که اگر برچسب صحیح، برچسبی غیر از برچسب پیش‌بینی شده(برچسب نمونه نماینده برنده) توسط سیستم تولید کننده هشدار باشد، یادگیری فعال اقدام به افزودن یک نمونه نماینده با نوع حمله درست در مجموعه ماتریس‌های مرتبط با نمونه نماینده شامل ماتریس مشخصه، ماتریس یال‌ها و غیره می‌نماید.

در صورتی‌که نوع حمله نمونه نماینده برنده صحیح باشد اقدام به افزودن یک فراهشدار جدید حاوی یک هشدار با نوع حمله‌ای که به صورت فعال در این سیستم آموخته است، می‌نماید.

یعنی اگر نمونه‌ای که توسط سیستم همبسته‌ساز تشخیص داده نشود به یک فراهشدار یا فرارفتار نرمال جدید ایجاد و ماتریس‌های مرتبط بروزرسانی و تکامل مفهوم صورت می‌گیرد، در غیر این‌صورت با رانش مفهوم نسبت به اصلاح و بروزرسانی سیستم تولید هشدار و ماتریس‌های مربوطه مبادرت می‌شود.

برای اینکه بر اساس دانش سیستم همبسته‌ساز خروجی سیستم تولیدکننده هشدار تصدیق و یا رد شود ارتباط سیستم تولید هشدار با سیستم همبسته‌ساز ضروری است و با این روش می‌توان چالش مثبت کاذب را تا حدی مرتفع نمود به این صورت که برچسب مثبت کاذب که آفت سیستم‌های تشخیص نفوذ است توسط سیستم تولید کننده هشدار یادگرفته و در پایگاه دانش بروز می‌شود. این امر موجب افزایش کارائی و دقت سیستم به مرور می‌شود و داده‌های به ظاهر پرت یا همپوشان به جای اینکه دور ریخته شوند و هزینه صرف شده توسط سیستم به هدر رود می‌توان نوعی کاربرد برای درک مفهوم "مثبت کاذب" توسط سیستم ایجاد شود و به اینصورت سیستم با ایجاد برچسبی مانند برچسب "مثبت کاذب" از اشتباهات خود نیز درس می‌گیرد و تنها از آن دانش برای تعیین تکلیف هشدار جاری استفاده نمی‌نماید.

مثبت کاذب می‌تواند ترافیک عادی باشد که به اشتباه هشدار تولید کرده است و یا ممکن است نوعی دیگری از هشدارها باشد که از اهمیت پائینی برخوردار است. در صورت تشخیص حالت اول می‌توان در قالب یادگیری فعال توسط متخصص امنیت از موضوع مطمئن شد تا اگر این موضوع صحیح باشد، دانش مورد استفاده در سیستم تولید هشدار اصلاح شود و اگراشتباه باشد آنگاه دانش مورد استفاده در سیستم همبسته‌ساز به روز شود. به این شیوه ایده تکامل و رانش مفهوم شکل می‌گیرد.

● تکامل و رانش مفهوم در یادگیری فعال

الگوریتم‌هایی مانند طبقه‌بندها، بعد از اینکه یادگیری را انجام دادند بایستی عملیاتِ پیش‌بینی و طبقه‌بندی را بر روی داده‌های جدید انجام دهند. حال اگر یک الگوریتم نتواند بگوید داده جدیدی که از راه رسیده به طور قطع متعلق به کدام طبقه است، می‌تواند این داده جدید را به یک خبره یا همان کسی که بتواند برچسب دقیق نمونه داده را مشخص نماید، ارسال و برچسب واقعی داده‌ها را دریافت کرده و مانند یک الگوریتم برخط نمونه‌ای که در طبقه‌بندیِ آن تردید وجود دارد را با دانستنِ کلاسِ واقعیِ آن (که توسط ناظر برچسب خورده است) به مدلِ آموزشی خود اضافه کند و یادگیری را به صورت پویا و برخط به انجام رساند. در واقع این‌جا نوعی یادگیری فعال برای الگوریتم اتفاق افتاده است.

واحد یادگیری فعال به کمک پیش‌بینی‌های ارائه شده و میزان ضریب انطباق‌پذیری که از سیستم‌های قبلی به عنوان یک تصمیم‌یار ارائه می‌شود و از طریق پرس‌وجو از متخصص انسانی یا هر نوع خبره دیگری، برچسب درست را در خروجی ارائه می‌کند.

اما مهمترین گام این سیستم آن است به جای تولید هشدار مثبت کاذب و به‌منظور جلوگیری از هدر دادن زمان و هزینه، از طریق دانش پرس و جو شده اقدام به تکامل و رانش مفاهیم قبلی و به اینصورت به ذخیره دانش کسب شده اقدام می‌نماید. با این رویکرد دانش مورد استفاده توسط سیستم تولیدکننده هشدار و یا سیستم همبسته‌ساز بروزرسانی می گردد. چرا که رانش مفهوم در طول زمان، نتایج طبقه‌بندی را تغییر می‌دهد که این موضوع به دلیل وقوع تغییرات در مفاهیم و الگوی داده‌ها است. از آنجاکه در دنیای واقعی معمولاً تغییرات بر اساس قواعد خاصی صورت می‌گیرد، لذا بدیهی است که رانش مفهوم نیز از قواعد خاصی پیروی کند.

رانش مفهوم در طول زمان، نتایج طبقه بندی را تغییر می‌دهد که این موضوع به دلیل وقوع تغییرات در مفهوم و الگوی داده ها است. به عبارت دیگر فرض کنید داده‌های با مقادیر خصوصیات مشخص متعلق به کلاس معینی باشد، چنانچه تغییر مفهوم رخ دهد، ممکن است این داده دیگر به کلاس قبلی تعلق نداشته باشد و به کلاس دیگری منتسب گردد. در اثر این رانش مفهوم، مدل ایجاد شده برای طبقه‌بندی، تناسب خود را با داده‌ها از دست خواهد داد و دقت آن کم خواهد شد و حال باید سیستم متناسب با وضعیت و شرایط محیط در هر لحظه بتواند وضعیت و شرایط آتی محیط را پیش‌بینی کند. با این شیوه سیستم یادگیری فعال فراخور نیاز، یکی اقدامات زیر را بر رفع این مسئله اتخاذ می‌نماید.

۱- ایجاد یک نمونه نماینده جدید در ماتریس نمونه‌های نماینده متعلق به سیستم تولیدکننده هشدار به عنوان یک نماینده جدید از یک نوع رفتار جدید

۲- تغییر برچسب یک نمونه نماینده به دلیل تغییر در مفاهیم انواع حملات و یا رفتار عادی در سیستم تولید کننده هشدار

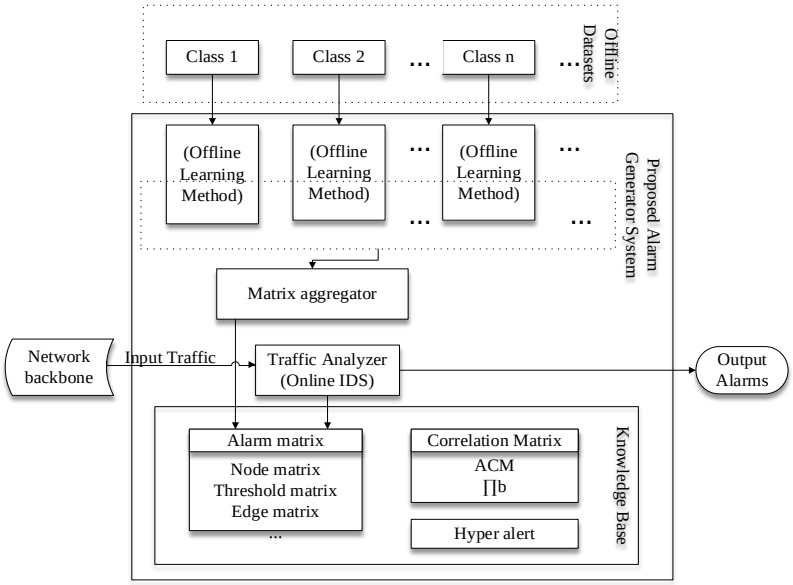
۳- ایجاد یک فراهشدار(فرارفتار) جدید به عنوان نماینده گونه ای جدید از جریان رفتار نرمال یا حمله در سیستم همبسته‌ساز

۴- اصلاح و یا پاک کردن کل و یا یک رفتاردر فراهشدار(فرارفتار) موجود در نمایه‌های رفتار عادی و یا حمله که به دلیل عدم مراجعه و پاسخ‌های غلط امتیازش پایین و منقضی شده است مثلاً:

- فرارفتار نرمالی که دیگر در سازمان وجود ندارد (مانند تغییر پست سازمانی یک کارمند).

- رفتاری که تبدیل به رفتار خصمانه گردیده است (مانند اخراج یک کارمند که منجر به رفتار بدخواهانه می‌گردد).

- فراهشداری که موجب هشدار کاذب بالا می شود و متواتر نیست (اشتباه غیرعمد یک کاربر مانند فراموشی گذرواژه که منجر به تولید هشدار می‌شود).



شکل(۵): سیستم تولید کننده هشدار

ارجاع ترافیک به سیستم همبسته‌ساز حمله یا رفتار عادی به برچسب تخصیص داده شده توسط سیستم تولیدکننده هشدار بستگی دارد و خروجی واحد یادگیری فعال وابسته به انطباق و یا عدم انطباق نتایج حاصل از دو واحد قبلی است. حالات مختلف در جدول (۲) نشان داده شده است:

جدول (۲): حالات ممکن مدل توسعه‌یافته پیشنهادی، اقدامات واحد یادگیری فعال و تغییر مفهوم

خروجی سیستم تولیدکننده هشدار	خروجی همبسته‌ساز (هشدار یا رفتار نرمال)	خروجی/ خروجی یادگیری فعال	ویرایش مفهوم	حالات پیش‌آمده
حمله	همبسته شده	حمله	-	به‌درستی حمله تشخیص داده شد (TP)
حمله	همبسته نشده	حمله	افزودن یک ابررفتار (ابر هشدار) جدید با نمونه ترافیک جدید در همبسته‌ساز	به‌اشتباه همبسته نشد (FN)

ادامه جدول (۲): حالات ممکن مدل توسعه‌یافته پیشنهادی، اقدامات واحد یادگیری فعال و تغییر مفهوم

حمله	همبسته نشده	عادی	افزودن به عمر نمونه نماینده (نهایتاً از بین بردن نودی که منجر به تولید هشدار کاذب می‌شود)	به‌اشتباه حمله تشخیص داده شد(FP)
عادی	همبسته شده	عادی	-	به‌درستی حمله تشخیص داده نشد(TN)
عادی	همبسته نشده	عادی	افزودن یک ابر رفتار (عادی) جدید با نمونه ترافیک جدید در همبسته‌ساز	به‌اشتباه همبسته نشده(FP)
عادی	همبسته نشده	حمله	افزودن به یک نمونه نمایند (نود) جدید در پایگاه دانش سیستم تولیدکننده هشدار	به‌اشتباه حمله تشخیص داده نشد(FN)

● **اهدف تکامل و رانش مفهوم در یادگیری فعال:** این روش علاوه بر استفاده از دانش پرس‌وجو شده در خروجی سیستم، اقدام به بروزرسانی مدل قدیمی نیز می‌نماید که دلیل آن مقابله با ماهیت این نوع از مسائل دنیای واقعی که با پیچیدگی‌های متعددی مانند روش‌های فریب، حملات صفر روزه و تکنیک‌های ناشناخته همراه است، است و همچون یک متخصص انسانی با تغییرات مفاهیم محیط پیرامون تکامل و بروزرسانی می‌گردد و گاهی هم اقدام به فروپاشی و ویرایش مدل‌های قدیمی می‌نماید.

