مجله علمی- پژوهشی "پدافند الکترونیکی و سایبری"
سال ششم، شماره ۲، تابستان ۱۳۹۷، ص ۱۰۸-۹۳

شناسایی جریان‌های مخرب در شبکه با به‌کارگیری اجماع

حمید پروین[۱]، وحیده رضایی[۲]، صمد نجاتیان[۳]*، روح‌اله امیدوار[٤]، میلاد یثربی[٥]

۱- استادیار دانشکده مهندسی برق و کامپیوتر دانشگاه آزاد اسلامی نورآباد ممسنی، ۲ و ۳- استادیار، دانشگاه آزاد اسلامی واحد یاسوج ٤- مربی، باشگاه پژوهشگران جوان ونخبگان واحد یاسوج، ۵- کارشناس ارشد، گروه کامپیوتر، دانشگاه آزاد اسلامی، واحد شیراز، ایران

(دریافت: ۹٦/۰٤/۲٤، پذیرش: ۹٦/۰۷/۰۱)

چکیده

مقوله امنیت در شرایط جدید جهانی ابعاد متفاوتی پیدا کرده است. یکی از حوزه‌های امنیتی که در شرایط جدید جهانی بسیار مورد اهمیت قرار گرفته است، حوزه امنیت سایبری است. در این تحقیق برای مطالعه بر روی حملات ناشناخته دو هانی‌نت آزمایشگاهی مجازی در دو مکان مختلف طراحی شده و همچنین از سایر مجموعه داده‌های علمی استفاده گردیده است. در داده‌های شبکه‌ای، مشکل داده‌های نامتوازن اغلب اتفاق می‌افتد و موجب کاهش کارایی در پیش‌بینی برای رده‌هایی که در اقلیت هستند، می‌گردد. در این مقاله برای حل این مشکل، از روش‌های یادگیری جمعی استفاده گردیده است تا بتوان مدلی خودکار ارائه نمود که با استفاده از فنون مختلف و با استفاده از یادگیری مدل، حملات شبکه به‌ویژه حملات ناشناخته را شناسایی نماید. روش‌های جمعی، بسیار مناسب برای توصیف مشکلات امنیتی رایانه‌ای می‌باشند زیرا هر فعالیتی که در سیستم‌های رایانه‌ای انجام می‌گیرد را می‌توان در سطوح چند انتزاعی مشاهده کرد و اطلاعات مرتبط را می‌توان از منابع اطلاعاتی چندگانه جمع‌آوری نمود. روش تحقیق بر اساس تحلیل‌های آماری جهت بررسی میزان صحت و درستی نتایج و میزان اتکاپذیری آن‌ها صورت گرفته است. در این مرحله به‌کمک فنون و آزمایش‌های آماری نشان داده شده که عملکرد الگوریتم طراحی شده با رأی‌گیری وزنی پیشنهادی بر اساس الگوریتم ژنتیک نسبت به دوازده طبقه‌بند دیگر بهتر می‌باشد.

واژه‌های کلیدی: هانی‌نت، حملات ناشناخته، یادگیری جمعی، داده‌های نامتوازن، رای‌گیری وزنی، آزمایش‌های آماری

۱- مقدمه

در حال حاضر مطالعات و بررسی‌های تجربی حملات اینترنتی به‌عنوان یکی از دامنه‌های پژوهشی فعال به شمار می‌آید، و در سال‌های اخیر توجه زیادی به خود جلب کرده است.

در دنیای امروز مجموعه داده‌های نامتوازن سهم زیادی از مجموعه داده‌های جهان واقعی را به‌خود اختصاص داده‌اند. در مجموعه مسائل داده‌های نامتوازن رده‌های اقلیت[1] اهمیت بیشتری نسبت به رده‌های اکثریت[2] دارند. از این رو رده‌بندی درست مجموعه داده‌های نامتوازن از اهمیت بیشتری برخوردار است. هنگام یادگیری و طبقه‌بندی مجموعه داده‌های نامتوازن، روش‌های استاندارد داده‌کاوی سعی در رسیدن به حداکثر دقت طبقه‌بندی دارند و همین امر باعث صرف‌نظر کردن از رده‌های اقلیت می‌شود که همین امر باعث می‌شود این روش‌ها برای مقابله با مشکل رده‌های نامتوازن روش‌های نامناسبی باشند.