۳-۲- سیستم تولید هشدار

همانطور که در بخش قبل عنوان شد سیستم تولید هشدار از دو ماژول یادگیری اولیه برون خط و یادگیری افزایشی برخط تشکیل شده است.

● **زیرسیستم یادگیری اولیه برون خط:**

این زیرسیستم همانطور که در شکل (۵) نشان داده شده است با استفاده از دو الگوریتم Kmeans و یادگیری شبکه عصبی مانند LSTM [۲۴] یا SOINN [۲۵] تنظیم شده اقدام به خوشه‌بندی کلاس‌های حملات متعارف می‌نماید.

هر یک از کلاس‌های تعریف شده، به‌صورت مجزا به الگوریتم شبکه عصبی پیشنهادی داده می‌شود. که این امر موجب می شود: ۱. کشف تعداد زیرخوشه‌های موجود در یک کلاس و ۲. مکان اولیه مراکز خوشه. این ویژگی منحصر به فرد باعث می شود که توپولوژی زیرکلاس‌های کلاس اصلی استخراج گردد، اما به دلیل ماهیت افزایشی شبکه عصبی، ترتیب داده‌های ورودی منجر به عدم دقت در مکان نمونه‌های نماینده می‌شود. اما پس از استفاده از شبکه عصبی، حالا که تعداد خوشه‌ها و مکان اولیه نمونه‌های نماینده مشخص شده است الگوریتم Kmeans می‌تواند به‌صورت کارامد نسبت به تعیین مکان دقیق نمونه‌های نماینده اقدام نماید.

خروجی این مرحله به یک ماژول تجمیع کننده ارائه می‌شود. این ماژول موظف است که اطلاعات تولید شده از گام پیشین را در ماتریس‌های مشخصه، آستانه شباهت، یالها و ماتریس چگالی مربوط به نمونه‌های نماینده که توسط هر کلاس به صورت مجزا تولید شده‌اند، تجمیع نماید و این دانش برون خط را در پایگاه دانش ذخیره نماید.

● **زیرسیستم یادگیری افزایشی برخط:**

پیش از شروع باید بگوییم که انواع مختلفی از شبکه‌های عصبی وجود دارند ولی در این بخش به‌دلیل مزیت‌ها شبکه های عصبی LSTM برای دسته‌بندی داده‌ها استفاده شده است. در این الگوریتم، لایه ورودی داده‌های خام شامل آدرس و پورت مبدا و مقصد، نوع حمله و زمان ورود بسته‌ها به شبکه را دریافت می‌کند، در این شبکه سه دروازه یا gate وجود دارد که از طریق آن شبکه نسبت به کنترل جریان داده درون خود اقدام می‌نماید.

دروازه فراموشی یا همان Forget gate، وظیفه کنترل جریان اطلاعات از گام زمانی قبلی را دارد. این دروازه مشخص می کند آیا اطلاعات حافظه از گام زمانی قبل مورد استفاده قرار گیرد یا خیر و اگر باید از گام زمانی قبل چیزی وارد شود به چه میزان باشد که این دروازه **رانش مفهوم** در یادگیری فعال را محقق می‌نماید.

دروازه بروزرسانی (Update gate) به دروازه ورودی یا Input gate هم معروف است که وظیفه کنترل جریان اطلاعات جدید را بر عهده دارد. این دروازه مشخص می‌کند آیا در گام زمانی فعلی باید اطلاعات جدید مورد استفاده قرار گیرد یا خیر و اگر بلی به چه میزان، که این دروازه نیز **تکامل مفهوم در یادگیری فعال** را محقق می‌نماید.

دروازه خروجی یا همان Output gate نیز مشخص می‌کند چه میزان از اطلاعات گام زمانی قبل با اطلاعات گام زمانی فعلی به گام زمانی بعد منتقل شود. برای اینکه مشخص کنیم در خروجی از چه محتوایی باید استفاده کنیم از دروازه خروجی بهره می‌بریم که این دروازه نیز عملاً منجر به تنظیم پارامترهای F5 و F6 می‌شود که جزء پارامترهای داخلی الگوریتم LSTM است، چرا که مدل بر اساس تعدد تکرار و میزان شدت شباهت ورودی‌ها بروزرسانی می‌شوند. نحوه عملکرد الگوریتم پیشنهادی برای همبسته‌ساز در الگوریتم (۱) آمده است.

الگوریتم (۱): کد پایتون الگوریتم پیشنهادی LSTM برای همبسته‌سازی

```
Algorithm Correlator
Input: Dataset D, learning rate, number of iterations.
Output: label of Cluster set C.
{
model = Sequential()
return_sequences=True,  model.add(LSTM(256, input_dim=6,
activation='tanh' model.add(Dropout(0.1))
model.add(LSTM(256, return_sequences=True))
model.add(Dropout(0.1)) model.add(LSTM(256,
return_sequences=False))
model.add(Dropout(0.1))
model.add(Dense(8, activation='softmax'))
model.compile(loss='mean_squared_error', optimizer='adam',
metrics=['accuracy'])
history = model.fit(x_train, y_train, batch_size = 500, epochs=20,
validation_data=(x_test, y test), verbose=1, callbacks=fcp, tb]).history
}
```

این نوع شبکه‌ها از مجموعه ای از لایه‌ها پشت سرهم تشکیل شده‌اند، برای ساخت این شبکه ها در کتابخانه کراس پایتون از مدل (model) که اصلی‌ترین ساختار در کراس است و لایه‌ها را

سازماندهی می کند استفاده شده است. اصلی‌ترین و ساده‌ترین نوع مدل هم در این ماژول مدل ترتیبی (Sequential) است که به‌صورت یک مجموعه ترتیبی و خطی از لایه ها که هر لایه تنها با لایه بعد از خود ارتباط دارد تعریف می‌شود.

Hidden _ nodes این آرگومان تعداد سلول‌های عصبی LSTM را مشخص می کند. اگر تعداد آن بیشتری باشد، شبکه قدرتمندتر می‌شود. هر چند، تعداد پارامترهای یادگیری نیز افزایش می یابد. این بدان معناست که برای آموزش شبکه به زمان بیشتری احتیاج دارد.

Timesteps تعداد زمان‌سنجی که می‌خواهید در نظر گرفته شود است. به عنوان مثال اگر می خواهید یک جمله را طبقه بندی کنید، این تعداد، کلمات در یک جمله خواهد بود.

Input _ dim ابعاد ویژگی ها، به عنوان مثال بازنمایی تعداد ویژگی‌ها در بردار خصیصه‌ها است.

Dropout _ value برای کاهش اتصالات، لایه dropout بخشی از اتصالات شبکه ممکن را به‌طور تصادفی می گیرد. این مقدار، درصد در نظر گرفتن اتصالات شبکه برای هر دوره / دسته است.

Learn _ rate نرخ یادگیری را نشان می‌دهد که هر وزن به چه میزان بروز شود.

Decay چقدر نرخ یادگیری با گذشت زمان کاهش یابد.

Return _ sequence اگر True را انتخاب کنیم بعد از هر ورودی مقدار را برمی‌گرداند و اگر false باشد بعد از آخرین ورودی یک بردار را بر می‌گرداند.

Loss میزان خطای شبکه را در هر مرحله نشان می‌دهد و loss function هم تابعی است که این میزان خطا را محاسبه می‌کند. یکی از loss functionهای خیلی رایج mean _ squared _ error هست. ورودی‌های هر loss function جواب مطلوب و همچنین محتوای آخرین لایه fully connected است و بعد سیگمای حاصله از این بردار را به Optimizer می‌دهد و جهت کمینه کردن میزان خطای loss در مراحل بعدی این پروسه بالا را تکرار می‌شود تا خطای شبکه کم شود و به حد آستانه مطلوب ما برسد. به این شیوه **مفهوم یادگیری فعال و ویرایش مفهوم** به وقوع می پیوندد.

در روابط فوق با تعیین batch _ size و تعیین تعداد مراحل اجرا یا epochs مدل خود را اجرا می‌کنیم.

Batch _ size در هر مرحله از یادگیری تعدادی از نمونه ها را آموزش داده و پس از آن پارامترهای شبکه تنظیم می‌شود به این تعداد نمونه batch _ size گفته می شود. البته در انتخاب batch _ size هم باید دقت کنیم که نه زیاد بزرگ باشد و نه زیاد کوچک. از طرفی اگر batch _ size برابر کل نمونه های آموزش باشد گرادیان پایدارتر می شود و شبکه به کندی همگرا می شود و از طرف دیگر اگر batch _ size خیلی کوچک انتخاب شود گرادیان ناپایدار خواهیم داشت که به طبع آن مجبوریم نرخ یادگیری را کاهش بدهیم.

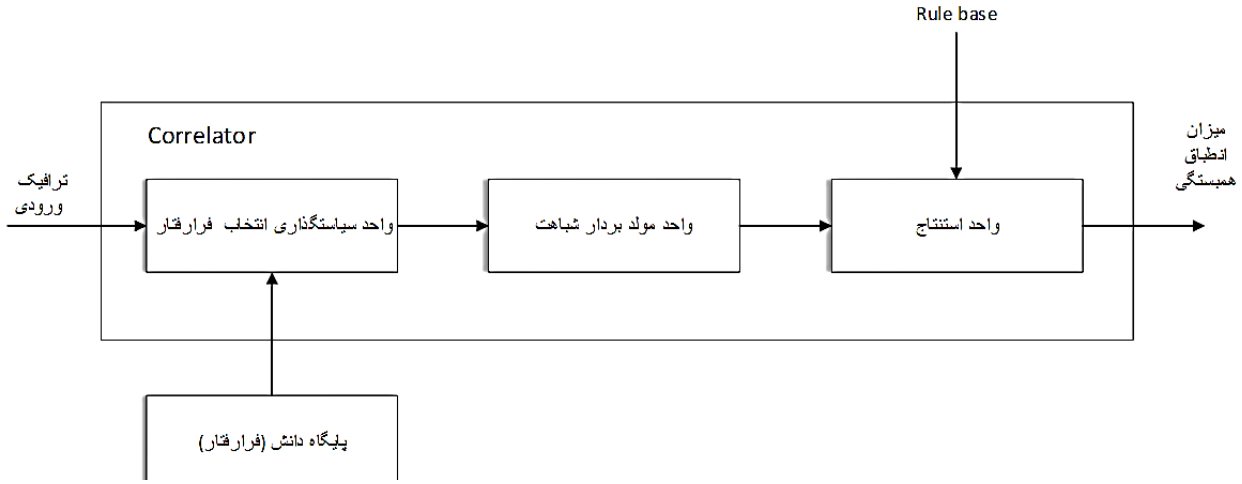
Epochs نیز برای بدست آوردن نتایج شبیه سازی برنامه را به طور مستقل در دفعات متعدد اجرا کرده که نتایج معادل میانگین ۲۰ اجرای مستقل برنامه است.

۳-۳- سیستم همبسته‌ساز

همانگونه که در ابتدا شرح داده شد سیستم همبسته‌ساز موظف است با دریافت هشدار و برچسب نوع حمله از سیستم تولیدکننده هشدار، احتمال همبستگی با هشدارهای پیشین را محاسبه کند. هشدارهای پیشین در پایگاه دانش ذخیره شده‌اند و به مرور زمان و با کسب دانش بیشتر، اطلاعات موجود در این ماژول نیز تکامل می‌یابد.

همانطور که قبلاً نیز اشاره شد در مدل توسعه یافته، علاوه بر هشدارها می‌توان رفتارهای عادی را نیز طبق همین روال همبسته نمود به‌طوری‌که ترافیک مزبور به سیستم همبسته‌ساز رفتار عادی و یا هشدار به فراخور تشخیصی که سیستم تولیدکننده هشدار داده است، ارجاع داده شود. لازم به ذکر است فرآیند همبسته‌سازی در هر یک از دو حالت هیچ تفاوتی نداشته، لذا در این بخش هرجا سخن از رفتار می‌شود منظور هشدار و یا رفتار عادی و بالعکس است و در هر یک از آنان تفاوتی در سیستم ارائه شده، وجود نخواهد داشت.

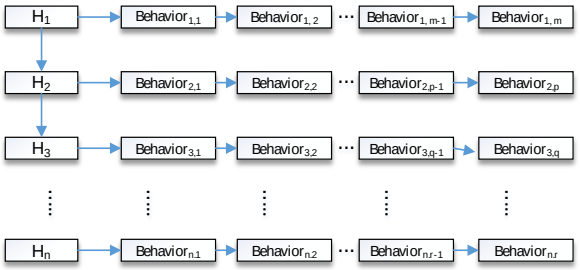
هشدارهای تولید شده به‌منظور کشف مثبت کاذب و خطاهای سیستم تشخیص نفوذ به بخش همبسته‌ساز منتقل می‌شود. سیستم همبسته‌ساز نسبت به بررسی برخی دیگر از ویژگی‌های هشدار تولید شده با کمک دانش ذخیره شده در پایگاه دانش اقدام نموده و میزان همبستگی آن را می‌سنجد. در صورتی‌که میزان همبستگی از حد آستانه بیشتر باشد، رفتار را (حمله و یا عادی) تشخیص می‌دهد. در غیر اینصورت با پرسش از خبره پاسخ مناسب به هشدار تولید شده می‌دهد و با کمک یادگیری فعال اقدام به ویرایش و ذخیره دانش جدید در پایگاه دانش می‌نماید. در بخش‌های بعدی جزئیات سیستم تولید کننده هشدار پیشنهادی شرح داده شده است.



شکل(۶): سیستم همبسته‌ساز رفتار

۳-۳-۱- واحد انتخاب فرارفتار

همانطور که در بخش گذشته شرح داده شد وقتی واحد سیاست‌گذاری انتخاب هشدارها، هشداری مانند b را برای بررسی میزان همبستگی با a معرفی می‌نماید. همانطور که در شکل (۷) نشان داده شده است زمانی‌که دو هشدار a و b همبسته تشخیص داده شوند هشدار a به فراهشدا (و یا فرارفتار) که b به آن تعلق داشته، اضافه خواهد شد.



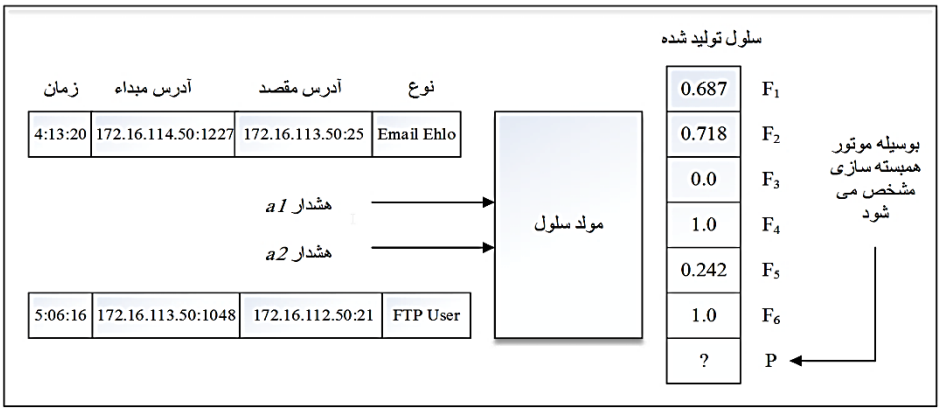
شکل(۷): نحوه ذخیره یک رفتار در ساختار فرفتارها

علاوه براین با توجه به اینکه a با کلیه هشدارهای موجود در یک فرارفتار تشکیل یک سناریو را می‌دهد. منطقی است که فرض کنیم a با بسیاری دیگر از هشدارهای موجود در این سناریو نیز دارای همبستگی بالایی می‌باشد. با توجه به این موضوع میزان همبستگی a با کلیه هشدارهای که یک فرارفتار نماینده آن است، محاسبه می‌شود به این‌ترتیب پس از شناسایی سناریویی (فرارفتاری) که هشدار a به آن تعلق دارد برای حفظ و ذخیره دانش کسب شده و توسعه آن و همچنین استفاده‌های بعدی از این دانش، مدل آموخته شده بروز می‌شود. با توجه به اینکه در فرآیند یادگیری بعضی از ویژگی‌ها از این ماتریس‌ها استخراج می‌شوند لذا دانش کسب شده از همبسته‌سازی a در همبسته‌سازی‌های بعدی استفاده خواهد شد که این به معنای یادگیری مستمر سیستم همبسته‌سازی در طول زمان و بکارگیری دانش‌های کسب شده خواهد بود. در صورتیکه a با هیچ هشداری در فراهشدارهای موجود همبسته نشود (کمتر از آستانه همبستگی باشد) یک فراهشدار جدید ساخته خواهدشد و a به عنوان تنها هشدار موجود در آن قرار خواهد گرفت. به این‌ترتیب یک سناریوی احتمالی جدید با هشدار a آغاز می‌شود. میزان همبستگی دو سلول a و b فقط در صورتی توسط لایه دوم همبسته‌سازی مشخص می‌شود که میزان انطباقش با c بیشتر از از حد آستانه باشد در غیر اینصورت برای تعیین وضعیت هشدار با استفاده از دانش متخصص انسانی (همان اطلاعات برچسب صحیح) به تکامل و یا ویرایش مفهوم فعلی در پایگاه دانش اقدام می‌نماید.

۳-۳-۲- واحد مولد بردار شباهت

فرض کنید $a2$آخرین هشداری است که توسط سیستم تشخیص نفوذ تولیدشده است (هشدار جاری مورد بررسی) و $a1$ نیز هشداری است که برای بررسی میزان همبستگی انتخاب شده است. طبیعتاً هر یک از دو هشدار $a1$ و $a2$ حاوی اطلاعات متعددی می‌باشند اطلاعاتی از قبیل زمان بروز هشدار، آدرس IP و آدرس پورت مقصد، IDS تولید کننده هشدار، آدرس فیزیکی مبدا و مقصد، شماره پروتکل، اولویت هشدار، نوع هشدار(نوع هشدار همان مقداری است که توسط IDS تعیین می‌شود) و غیره.

همانطور که در شکل (۸) نشان داده شده است، واحد مولد بردار شباهت از شش ویژگی مورد استفاده در [۱۶] برای ساختن بردار شباهت استفاده کرده است.



شکل(۸): واحد مولد بردار شباهت

از بین شش ویژگی مورد استفاده چهار ویژگی آن مستقیماً از دو هشدار$a1$ و$a2$ استخراج می‌شوند ولی دو ویژگی دیگر با بررسی دانش ذخیره شده در پایگاه دانش محاسبه می‌شوند. این دو ویژگی باعث می‌شوند که سیستم به کاوش در سابقه کارکرد خود بپردازد و از آن برای همبسته کردن هشدارهای جدید استفاده کند. می‌توان گفت به کمک استفاده از این دو ویژگی به نوعی مفهوم داده‌کاوی و استفاده ازروش‌های آماری نیز وارد محاسبه احتمال همبستگی می‌شود ویژگی‌های مورد استفاده در تولید هر بردار شباهت عبارتند از:

F1: میزان شباهت آدرس IP مبدا $a1$ و$a2$ است که مقدار آن بین صفر و یک است.

F2: میزان شباهت آدرس IP مقصد $a1$ و$a2$ است که مقدار آن بین صفر و یک است.

F3: برابر بودن پورت مقصد $a1$ و$a2$ است که مقدار آن صفر یا یک است.

F4: برابر بودن آدرسIP مبدا هشدار دوم $(a2)$ با آدرس IP مقصد هشدار اول $(a1)$ است که مقدار آن صفر و یا یک است.

F5: شدت همبستگی سابقه‌نگر بین هشدارهای قبلی از نوع 1 و $a2$ که بر اساس شدت میزان شباهت دو رفتار براساس سابقه‌ای که در مدل یادگرفته شده وجود دارد، استخراج می‌شود و مقدار آن بین صفر و یک می‌باشد.

F6: میزان تعداد دفعاتی است که هشدارهای قبلی از نوع 1 و $a2$ با هم همبسته شده‌اند که براساس سابقه‌ای که در مدل یادگرفته شده وجود دارد، استخراج می‌شود، این مقدار در ابتدا صفر است و به مرور و با کسب دانش می‌تواند تا ۱ افزایش یابد.

برای محاسبه میزان شباهت دو آدرس IP می توان از تعداد بیت‌های مشابه در دو آدرس با شمارش از با ارزش‌ترین بیت استفاده کرد. اگر این تعداد را برابر n فرض کنیم مقدار شباهت دو آدرس IP در IPv4 از رابطه ی زیر به دست می‌آید:

$$Simil\ rity(IP_1, IP_2) = n/32 \qquad (۱)$$

مثلاً نمایش‌های دودویی دو آدرس 192.168.10.60 و 192.168.42.25 که به صورت زیر می‌باشند.

$$\begin{matrix} 192.168.10.60: & 11000000.10101000.00001010.00111100 \\ 192.168.42.25: & 11000000.10101000.00101010.00011001 \end{matrix} \qquad (۲)$$

دارای ۱۸ بیت مشابه در بیت‌های پر ارزش خود هستند لذا میزان شباهت آنها 18/32 و برابر ۰،۵۶ است. بدیهی است در صورتیکه شباهت دو آدرس IP کامل باشد مقدار جواب یک و در صورتیکه کاملاً متفاوت باشند مقدار جواب صفر خواهدبود.

قبل از استفاده از هر آسیب‌پذیری موجود در یک سرویس، مهاجم بایستی پورت مربوطه را بررسی کند تا از باز بودن یا بسته بودن آن اطلاع حاصل نماید. چون حملات معمولاً بر اساس آزمون و خطا صورت می‌گیرد لذا تعداد زیادی هشدار تولید می‌کند که درصورتی که در دو هشدار یک شماره پورت مقصد مورد بررسی قرارگرفته باشد می‌تواند نشان‌دهنده پویش شبکه برای آن شماره پورت باشد. لذا برابر بودن شماره پورت مقصد دو هشدار نیز می‌تواند گویای ارتباط دو هشدار باشد. مقدار این ویژگی در صورت یکسان بودن پورت مقصد دو هشدار برابر یک و در غیر این صورت برابر صفر می‌باشد.