یکی از بهترین ابزارهای جمع‌آوری اطلاعات جهت بررسی و تحقیق درباره حملات ناشناخته هانی‌نت‌ها می‌باشند. یک هانی‌نت شامل شبکه‌ای از هانی‌پات‌هایی با تعامل بالا می‌باشند و یک شبکه واقعی را شبیه‌سازی می‌کند [۱]. در داده‌های شبکه‌ای، مشکل داده‌های نامتوازن اغلب اتفاق می‌افتد و موجب کاهش کارایی در پیش‌بینی برای رده‌هایی که در اقلیت هستند، می‌گردد [۳-۲]. برای حل مشکل به‌وجودآمده در داده‌های نامتوازن از روش‌های یادگیری جمعی استفاده گردیده است. مشخص شده که روش‌های جمعی، بسیار مناسب برای توصیف مشکلات امنیتی رایانه می‌باشند زیرا هر فعالیتی که در سامانه‌های رایانه‌ای انجام می‌گیرد را می‌توان در سطوح چند انتزاعی مشاهده کرد و اطلاعات مرتبط را می‌توان از منابع اطلاعاتی چندگانه جمع‌آوری

*رایانامه نویسنده مسئول: nejatian@iauyasooj.ac.ir

1- Minority Class
2- Majority Class

نمود [۴]. تحقیقات مختلفی درباره کاربرد روش‌های جمعی جهت تشخیص نفوذ صورت گرفته است [۸-۵].

هدف از این مطالعه، یافتن و ارائه مدلی برای خودکار نمودن سیستم‌های تشخیص نفوذ می‌باشد بدین شکل که براساس یادگیری، جریان‌های ناشناخته مخرب را بتوان شناسایی نمود. برای حل مشکل داده‌های نامتوازن در داده‌های شبکه، از یادگیری جمعی استفاده شد. یک طبقه‌بندی مبتنی بر یادگیری جمعی، شامل مجموعه‌ای از طبقه‌بندها است که بر روی داده‌های آموزشی، آموزش داده شده‌اند. در این مقاله سعی بر آن است تا با ارائه راه‌حلی کارآمد مسائل و مشکلات مطرح شده را رفع و دقت تشخیص را در حملات مختلف افزایش دهیم. به‌ویژه تاکید ما بر تشخیص سریع، دقیق و خودکار حملات می‌باشد. انتظار این است که سیستم پیشنهادی نه‌تنها این حملات را با دقت مناسب تشخیص دهد بلکه نرخ اشتباه در آن‌ها را نیز به‌حداقل برساند. برای اجرای این کار با استفاده از یادگیری جمعی هم مشکل داده‌های نامتوازن شبکه را حل نموده و هم مدلی کاراتر و دقیق‌تری برای شناسایی حملات به‌دست آمد. به طور خاص هدف نویسندگان این مقاله برآورده کردن اهداف زیر می‌باشد:

۱- شناسایی جریان‌های ناشناخته و مخرب در شبکه.

۲- طراحی مدلی برای خودکار نمودن سیستم‌های تشخیص نفوذ بر اساس یادگیری.

۳- پیاده‌سازی شبکه هانی‌نت آزمایشگاهی بهینه‌سازی‌شده بر بستر مجازی سازی.

۴- ارتقا فنون شناسایی در سیستم‌های تشخیص نفوذ.

۵- افزایش امنیت شبکه‌ها از طریق امکان شناسایی حملات ناشناخته.

در این تحقیق ابتدا به توضیح نتایج جدیدترین کارهای مرتبط پرداخته‌ایم شده است. در بخش بعدی هانی‌نت آزمایشگاهی طراحی شده تشریح گردیده است و پس از آن الگوریتم پیشنهادی به همراه روش وزن‌دهی پیشنهادی جهت انتخاب مدل نهایی تشریح گردیده است. در ادامه روش‌های ارزیابی‌شده و تشریح نتایج به‌دست‌آمده مورد بررسی قرار گرفته و در پایان نتایجی که از تحقیق به‌دست آمده جهت تحقیق‌های آینده تشریح گردیده است.

۲- کارهای مرتبط

روش‌های جمعی به‌طور موفقیت آمیز برای ایجاد تغییر و تمایز در امور جهان واقعی به‌کار گرفته شدند. به راستی این روش‌ها تقریباً در هر جایی که راه‌کارهای یادگیری به کار می‌رود مفید می‌باشند.