ویژگی F5 که بین صفر و یک تغییر می‌کند، شدت همبستگی است که در اصل میزان همبسته شدن دو هشدار را بر اساس تجربه سیستم از همبسته شدن هشدارهای مشابه قبلی نشان می‌دهد این ویژگی از ماتریسی با عنوان ماتریس شدت همبستگی که در . در ابتدای شروع به کارِ سیستم، مقدار شدت همبستگی صفر است و در تعیین احتمال همبسته شدن تاثیر ندارد ولی به مرور زمان و بر اثر مشاهده هشدارها این مقدار ممکن است افزایش یابد وحتی به ۱ برسد به خاطر طبیعتِ همبسته‌سازی، این امکان وجود دارد که دو هشدار بدون داشتن مقدار بالایی از شدت همبستگی و به خاطر سایر ویژگی‌هایشان با یکدیگر همبسته شوند. شدت همبستگی در زمانی که عدم قطعیت وجود دارد می‌تواند راهگشا باشد. مثلاً میزان شباهت آدرسIP مبدا به علت استفاده از آدرس‌های جعل شده پایین است اما سابقه همبسته‌سازی در قالب شدت همبستگی گویای

احتمال بالای همبستگی دو نوع هشدار است.

مثال فرض‌کنید در یک سیستم دو هشدار مانند $a1$ و $a2$ از نوع$t1$ و $t2$با مشخصات زیر دریافت می‌شوند و سیستم سعی دارد همبستگی آنها را ارزیابی کند.

$$(a_1.SrcIP = a_2.SrcIP)and \\ (a_1.DestIP = a_2.DestIP)and \\ (a_1.DestPort = a_2.DestPort) \quad (۳)$$

بدیهی‌است که تصمیم برای همبستگی دو هشدار فوق علیرغم کم بودن شدت همبستگی سابقه‌نگر تصمیم روشنی می‌باشد. در دفعات بعد و با دریافت مجدد هشدارهایی از نوع $t1$و $t2$و در شرایط عدم قطعیت، (مثلا اگر مقدار شباهت آدرس IP برابر ۰,۵ باشد) تصمیم‌گیری به‌کمک شدت همبستگی سابقه‌نگر امکان‌پذیر می‌شود و چون سابقه قبلی همبستگی بین هشدارهای از نوع $t1$و $t2$ بالا بوده است این ویژگی کم بودن شباهت آدرس‌هایIP مبدا را جبران می‌کند.

شدت همبستگی بین دو نوع هشدار نقشی اساسی در تحلیل الگوی حملات دارد و رابطه ی سببی بین دو هشدار را مشخص می‌کند. در [۲۶] قبل از آنکه هر هشداری بتواند در سیستم همبسته‌ساز مورد پردازش قرارگیرد لازم است تا از این هشدار و هشدارهای قبلی اطلاعاتی استخراج و میزان شباهتش با آنها سنجیده شود. داده‌های آموزشی بکاررفته در همبسته‌ساز مجموعه‌ای محدود از دانش اولیه است که برچسب گذاری شده است. این دانش مجموعه‌ای از حالات ساده همبستگی است که با دانش ابتدایی از مفاهیم امنیت و سناریوهای حملات قابل توصیف می‌باشند.

ویژگی ششم F6 فرکانس یا تعدد نسبی همبسته شدن این دو نوع هشدار را بنا به تجربه سیستم نشان می‌دهد. این ویژگی یک ویژگی آماری است و باز هم در شرایط عدم قطعیت راهگشا خواهد بود. این ویژگی همچنین یک نقش تنظیم کننده برای F5 را نیز به عهده دارد به این معنا که اجازه نمی‌دهد که پس از یک یا دو بار همبسته شدنِ دو هشدار، تاثیر F6 زیاد باشد زیرا با چند بار همبستگی دو هشدار هنوز از لحاظ آماری نمی‌توان آنها را وابسته فرض کرد و به مقدار F5 اعتماد کرد ولی با تکرار این همبستگی، می‌توان قبول کرد که دو هشدار واقعاً وابسته هستند. در این زمان ویژگی F6 نیز افزایش یافته و به ۱ نزدیک شده است و ویژگی F5 کاملا بالغ شده و می‌توان به آن اعتماد نمود. طبیعتاً F6 نیز در ابتدا صفر می‌باشد ولی با تکرار همبسته شدنِ هشدارها مقدار آن افزایش می‌یابد. سرعت این افزایش یکی از پارامترهای قابل کنترل در سیستم می‌باشد. اطلاعات دو ویژگی اخیر در قالب ماتریسی ذخیره می‌شوند که هر درایه ماتریس شدت همبستگی و فرکانس همبستگی دو هشدار را نشان می‌دهد.

۳-۳-۳- موتور همبسته‌سازی

داده‌های آموزشی به‌کاررفته در همبسته‌ساز مجموعه‌ای محدود از دانش اولیه است که به‌صورت دستی تعریف شده و برچسب گذاری شده است. این دانش مجموعه‌ای از حالات ساده همبستگی است که با دانش ابتدایی از مفاهیم امنیت و سناریوهای حملات قابل توصیف می‌باشند.

تعداد داده‌های آموزشی تعریف شده در همبسته‌ساز بسیار محدود می‌باشد به طوری که با استفاده از ۲۱ داده آموزشی عمومی ترین حالت‌های همبستگی بین دو هشدار تعریف می‌شوند. از ۲۱ داده آموزشی ۱۸ داده از داده‌های بکار رفته در [۱۶] می‌باشد و سه داده ی جدید برای افزایش دقت به آنها اضافه شده است. برای استفاده از این داده‌ها در لایه ی همبسته‌سازی مبتنی بر قوانین فازی، همین داده‌ها با تغییرات اندکی مورد استفاده قرار می‌گیرند به طوری که از واژگان فازی مانند *"low", "high", "medium"* و *"low" "Moderate"* برای نشان دادن مقدار هر یک از ویژگی‌ها استفاده می‌شود.

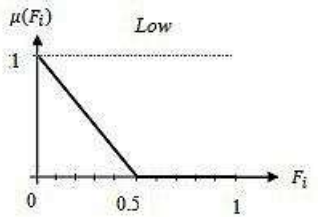
جدول (۳): قوانین استفاده شده در سیستم همبسته‌ساز

	F1	F2	F3	F4	F5	F6	Class
1	high	high	1	0	high	high	20
2	low	low	1	1	low	high	15
3	high	high	0	0	low	low	16
4	low	low	1	0	low	high	9
5	high	high	0	0	medium	medium	18
6	medium	high	0	0	medium	medium	17
7	medium	medium	1	1	medium	medium	19
8	medium	medium	0	0	low	mod_low	3
9	low	high	0	0	low	mod_low	5
10	high	medium	1	0	medium	mod_low	14
11	low	low	0	0	low	low	1
12	medium	high	0	0	high	high	18
13	medium	medium	1	0	high	high	17
14	high	high	1	0	low	low	19
15	medium	medium	0	0	medium	low	4
16	low	low	1	0	high	high	14
17	low	low	1	1	high	high	19
18	low	low	0	1	medium	low	17
19	low	low	1	1	medium	medium	18
20	low	low	0	1	low	low	17
21	medium	medium	0	0	medium	high	7

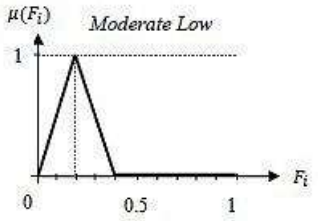
جدول (۳) بیست و یک قانون تعریف شده در اولین لایه همبسته‌سازی را نشان می‌دهد. همان‌طور که مشخص است هر قانون از یک مقدم، شامل بررسی وضعیت شش ویژگی f1 تا f6 و یک تالی شامل شماره کلاس انتساب یافته به آن وضعیت، تشکیل شده است. فرم کلی قوانین مورد استفاده به‌صورت زیر می‌باشد.

$$if(F_1 = V_1) and (F_2 = V_2) and \dots (F_6 = V_6)\ Then\ (Class = C) \quad (۴)$$

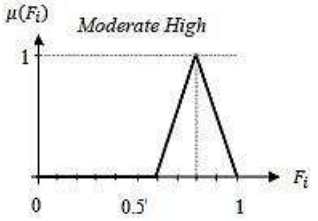
تابع عضویت فازی خطی هر یک از واژگان مورد استفاده در تعریف قوانین در شکل (۹) قابل مشاهده است.



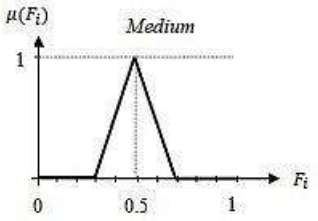
$$\mu(F_i) = \begin{cases} 1 - 2 * F_i & if\ 0.0 \le F_i < 0.5 \\ 0 & if\ 0.5 \le F_i \le 1.0 \end{cases}$$



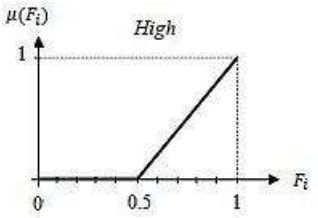
$$\mu(F_i) = \begin{cases} 5 * F_i & if\ 0.0 \le F_i < 0.2 \\ 2 - 5 * F_i & if\ 0.2 \le F_i < 0.4 \\ 0 & if\ 0.4 \le F_i \le 1.0 \end{cases}$$



$$\mu(F_i) = \begin{cases} 0 & if\ 0.0 \le F_i < 0.3 \\ 5 * F_i - 1.5 & if\ 0.3 \le F_i < 0.5 \\ 3.5 - 5 * F_i & if\ 0.5 \le F_i < 0.7 \\ 0 & if\ 0.7 \le F_i \le 1.0 \end{cases}$$



$$\mu(F_i) = \begin{cases} 0 & if\ 0.0 \le F_i < 0.6 \\ 5 * F_i - 3 & if\ 0.6 \le F_i < 0.8 \\ 5 - 5 * F_i & if\ 0.8 \le F_i \le 1.0 \end{cases}$$



$$\mu(F_i) = \begin{cases} 0 & if\ 0.0 \le F_i < 0.5 \\ 2 * F_i - 1 & if\ 0.5 \le F_i \le 1.0 \end{cases}$$

شکل(۹): تابع عضویت هر یک از واژگان

برای تعیین میزان سازگاری یک بردار شباهت مانند x بامقادیر *(v1, v2, v3, v4, v5, v6)* برای ویژگی‌هایش، با یک قانون مانند Rj که با مقادیر فازی برای ویژگی‌هایش، با یک قانون مانند *(v1, v2, v3, v4, v5, v6)* تعریف شده است از میانگین تعلق مقادیر ویژگی‌ها در آن سلول به مجموعه‌های موجود در قانون Rj استفاده می‌شود. رابطه محاسبه میزان سازگاری x با Rj به صورت رابطه ۹ بیان می‌شود [۲۷].

$$Compatibility(x, Rj) = \frac{1}{n}\sum_{i=1}^{n} \mu(v_i, V_i) \quad (۵)$$

درمثال زیر، با استفاده از رابطه ۹ و توابع عضویت شکل ۹ میزان سازگاری بردار شباهت x با R1 (قانون ۱ از جدول ۳) برابر ۰,۶۴۴ محاسبه می‌شود.

$$x: (v_1 = 1, v_2 = 0.81, v_3 = 0, v_4 = 0, v_5 = 0.62, v_6 = 1)$$
$$R_1: \begin{pmatrix} V_1 = high, V_2 = high, V_3 = 1, V_4 = 0, V_5 = \\ high, V_6 = high, Class = 19 \end{pmatrix} \quad (۶)$$

$$\mu(v_1, V_1) = \mu(1, high) = 1 \qquad (v_4 = V_4) \Rightarrow \mu = 1$$
$$\mu(v_2, V_{21}) = \mu(0.87, high) = 0.625 \quad \mu(v_5, V_5) = \mu(0.62, high) = 0.24$$
$$(v_3 \neq V_3) \Rightarrow \mu = 0 \qquad \mu(v_6, V_6) = \mu(1, high) = 1$$
$$Compatibility(x, R_j) = \frac{1 + 0.625 + 0 + 1 + 0.24 + 1}{6} = 0.644 \quad (۷)$$

میزان سازگاری بردار شباهت x با تک‌تک قوانین جدول ۲ به همین شکل بررسی می‌شود تا بتوان سازگارترین قانون با x را یافت. میزان سازگاری x با قوانین بیست و یک گانه مورد استفاده به ترتیب عبارت است از:

$$R_1: 0.644, R_2: 0.167, R_3: 0.603, R_4: 0.333, R_5: 0.67, R_6: 0.503, R_7: 0.067, R_8: 0.333,$$
$$R_9: 0.437, R_{10}: 0.4, R_{11}: 0.333, R_{12}: 0.644, R_{13}: 0.373, R_{14}: 0.437, R_{15}: 0.4, R_{16}: 0.373 \quad (۸)$$
$$R_{17}: 0.207, R_{18}: 0.233, R_{19}: 0.067, R_{20}: 0.167, R_{21}: 0.567$$

پس از بررسی بردار شباهت x باهمه قوانین موجود سه قانون با بالاترین میزان سازگاری با x انتخاب می‌شوند. در ادامه شماره کلاس انتسابی به سلول x از طریق شماره کلاس موجود در تالی قوانین مذکور به دست می‌آید.

پس از تعیین شماره کلاسی نهایی (C) در هر مرحله لازم است شماره کلاس تبدیل به احتمال شود برای این کار از رابطه ۹ استفاده می‌شود:

$$P = \frac{C - 1}{\lambda} + \frac{1}{2\lambda} \quad (۹)$$

که در آن λ تعداد کل کلاس‌ها هست (۲۰ =λ) با توجه به نحوه تعریف بازه‌های مربوط به هر شماره کلاس در رابطه ۹ از جمله $\frac{۱}{۲\lambda}$ برای انتقال مقدار احتمال به وسط بازه تعریف شده استفاده می‌شود.

و در نهایت هشدار مورد نظر به همراه احتمال همبستگی و میزان سازگاری به لایه سوم یادگیری فعال به‌منظور تصمیم‌گیری ارجاع داده می‌شود.

۴- نتایج آزمایشگاهی

۴-۱- مجموعه داده‌ها

مجموعه داده‌های زیادی در حوزه تشخیص نفوذ ارائه شده است اما بزرگترین چالش این حوزه عدم وجود یک مجموعه داده فراگیر به‌منظور مقایسه روش‌های موجود است. یکی از مهمترین چالش‌ها درتحقیقات و توسعه سیستم‌های جدید تشخیص نفوذ مبتنی بر ناهنجاری فقدان مجموعه داده مناسب و قابل دسترسی به‌صورت عمومی است. این چالش باعث شده آموزش، ارزیابی و مقایسه ی نوآوری‌ها و روش‌های جدید در زمینه تشخیص نفوذ مبتنی بر ناهنجاری مشکل شود. در ادامه مجموعه‌ای از پرکابردترین مجموعه داده‌ها که مقالات معتبر از آنها استفاده نموده‌اند توضیح داده می شود.

DARPA2000. گروه تکنولوژی و سیستم‌های سایبری در آزمایشگاه لینکلن در دانشگاه ام‌آی‌تی [۲۸] با حمایت آژانس پروژهشهای پیشرفته دفاعی و آزمایشگاه تحقیقاتی نیروی هوایی ایالات متحده در سال‌های ۱۹۹۸ و ۱۹۹۹ دست به جمع‌آوری و انتشار اولین داده‌های استاندارد برای ارزیابی IDSهای شبکه زد این ارزیابی احتمال تشخیص و احتمال هشدارهای نادرست را برای IDSهای تحتِ بررسی اندازهگیری نمود و نقش قابل توجه در پیشرفت تحقیقات مرتبط با IDS و هدفمندسازی آن‌ها داشت. نتیجه این کار دو مجموعه دادهای DARPA 1999 و DARPA1998 است که هنوز هم به عنوان شناخته شده‌ترین و پرکاربردترین مجموعه‌های دادهای در ارزیابی IDSها استفاده می‌شوند. داده‌های DARPA 2000 با تاکید بیشتر بر حملات چند مرحله‌ای و شناسایی سناریوی این حملات تدارک دیده شده است. دو حمله چند مرحله ای بنام LLDOS1.0 و LLDOS2.0 در این مجموعه تعبیه شده است [۲۸]. در این مقاله با استفاده از این دو سناریو به ارزیابی ماژول همبسته‌ساز از مدل پیشنهادی پرداخته می‌شود.

KDD CUP99. مجموعه داده KDD CUP99 یکی از پراستفاده ترین مجموعه داده‌ها در زمینه تشخیص نفوذ است که به صورت عمومی در دسترس است.

مجموعه داده KDD CUP 99 دارای ۴۱ خصیصه است [۲۹]. تعدادی از این خصیصه‌ها به صورت پیوسته و تعدادی از آنها به صورت گسسته هستند. تعداد خصیصه‌های مورد استفاده به‌منظور ارزیابی تکنیک‌های داده کاوی بر روی این مجموعه داده متفاوت است، گاهی تمامی این خصیصه‌ها و گاهی زیرمجموعه‌ای از این خصیصه‌ها مورد استفاده قرار می‌گیرد.. این خصیصه‌ها به سه دسته اصلی، محتوی و ترافیک تقسیم‌بندی شده‌اند از میان خصیصه‌های موجود در این مجموعه داده Service Type و Flag و Protocol Type مقادیر عددی دارند که بایستی به جای آن تنها سه دسته خصیصه مقادیر عددی قرار داده شود. یکی از روش‌هایی که می‌توان برای این منظور استفاده کرد این است که در تمام داده‌ها فرکانس هر یک از مقادیر موجود در این سه خصیصه را محاسبه کرد و به‌جای این مقادیر استفاده کرد و یا اینکه می‌توان از این سه خصیصه در اجرای الگوریتم استفاده نکرد. اما در این مقاله به‌منظور حفظ خواص فیلدهای نامینال(داده‌های طبقه‌بندی) و استفاده حداکثر از ظرفیت دیتاست از تابع get_dummies در پایتون اقدام شده است.

N-BaIoT. [۳۰] این مجموعه داده مربوط به داده‌های ترافیک واقعی جمع آوری شده از ۹ دستگاه تجاری اینترنت اشیا است که در مخزن داده یادگیری ماشین دانشگاه کالیفرنیا بارگذاری شده است و توسط گروه مهندسی نرم‌افزار و سیستم‌های‌اطلاعاتی دانشگاه بن گوریون در اسرائیل در سال ۲۰۱۸ ایجاد شده است که یک مجموعه داده بروز و استاندارد است و به همین دلیل به عنوان مطالعه موردی، در این تحقیق شبیه‌سازی آن صورت گرفته است. این داده‌ها حاوی آمار ترافیک برای شبکه اینترنت اشیا با ترافیک عادی و حمله است این ۹ دستگاه شامل ترموستات، مانیتور کودک، یک وب‌کم، دو درگاه و چهار دوربین امنیتی است که از طریق وای فای به چندین نقطه دسترسی وصل می شوند. باتنت مهاجمین دستگاه‌های آسیب پذیر را در شبکه جستجو می‌کنند و بدافزارها را به دستگاه‌های آسیب‌پذیر تزریق می‌کنند سپس کنترل دستگاه‌های به خطر افتاده را به دست می گیرند و از آنها به عنوان بخشی از شبکه باتنت برای عمل انجام حملات در مقیاس بزرگ مانند حملات انکار سرویس بر روی کل شبکه استفاده می نمایند هر نمونه داده ۱۱۵ ویژگی دارد. برای هر دستگاه داده‌ها در هر دو شرایط عادی و چندین نوع حمله شامل بش لایت و میرای انجام می‌شود.

همانطور که گفته شد مجموعه داده N-BaIoT از یک مجموعه‌داده ترافیک نرمال و دو مجموعه داده از باتنت‌های مشهور میرای و بش لایت بر روی ۹ دستگاه تجاری تشکیل شده است در این پژوهش از مجموعه داده دستگاه نظارت کودک برای شبیه‌سازی مدل پیشنهادی استفاده شده که شامل یک کلاس داده نرمال، سه کلاس حملات میرای شامل (Ack, scan, syn) و چهار کلاس حملات بش لایت شامل (combo, tcp, junk, scan) هست این داده‌ها دارای ۲۳ ویژگی می‌باشند که در پنج پنجره زمانی استخراج و سپس پنج بردار بیست و سه بعدی از هر پنجره در یک بردار منفرد ۱۱۵ بعدی قرار می‌گیرد.

۴-۲- معیارهای ارزیابی

۴-۲-۱- معیار ارزیابی همبسته‌ساز فازی افزایشی

معمولا سه معیار درستی[1]، تمامیت[2] و نرخ خطا[3] در بررسی سیستم‌های همبسته‌ساز استفاده می‌شود [۱۷]. درستی نشان‌دهنده میزان صحتِ هشدارهای انتخاب شده در سناریوی استخراجی است. تمامیت نشان‌دهنده میزان کامل بودن سناریوی تولیدشده توسط همبسته‌ساز نسبت به سناریوی مطلوب هست. نرخ خطا در همبسته‌سازی نشان‌دهنده میزان خطا در انتخاب هشدارها برای همبسته‌سازی است.

$$Soundness = \frac{TC}{TC + FC} \quad (۱۰)$$

$$Completeness = \frac{TC}{TC + FN} \quad (۱۱)$$

$$False\ Correlation\ Rate = \frac{FC}{TC + FN} \quad (۱۲)$$

TC هشدارهایی که به درستی همبسته شده‌اند، TN نشان دهنده هشدارهایی است که که به درستی با هم همبسته نشده است و FC همه ی هشدارهایی که به غلط همبسته و FN هشدارهایی که به غلط همبسته نشده‌اند را نشان می‌دهد.