تحقیقی در زمینه شناسایی ویروس‌ها با استفاده از اجماعی از شبکه عصبی و SVM صورت گرفته است. در این تحقیق روشی جدید با استفاده از یکپارچه‌سازی شناسایی پویا و ایستا ویروس‌ها ارائه گردیده است. نتایج حاصل نشان می‌دهد که روش جمعی پیشنهادی بسیار کارا به شناسایی ویروس‌های ناشناخته و تغییریافته می‌پردازد [۹].

مطالعه دیگری در زمینه به‌کارگیری روش جمعی برای تصحیح خطای خروجی جهت تشخیص نفوذ انجام شده است. هدف این تحقیق رفع مشکل داده‌های نامتوازن، افزایش نرخ تشخیص هر رده و کمینه کردن هشدارهای کاذب در سیستم تشخیص نفوذ می باشد. با استفاده از روش‌های جمعی بگینگ و ادابوست مورد آزمایش قرار گرفته است. روش جمعی پیشنهادی باعث بهبود میزان دقت تا ۹۹/۷٪ می‌شود [۱۰].

تحقیقی با عنوان سیستم‌های تشخیص نفوذ با استفاده از انتخاب جمعی بگینگ صورت گرفته که با استفاده از روش‌های جمعی موفق به طراحی الگوریتمی با کارایی بسیار بالا در پیش‌بینی مسائل شده و اشکالات کارهای گذشته را رفع نموده است [۱۱].

روش جدید دیگری از روش‌های تشخیص نفوذ با استفاده از روش‌های جمعی در داده‌کاوی ارائه شده است. در این مطالعه با استفاده از روش اجماع bagging و رده پایه REPTree سیستم تشخیص نفوذ پیاده سازی گردیده است. در این تحقیق دلیل استفاده از روش bagging زمان کم‌تر برای ایجاد مدل می‌باشد. روش پیشنهادی به‌صورت کاملا رقابتی نرخ مثبت کاذب کم‌تری نسبت به سایر روش‌های یادگیری داشته و بالاترین دقت طبقه‌بندی را نشان می‌دهد [۱۲].

مدل دیگری از تشخیص نفوذ مبتنی بر یادگیری جمعی برای حملات U2R و R2L ارائه شده است. هدف شناسایی داده‌های مربوط به حملات شبکه است که شناسایی آن‌ها دشوار می‌باشد. نتایج حاصل نشان داده است که استفاده از الگوریتم ادابوست در یادگیری جمعی موجب افزایش کارایی تشخیص حملات طبقه بندها از میان طبقه‌بندهای ضعیف می‌شود [۱۳].

مطالعه دیگری نیز اشتباه طبقه‌بندی در سیستم‌های تشخیص نفوذ را مورد بررسی قرار داده است. در این تحقیق یک ساختار جدید مبتنی بر اجماع فازی ارائه گردیده است. این ساختار چند مرحله‌ای ابتدا انتخاب ویژگی را مورد نظر دارد و سپس در مرحله ترکیب طبقه‌بندها روش اجماع فازی را اعمال

می‌نماید. روش وزن‌دهی فازی جمعی پیشنهادی عملکرد بسیار مناسب‌تری نسبت به سایر روش‌ها داشته است [۱۴].

مطالعات دیگری نیز در این تحقیق مورد مطالعه گرفته است [۱۵-۱۸].

در مرجع [۳۰] یک روش استفاده از بینایی استریو به‌منظور ارائه یک سامانه واقعیت افزوده مبتنی بر بینایی ارائه شده است. در این مقاله یک سامانه واقعیت افزوده ارائه شده است که مبتنی بر بینایی است و از بینایی استریو برای به‌دست آوردن اطلاعات عمق استفاده می‌کند. در این سامانه از نشانک‌های غیرفعال دایره‌ای با شعاع ۱/۵ cm استفاده شده است که به منبع انرژی اضافی خارجی برای راه‌اندازی آن‌ها نیازی نیست. در مرجع [۳۱] یک روش استفاده از خوشه‌بندی و مدل مارکوف جهت پیش‌بینی درخواست آتی کاربر در وب ارائه شده است. این الگوریتم از یک روش پیش‌بینی استفاده می‌کند تا فعالیت‌های آتی کاربر را پیش‌بینی کند.