۴-۲-۲- معیار ارزیابی مدل پیشنهادی تشخیص نفوذ

روشن است که مقدار دو معیار اول بین صفر و یک متغیر است و هرچه به یک نزدیکتر باشد دقت سناریوی استخراجی بیشتر بوده است. مقدار معیار سوم بزرگتر یا مساوی صفر است و هرچقدر میزان آن کمتر باشد نشان دهنده ی دقت بیشتر در سناریوی استخراجی است. معیارهای معرفی شده در کنار هم می‌توانند نشان دهنده دقت یک سناریو می‌باشند و هر یک جنبه‌ای از دقت آن را بیان می‌کنند.

دقت به‌دلیل توزیع نامتعادل رده‌ها یک معیار قابل اتکا برای ارزیابی سیستم‌های تشخیص نفوذ نیست. دستیابی به دقت بالا برای کلاس اقلیت که وابسته به اندازه مجموعه آموزش و تست می‌باشد، آسان است. در ادامه تعدادی از معیارهای مفید برای ارزیابی سیستم‌های تشخیص نفوذ که مستقل از اندازه مجموعه داده‌های تست و یادگیری می‌باشد، نشان داده شده است.

$$DR = \frac{TP}{TP + FN} \quad (۱۳)$$

$$FAR = \frac{FP}{TN+FP}$$
$$PPV = \frac{TP}{TP+FP} \quad (۱۴)$$

$$MCC = \frac{TN * TP - FN * FP}{\sqrt{(TN + FP)(TN + FN)(TP + FN)(TP + FP)}} \quad (۱۵)$$

$$F - value = \frac{2 * PR * DR}{(PR + DR)} \quad (۱۶)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (۱۷)$$

در معادلات (۶-۴) تا (۱۱-۴) TP هشدارهایی است که به درستی به عنوان حمله شناسایی شده است، TN نشان دهنده هشدارهایی است که که به درستی به عنوان رفتار عادی شناسایی شده است و FP همه هشدارهایی است که به غلط به عنوان حمله شناسایی شده است و FN هشدارهایی است که به غلط به عنوان رفتار عادی شناسایی شده است روشن است که مقدار دو معیار اول بین صفر و یک تغییر می‌کند و هرچه به یک نزدیکتر باشد دقت سناریوی استخراجی بیشتر بوده است. مقدار معیار سوم بزرگتر یا مساوی صفر است و هرچقدر میزان آن کمتر باشد نشان دهنده ی دقت بیشتر در سناریوی استخراجی است.

متاسفانه در مقالات از اسامی مختلفی برای معیارهای معادلات ۱۶ تا ۲۱ عنوان شده است و این موضوع کار را برای مقایسه در حوزه یادگیری ماشین و تشخیص نفوذ سخت می‌نماید، لذا در جدول (۴) نام‌های متفاوتی که برای معیارهای معادلات ۱۶ تا ۲۱ عنوان شده آورده شده است.

جدول (۴): نام‌های متفاوت معیارهای ارزیابی

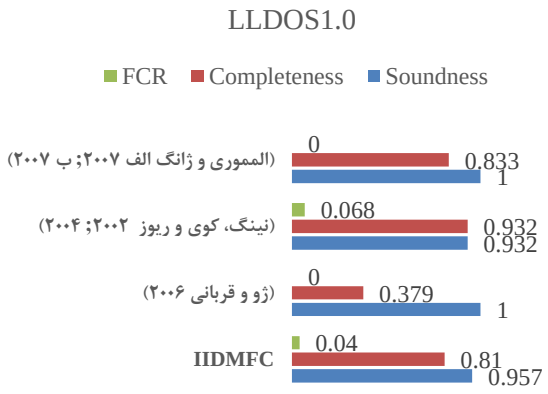
DR	*Detection rate, True Positive rate, tp rate, hit rate, Sensitivity, Recall, Effectiveness*
FAR	*False alarm rate, FPR, fp rate, fall out*
PPV	*Positive predictive value, PR, Precision, Efficienc*
MCC	*Matthews correlation coefficient, MCC*
F-value	*F-score, F-measure, F1*
Accuracy	

معیارهای معرفی شده در کنار هم می‌توانند نشان دهنده ی دقت یک سناریو باشند و هر یک جنبه‌ای از دقت آن را بیان می‌کنند. به عنوان مثال در ارزیابی مسائلی همچون تشخیص نفوذ که عدم توازن در برچسب‌ها به شدت وجود دارد معیاری مثل Accuracy نمی‌تواند خیلی مفید باشد چرا که معمولاً در یک مجموعه‌داده حجم بسیار کمتری از حملات نسبت به تعداد رفتارهای نرمال وجود دارد و بدلیل عدم تعادل داده‌ها در این نوع مسائل نیاز به معیارهایی همچون PPV(Precision) و F1(value) بخاطر در نظر گرفتن نرخ هشدار مثبت کاذب می‌تواند نقش مکمل به‌سزایی را در ارزیابی دقت داشته باشد.

[1] Soundness
[2] Completeness
[3] Error Rate

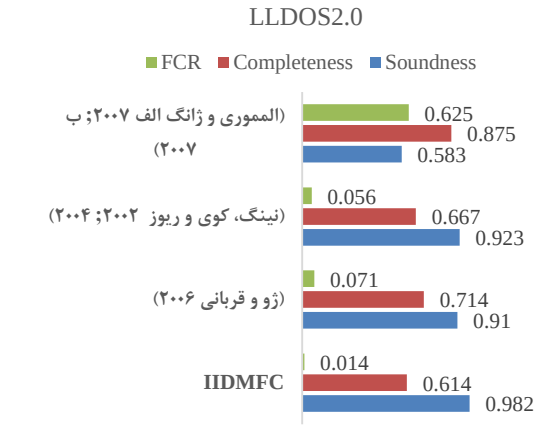
۴-۳- ارزیابی و مقایسه

آزمایش۱: در این آزمایش به بررسی و مقایسه میزان دقت همبسته‌سازی در سیستم همبسته‌ساز پیشنهادی با تعدادی از کارهای شناخته شده با استفاده از سه معیار درستی، تمامیت و نرخ خطا پرداخته شده است، همانطور که در شکل (۱۰ و ۱۱) نمایش داده است نتایج حاصل توسط همبسته‌ساز فازی پیشنهادی نسبت به روش‌های ایستا که در [۱۶،۱۷،۱۸،۱۹ و ۲۰] ارائه شده است و از قبل و به کمک ده‌ها قانون نحوه حمله LLDOS1.0 را می‌دانند نتایج مناسبی را ارائه نموده است.



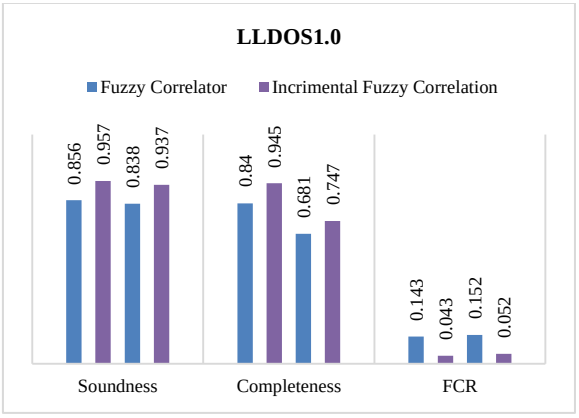
شکل (۱۰): مقایسه دقت همبستگی فازی افزایشی (IFC) با چندین تحقیق مشابه با سناریو LLDoS1.0

چرا که روش پیشنهادی در اینکارها مبتنی بر قوانین پیشنیاز و پی‌آمد بوده است. لذا لازم بوده است تا قبل از همبسته‌سازی، ارتباط بین مراحل هر حمله تعریف و در قالب ده‌ها قانون در پایگاه داده سیستم ذخیره شود، به نظر می‌رسد نتایج قابل قبولی را دارد. نتیجه تولیده شده توسط همبسته‌ساز پیشنهادی با نتایج گزارش شده در [۱۷-۱۸] و که از شناخته شده ترین و پر استنادترین کارهای انجام شده روی این داده‌هاست، بسیار نزدیک است.

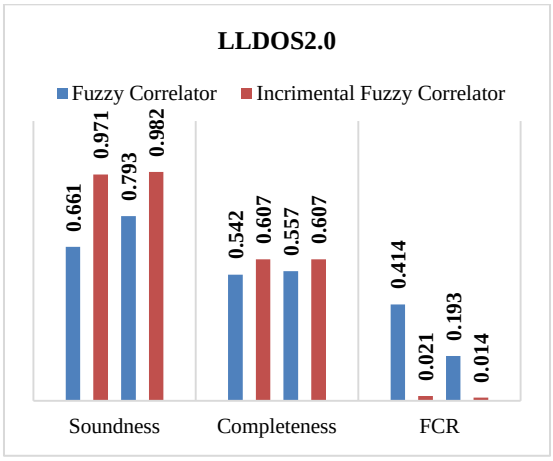


شکل (۱۱): مقایسه دقت همبستگی فازی افزایشی (IFC) با چندین تحقیق مشابه با سناریو LLDoS2.0.

همچنین نتایج به دست آمده توسط همبسته‌ساز فازی پیشنهادی به نتایج گزارش شده در [۱۹-۲۰] و نیز بسیار نزدیک است ضمن آنکه هرچند نتایج گزارش شده در این دو کار برای حمله LLDOS1.0 قوی به نظر می‌رسد اما در استخراج حمله ی LLDOS2.0 بسیار ضعیف عمل می‌کند. علاوه براین، این دو کار نیز مانند [۱۹-۲۰] و مبتنی بر قوانین ایستا و گردآوری پیشاپیش دانش مورد نیاز بوده است و توانایی تشخیص حملات جدید را ندارد. در مقایسه با روش‌های پویا نیز همبسته‌ساز فازی پیشنهادی از دقت بالایی برخوردار است. برای مثال در [۱۶] که مانند همبسته‌ساز فازی مبتنی بر یادگیری فعال پیشنهادی، به صورت پویا عمل می‌کند معیار تمایت به شکل معنی داری کمتر از همبسته‌ساز فازی مبتنی بر یادگیری فعال است. در شکل (۱۲ و ۱۳) سه معیار اصلی مورد استفاده در ارزیابی دقت که عبارتند از درستی، تمامیت و نرخ خطا در همبسته‌سازی مورد بررسی قرار گرفت.



شکل(۱۲): ارزیابی دقیق سناریوی LLDoS1.0



شکل(۱۳): ارزیابی دقیق سناریوی LLDoS2.0

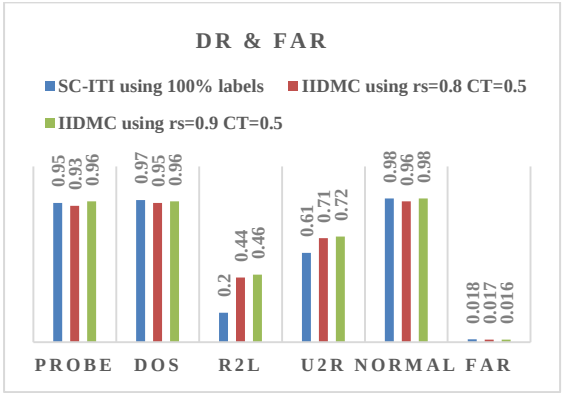
آزمایش ۲: در این آزمایش به ارزیابی سیستم تولید کننده هشدار به‌صورت مجزا پرداخته شده است، از آنجا که یادگیری افزایشی برخط در سیستم تولید کننده هشدار بدون در نظر گرفتن دو لایه بعدی یعنی لایه همبسته‌ساز و لایه یادگیری فعال

موضوعیت کمی پیدا می‌کند لذا به ارزیابی و مقایسه سیستم تولیدکننده هشدار به کمک یادگیری فعال با دو تا از بهترین روشها افزایشی در این حوزه پرداخته شده است.

مقایسه مدل پیشنهادی با روش SC-ITI: همانطور که در کارهای مرتبط عنوان شد، روش SC-ITI از معدود روش‌های افزایشی تشخیص نفوذ است که حداقل ۱۰ بار مورد ارجاع قرار گرفته است [۳۱] و در آزمایشات آن، از مجموعه آموزشی ساراساما[1] استفاده شده است [۳۲] این مجموعه آموزش شامل ۱۶۹/۰۰۰ نمونه از مجموعه داده "KDD ۱۰٪" است و مجموعه آزمون شامل ۳۱۱/۰۲۹ نمونه از مجموعه داده "KDD اصلاح شده[2]" است. آنها برای ارزیابی IDS پیشنهادی خود از دو معیار، DR و FAR استفاده کرده‌اند.

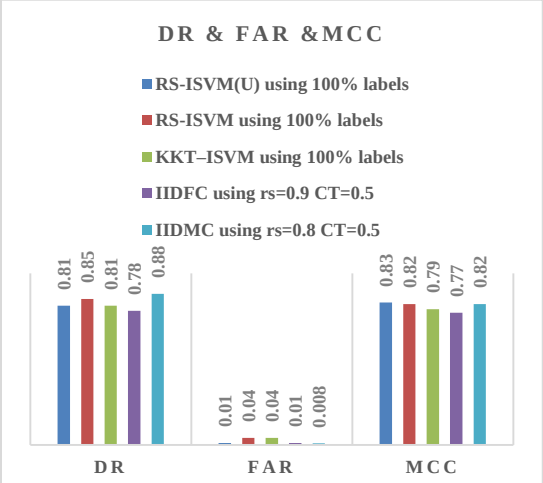
همانطور که در شکل (۱۴) نشان داده شده است، DRهای حملات R2L و U2R درIIDMFC با استفاده از ۸۰٪ و ۹۰٪ نرخ انتخاب قانون (rs) نسبت به SC-ITI با استفاده از داده‌ها با ۱۰۰٪ برچسب بطور قابل ملاحظه ای بیشتر است (برای سایر حملات، DRها تقریباً یکسان هستند).

نکته قابل توجه، میانگین FARهای روش SC-ITI و ISF-NIDS با استفاده از داده‌های دارای ۱۵٪ و ۱۰۰٪ برچسب به ترتیب ۰/۰۱۸، ۰/۰۱۷ و ۰/۰۱۶ است.



شکل (۱۴): مقایسه مقادیر DR، FAR با رویکردهای SC-ITI و مدل پیشنهادی افزایشی تشخیص نفوذ مبتنی بر همبسته‌ساز فازی (IIDMFC)

مقایسه مدل پیشنهادی با روش SVM افزایشی (RS-ISVM): روش ماشین بردار پشتیبان افزایشی نیز از روش‌های معدود افزایشی تشخیص نفوذ که بر روی مجموعه داده KDD99 آزمایش شده است، این مقاله در ژورنال غیر امنیتی چاپ شده اما در ژورنال‌های معتبر امنیتی حداقل ۳۵ بار مورد ارجاع قرار گرفته است.



شکل(۱۵): مقایسه مقادیر DR، FAR و MCC با رویکردهای مختلف SVM افزایشی و مدل پیشنهادی افزایشی تشخیص نفوذ مبتنی بر همبسته‌ساز (IIDMFC)

در [۱۲] ، نویسندگان حدود ۱۰٪ از مجموعه داده KDD99 را به عنوان مجموعه داده منبع انتخاب می‌کنند. همانطور که در شکل ۱۵ نشان داده شده است، بهترین DRها به IIDMFC با استفاده از rs = ۰/۸ تعلق دارند. DRهای IIDMFC با استفاده از rs = ۰/۹ تقریباً مشابه RS-ISVM است که بالاترین DR را در بین روش‌های آزمایش شده در (یی، وو و زو ۲۰۱۱) را دارد. همچنین، شکل (۱۵) FARهای روش‌های آزمایش شده را نشان می‌دهد. FARهای IDFC با استفاده از rs = ۰/۸ همانند حداقل FAR سایر روشهای آزمایش شده در [۱۲] است. حداقل FAR متعلق به IDFC با rs = ۰/۸ است. همچنین DRهای IDFC با استفاده از rs = ۰/۸ تقریباً مشابه RS-ISVM(U) و RS-ISVM است که بالاترین DR را در بین روشهای آزمایش شده در [۱۲] را دارد. مقادیر دو پارامتر Accuracy و f1 برای IIDMFC نیز به ترتیب برابر ۰/۹۹۵۴ و ۰/۹۹۵۱ است که نشان دهنده دقت بالای نرخ تشخیص است.

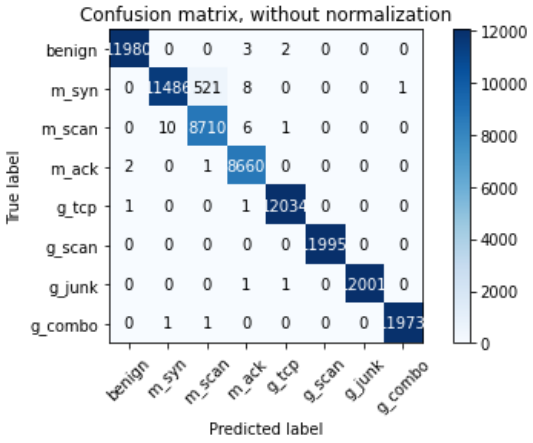
Accuracy for testing = 0.995387613773346

Average F1 score is 0.9951164636460431

شبیه‌سازی فاز آفلاین مدل پیشنهادی توسط مجموعه‌داده N-BaIoT: در این بخش به بررسی نتایج فاز آفلاین مدل پیشنهادی با استفاده از الگوریتم مبتنی بر یادگیری عمیق بر روی مجموعه داده بروز N-BaIoT پرداخته می شود. میزان دقت آن برابر ۹۹,۳۰٪ است که در بهترین حالت این مقدار به ۹۹,۵۱٪ هم می‌رسد.

همانطور که در ماتریس بهم‌ریختگی برای انواع حملات و مقادیر f1-score، PR(precision)، DR(recall) وaccuracy نشان داده شده در فاز آفلاین از دقت بالای ۹۹ درصد برخوردار است.

[1] Sarasamma
[2] Corrected KDD



شکل (۱۶): ماتریس بهم‌ریختگی حاصل از اجرای مرحله برون‌خط

جدول (۵): مقادیر معیار سنجش دقت بر روی مجموعه داده N-BaIoT

Classification Report

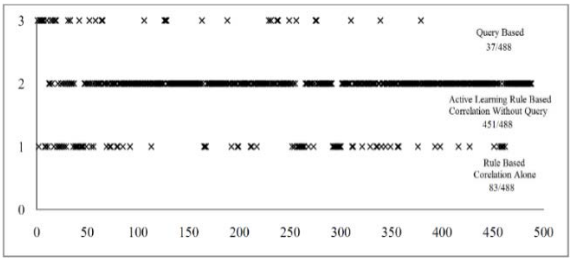
	precision	recall	f1-score	support
0	1.00	1.00	1.00	11985
1	1.00	0.96	0.98	12016
2	0.94	1.00	0.97	8727
3	1.00	1.00	1.00	8663
4	1.00	1.00	1.00	12036
5	1.00	1.00	1.00	11995
6	1.00	1.00	1.00	12003
7	1.00	1.00	1.00	11975
accuracy			0.99	89400
macro avg	0.99	0.99	0.99	89400
weighted avg	0.99	0.99	0.99	89400

۵- بحث و نتیجه گیری

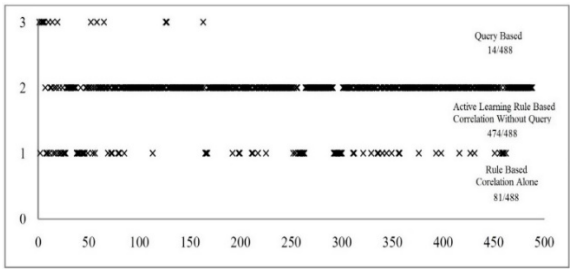
۵-۱- بحث

مقادیر دو پارامتر مهم آستانه همبستگی و حساسیت همبستگی نیز به‌طور شهودی به‌ترتیب برابر ۰/۵ و ۰/۱ انتخاب شده‌اند و بررسی‌های بعدی نیز این مقادیر را تایید کردند. وقتی دو هشدار در قالب یک بردار شباهت مورد بررسی قرار می‌گیرند و احتمال همبستگی آنها تعیین می‌شود برای آنکه واقعا همبسته شوند بایستی حداقلی از احتمال همبستگی را تامین نمایند.

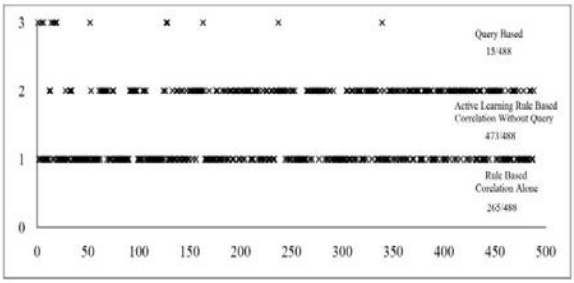
به نظر نمی‌رسد اگر احتمال همبستگی دو هشدار کمتر از ۰,۵ باشد صرف وقت بیشتر برای آن منطقی باشد بر همین اساس مقدار آستانه ی همبستگی برابر ۰/۵ انتخاب شده است. در کارهای مشابه نیز از همین عدد برای مقدار این پارمتر استفاده شده است. اما وقتی دو هشدار همبسته تشخیص داده شدند یعنی احتمال همبستگی آنها بیش از ۰/۵ است کلیه هشدارهایی که در همان فراهشدار قرار دارند و احتمال همبستگی آنها اختلافی کمتر از ۰/۱ دارد با این هشدار نیز همبسته فرض می‌شوند تا مقادیر موجود در ماتریس‌های همبستگی نیز بروز شود. این کار باعث خواهد شد تا سیستم از موفقیت فعلی خود تجربه ی بیشتری کسب کرده و آنرا برای همبستگی‌های بعدی بخاطر بسپارد. هرچقدر این پارامتر بزرگتر انتخاب شود دامنه ی تجربه کسب شده به تعداد بیشتری از هشدارها گسترش می‌یابد.