در[۳۲]، تحقیقی با عنوان رویکرد یادگیری مبتنی بر نمونه برای شناسایی و دسته‌بندی جریان‌های مخرب شبکه در هانی‌نت‌ها انجام شده و به این نتیجه رسیدند که با استفاده از الگوریتم‌های یادگیری ماشین به خصوص نمونه‌های مبتنی بر یادگیری به‌صورت قطعی و خودکار می‌توان یک مدل متشکل از نمونه‌ها طراحی نمود. همچنین نشان داده‌اند که شناسایی بر اساس نمونه‌های جدید و دسته‌بندی آن‌ها افزایش دقت و صحت عمل شناسایی را در پی داشته است.

تحقیقی با عنوان پیشنهاد یک روش جدید برای داده‌های نامتوازن غیرمرتبط در [۳۳] انجام شده است. نویسندگان در این مقاله به ارایه الگوریتمی پرداخته‌اند که مناسب و قابل اعمال برای مجموعه‌ای از داده‌های نامتوازن غیرمرتبط می‌باشد. آن‌ها نشان داده‌اند که این الگوریتم برای سرعت و کارایی یادگیری، کارآمد می باشد. نتایج نشان می‌دهد که کارایی الگوریتم پیشنهادی نسبت به روش‌هایی که مورد بررسی قرار دادند، بهتر می‌باشد.

در[۳۴]، روش داده وزن‌دار برای خوشه‌بندی جمعی انجام شده است. در این مطالعه از تکنیک جدیدی برای خوشه‌بندی جمعی با استفاده از نمونه‌گیری بوستینگ داده اصلی استفاده نموده‌اند. همچنین نشان داده‌اند که بخش مهمی از نقاط داده‌ها وجود دارند که از آن‌ها می‌توان ساختار کل مجموعه داده را شناسایی نمود.

تجزیه و تحلیل مقایسه‌ای طرح‌های رای‌گیری شناسایی بدافزارها مبتنی بر یادگیری در [۳۵] انجام شده است. آن‌ها یک روش شناسایی مبتنی بر یادگیری نظارتی ارایه کرده و با استفاده از دو الگوریتم یادگیری جمعی و دو تکنیک پیش‌پردازش با الگوریتم‌های مطرح به ارائه روشی برای شناسایی بدافزارها پرداخته‌اند. مدل آن‌ها برای شناسایی بدافزارهای مختلف و برای کلاس‌های پیش‌بینی مختلف مورد آزمایش قرار گرفته است.

نویسندگان مرجع [۳۶] روش‌های یادگیری ماشین را برای تشخیص نفوذ، با توجه به پارامترهایی مانند پیچیدگی الگوریتم، چالش‌های بهبود امنیت و غیره مورد بررسی قرار دادند. نویسندگان معیارهای مختلفی نظیر دقت، پیچیدگی الگوریتم و پیچیدگی زمان را برای انتخاب روش موثر تشخیص نفوذ مورد بررسی قرار داده‌اند.

نویسندگان مرجع [۳۷] روش جدیدی طبقه‌بندی را برای سیستم تشخیص نفوذ ارائه دادند تا دقت را افزایش دهند. در این روش برای وزن‌دهی ویژگی‌ها از الگوریتم ازدحام ذرات استفاده شده است. نتایج این روش با الگوریتم WMA مقایسه شده و نشان داده شده که این روش نتایج مناسب‌تری را تولید نموده است.

در تحقیق که در مجله پدافند الکترونیکی و سایبری منتشر شده به وسیله نویسندگان [۳۸] راه‌کاری نویر مبتنی بر روش آشکارسازی ترکیبی با یک معماری چهارلایه‌ای پیشنهاد شده است. لایه اول از واحد تحلیل‌گر جریان داده‌ها و واحد طبقه‌بندی تشکیل شده است که برای طبقه‌بندی نوع سرویس‌های شبکه از ترکیتب روش آمتاری n گرام و الگوریتم ژنتیک استفاده می‌کند. در لایه تشخیص نفوذ، یک واحد آشکارساز مبتنی بر امضاء و واحدهای آشکارساز مبتنی بر - ناهنجاری به شکل ترکیبی پیاده‌سازی شده‌اند که متناسب با برچسب نوع سرویس‌ها فراخوانی می‌شوند. سپس، در نتیجه پردازش ایر واحدها، لایه تصمیم‌گیری فراخوانی می‌شود. ایر لایه ماهیت حمله و نوع پاسخ را تشخیص داده و لایه مدیریت وقایع را فرامتنی خوانند. در این لایه ضمن اطلاع‌رسانی هشدارها به مدیر شبکه، در صورت نیاز، اعمال واکنشی و اقدامات امنیتی لازم نیز انجام خواهد شد.