شکل (۱۷): تعداد هشدارهای همبسته توسط قوانین فازی به‌تنهایی، مبتنی بر قواعد با استفاده از یادگیری فعال بدون احتساب تعداد پرس و جوها و تعداد پرس و جوهای متخصص با rs =۰/۷ در LLDoS2.0



شکل (۱۸): تعداد هشدارهای همبسته توسط قوانین فازی به‌تنهایی، مبتنی بر قواعد با استفاده از یادگیری فعال بدون احتساب تعداد پرس و جوها و تعداد پرس و جوهای متخصص با rs =۰/۸ در LLDoS2.0



شکل (۱۹): تعداد هشدارهای همبسته توسط قوانین فازی به‌تنهایی، مبتنی بر قواعد با استفاده از یادگیری فعال بدون احتساب تعداد پرس و جوها و تعداد پرس و جوهای متخصص با rs =۰/۹ در LLDoS2.0

شکل (۱۹- ۱۷) درصد هشدارهایی را نشان می‌دهد که در هر یک از دو لایه یک لایه و دو لایه با مقادیر مختلف این دو پارامتر با یکدیگر همبستگی دارند. در rs = ۰/۸، تعداد هشدارهای همبسته در همبستگی مبتنی بر قانون تک لایه بیشتر از rs = ۰/۹ است، اما در حالت دو لایه یادگیری فعال، تعداد هشدارهای مورد نیاز فقط یک است، اما، همانطور که در شکل‌های (۱۹- ۱۷) نشان داده شده است، دقت به درستی

افزایش می‌یابد، بنابراین این مسئله نشان می‌دهد کـه دانـش بـه دست آمده به طور مناسب منجر به دقت در همبستگی‌های آینده می‌شود.

۵-۲- نتیجه‌گیری

در این مقاله یک مدل تشخیص نفوذ مبتنی بریادگیری افزایشی و بر پایه همبستگی به اختصار (IIDMFC) ارائه شده است، که قادر است به صورت افزایشی و به کمک دانش خبره فرایند یادگیری و آزمون را در یک زمان و به صورت برخط انجام دهد. مقاله حاضـر بـا اسـتفاده از دو مفهـوم یـادگیری فعـال و سیسـتم مبتنـی بـر همبستگی هشدارها سعی دارد با کمترین نیاز به دانـش خبـره و استفاده بهینه از آن یک مـدل کارامـد را پیـاده‌سـازی نمایـد بـه طوری‌که هم خطای انسانی کاهش یابد و هم سیستم نسـبت بـه بروزرسانی برخط خود با استفاده از دانش حاصل از همبسته‌سازی و متخصص امنیت مبادرت نمایـد. همچنـین بـا اسـتفاده از یـک ساختار حافظه‌ای کارامد و سریع، دانـش در یـک سـاختار نیمـه نظارتی ذخیره می‌شود و اینگونـه بیشـترین بهـره وری از دانـش کسب شده صورت گیرد و با بررسی حالات متفاوت و خرد جمعی در شرایط ابهام با ارائه پیشـنهاداتی، بـه خبـره در تصـمیم‌گیـری یاری می‌رساند. بدین طریق مفهوم سیستم‌های تصمیم‌یار نیـز در مدل پیشنهادی گنجانده شده است. روش ارائـه شـده روی چنـد مجموعه داده تست معتبر آزمایش شـده و نتـایج حاصـل بیـانگر کارآمدی مدل پیشنهادی با دقت بالای ۹۹ درصد و با نرخ مثبت کاذب بسیار پایین است. به‌عنوان کارهای آینده نیز پیشنهاد مـی شود تا اسـتفاده از روش‌هـای ویـرایش مفهـوم ماننـد روش‌هـای مبتنی بر DDM برای تقویت تکامـل و ویـرایش مفهـوم در مـدل پیشنهادی و به‌منظور ارتقا در بروزرسانی اطلاعات توسط سیستم یادگیری فعال مورد آزمایش قرار گیرد.

۵- مراجع

[1] H.-J. Liao, C.-H. R. Lin, Y.-C. Lin, and K.-Y. Tung, "Intrusion detection system: A comprehensive review," *Journal of Network and Computer Applications*, vol. 36, no. 1, Art. no. 1, 2013.

[2] B. Morin and H. Debar, "Correlation of Intrusion Symptoms: An Application of Chronicles," in *Recent Advances in Intrusion Detection*, Berlin, Heidelberg, 2003, pp. 94–112, doi: 10.1007/978-3-540-45248-5_6.

[3]A. A. Ghorbani, W. Lu, and M. Tavallaee, *Network Intrusion Detection and Prevention: Concepts and Techniques*. Springer Science & Business Media, 2010.

[4] C. V. Zhou, C. Leckie, and S. Karunasekera, "A survey of coordinated attacks and collaborative intrusion detection," *Computers & Security*, vol. 29, no. 1, pp. 124–140, Feb. 2010, doi: 10.1016/j.cose.2009.06.008.

[5] Y. Bai and H. Kobayashi, "Intrusion Detection Systems: technology and development," in *17th International Conference on Advanced Information Networking and Applications, 2003. AINA 2003.*, Mar. 2003, pp. 710–715, doi: 10.1109/AINA.2003.1192972.

[6] A. S. Sodiya, H. O. D. Longe, and A. T. Akinwale, "A new two- tiered strategy to intrusion detection," *Information Management & Computer Security*, vol. 12, no. 1, Art. no. 1, Jan. 2004, doi: 10.1108/09685220410518810.

[7] S. Duque and Mohd. N. bin Omar, "Using Data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS)," *Procedia Computer Science*, vol. 61, pp. 46–51, Jan. 2015, doi: 10.1016/j.procs.2015.09.145.

[8] N. K. Kanakarajan and K. Muniasamy, "Improving the Accuracy of Intrusion Detection Using GAR-Forest with Feature Selection," in *Proceedings of the 4th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA) 2015*, New Delhi, 2016, pp. 539–547, doi: 10.1007/978-81-322-2695-6_45.

[9] J. A. Khan and N. Jain, "A survey on intrusion detection systems and classification techniques," *Int. J. Sci. Res. Sci., Eng. Technol.*, vol. 2, no. 5, pp. 202–208, 2016.

[10] D. Gupta, S. Singhal, S. Malik, and A. Singh, "Network intrusion detection system using various data mining techniques," in *2016 International Conference on Research Advances in Integrated Navigation Systems (RAINS)*, May 2016, pp. 1–6, doi: 10.1109/RAINS.2016.7764418.

[11] W.-Y. Yu and H.-M. Lee, "An incremental-learning method for supervised anomaly detection by cascading service classifier and ITI decision tree methods," in *Pacific-Asia Workshop on Intelligence and Security Informatics*, 2009, pp. 155–160.

[12] Y. Yi, J. Wu, and W. Xu, "Incremental SVM based on reserved set for network intrusion detection," *Expert Systems with Applications*, vol. 38, no. 6, pp. 7698–7707, 2011.

[13] K. K. Gupta, B. Nath, and R. Kotagiri, "Layered Approach Using Conditional Random Fields for Intrusion Detection," *IEEE Transactions on Dependable and Secure Computing*, vol. 7, no. 1, pp. 35–49, Jan. 2010, doi: 10.1109/TDSC.2008.20.

[14] R. Sadoddin and A. A. Ghorbani, "An incremental frequent structure mining framework for real-time alert correlation," *Computers & Security*, vol. 28, no. 3, pp. 153–173, 2010, doi: 10.1016/j.cose.2008.11.010.

[15] G. P. Spathoulas and S. K. Katsikas, "Reducing false positives in intrusion detection systems," *Computers & Security*, vol. 29, no. 1, pp. 35–44, Feb. 2010, doi: 10.1016/j.cose.2009.07.008.

[16] B. Zhu and A. A. Ghorbani, "Alert correlation for extracting attack strategies," *IJ Network Security*, vol. 3, no. 3, pp. 244–258, 2006.

[17] P. Ning, Y. Cui, and D. S. Reeves, "Constructing attack scenarios through correlation of intrusion alerts," in *Proceedings of the 9th ACM conference on Computer and communications security*, Washington, DC, USA, Nov. 2002, pp. 245–254, doi: 10.1145/586110.586144.

[18] P. Ning, Y. Cui, D. S. Reeves, and D. Xu, "Techniques and tools for analyzing intrusion alerts," *ACM Trans. Inf. Syst. Secur.*, vol. 7, no. 2, pp. 274–318, May 2004, doi: 10.1145/996943.996947.

[19] S. O. Al-Mamory and H. L. Zhang, "Building Scenario Graph Using Clustering," in *2007 International Conference on Convergence Information Technology (ICCIT 2007)*, Nov. 2007, pp. 799–804, doi: 10.1109/ICCIT.2007.51.

[20] S. O. Al-Mamory and H. L. Zhang, "Scenario Discovery Using Abstracted Correlation Graph," in *2007 International Conference on Computational Intelligence and Security (CIS 2007)*, Dec. 2007, pp. 702–706, doi: 10.1109/CIS.2007.21

[21] A. Milenkoski, M. Vieira, S. Kounev, A. Avritzer, and B. D. Payne, "Evaluating Computer Intrusion Detection Systems: A Survey of Common Practices," *ACM Comput. Surv.*, vol. 48, no. 1, p. 12:1–12:41, Sep. 2015, doi: 10.1145/2808691.

[22] R. Kandhari, V. Chandola, A. Banerjee, V. Kumar, and R. Kandhari, "Anomaly detection," *ACM Comput. Surv*, vol. 41, no. 3, pp. 1–6, 2009.

[23] J. Arshad, P. Townend, and J. Xu, "A novel intrusion severity analysis approach for Clouds," *Future Generation Computer Systems*, vol. 29, no. 1, pp. 416–428, Jan. 2013, doi: 10.1016/j.future.2011.08.009.

[24] F. Shen and O. Hasegawa, "A fast nearest neighbor classifier based on self-organizing incremental neural network," *Neural Networks*, vol. 21, no. 10, pp. 1537–1547, Dec. 2008, doi: 10.1016/j.neunet.2008.07.001.

[25] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," 1999.

[26] A. Valdes and K. Skinner, "Probabilistic Alert Correlation," in *Recent Advances in Intrusion Detection*, Berlin, Heidelberg, 2001, pp. 54–68, doi: 10.1007/3-540-45474-8_4.

[27] K. Polat and S. Güneş, "Principles component analysis, fuzzy weighting pre-processing and artificial immune recognition system based diagnostic system for diagnosis of lung cancer," *Expert Systems with Applications*, vol. 34, no. 1, pp. 214–221, Jan. 2008, doi: 10.1016/j.eswa.2006.09.001.

[28] "DARPA 2000 Intrusion Detection Scenario Specific Datasets | MIT Lincoln Laboratory." https://www.ll.mit.edu/r-d/datasets/2000-darpa-intrusion-detection-scenario-specific-datasets (accessed Aug. 07, 2020).

[29] "KDD Cup 1999 Data." http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html (accessed Aug. 07, 2020).

[30] "Learning to Forget: Continual Prediction with LSTM | Neural Computation | MIT Press Journals." https://www.mitpressjournals.org/doi/abs/10.1162/089976600300015015 (accessed Aug. 28, 2020).

[31] W.-Y. Yu and H.-M. Lee, "An incremental-learning method for supervised anomaly detection by cascading service classifier and ITI decision tree methods," in *Pacific-Asia Workshop on Intelligence and Security Informatics*, 2009, pp. 155–160.

[32] S. T. Sarasamma and Q. A. Zhu, "Min-max hyperellipsoidal clustering for anomaly detection in network security," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 4, pp. 887–901, Aug. 2006, doi: 10.1109/TSMCB.2006.870629.