۳- هانی‌نت‌ها و هانی‌پات‌ها

پروژه هانی‌نت [۱۹]، ابزارهای نرم‌افزاری و سخت‌افزاری انبوه را پیشنهاد می‌دهند که اعضای اتحادیه، می‌توانند این ابزار را به کار گیرند و همچنین پروژه‌های هانی‌نت جهت خلق شبکه‌های رایانه-ای (واقعی و مجازی) مورد استفاده قرار می‌گیرند و این شبکه‌ها به نحوی گسترش می‌یابند که بتوانند به جمع‌آوری اطلاعات درباره انواع ترافیک‌های رسیده به هانی‌پات بپردازند. در این تحقیق، هانی‌نت محدود به معنای محدود آن نمی‌شود. تعریف پیشنهادی نویسندگان این مقاله این است که هانی‌نت یک شبکه از هانی‌پات‌ها می‌باشد که همبندی شبکه خاصی را دنبال می‌کند.

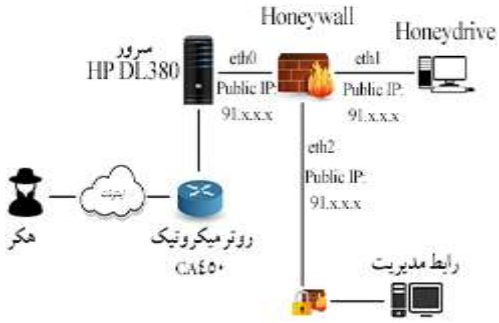
هانی‌نت یک شبکه بسیار کنترل‌شده از هانی‌پات‌ها می‌باشد [۲۰] که سرویس‌های شبکه‌ای را شبیه‌سازی می‌نماید تا این سرویس‌ها جهت تقلید یا شبیه‌سازی برنامه‌های واقعی و یا حتی کل سیستم طراحی شوند و ممکن است برای یک مهاجم بالقوه جذاب به‌نظر برسند، زمانی که مهاجم توسط این سرویس‌ها فریفته گردید، ممکن است هانی‌پات با مزاحم بالقوه تعامل برقرار کند.

ماشین فیزیکی که مثل یک هانی‌پات کار کند هانی‌پات فیزیکی نامیده می‌شود، در حالیکه یک ماشین مجازی که به‌عنوان هانی‌پات کار می‌کند، هانی‌پات مجازی نامیده می‌شود [۲۱].

۳-۱- طراحی هانی‌نت مجازی آزمایشگاهی

برای جمع آوری اطلاعات حملات و ایجاد یک مجموعه داده (به‌عنوان یکی از مجموعه داده‌ها) برای آزمایش الگوریتم پیشنهادی از دو هانی‌نت آزمایشگاهی مجازی پیاده‌سازی‌شده در دو مکان متفاوت استفاده شده است.

یکی از مهم‌ترین تفاوت‌های هانی‌نت طراحی‌شده این مقاله با سایر مدل‌ها استفاده از Honeydrive (که خود شامل تعداد زیادی هانی‌پات می‌باشد) به جای استفاده از هانی‌پات‌های جداگانه می‌باشد. Honeydrive یکی از بهترین هانی‌پات‌های توزیع لینوکسی می‌باشد. نسخه‌ای که در این مقاله استفاده شده لینوکس دسکتاپ Xubuntu ورژن ۱۲,۰۴,۴ می‌باشد. این هانی‌پات شامل بیش از ۱۰ هانی‌پات و ابزار تجزیه و تحلیل از قبل نصب‌شده و تنظیم‌شده می‌باشد. شکل (۱) هانی‌نت طراحی‌شده که الگو گرفته از ساختار هانی‌نت نسل سوم می‌باشد را نشان می‌دهد. همانطور که در این شکل قابل ملاحظه می‌باشد مهاجم از طریق اینترنت به سروری که هانی‌نت مجازی بر روی آن پیاده‌سازی شده است متصل می‌گردد. طبق یک قانون کلی، هرگونه اتصال به هانی‌نت مشکوک در نظر گرفته می‌شود. واسط eth0 وظیفه اتصال مهاجم به هانی‌وال را دارد. واسط eth1 وظیفه اتصال Honeywall به شبکه محلی داخلی که شامل هانی‌پات‌ها می‌باشد را بر عهده دارد. همچنین واسط eth2 این امکان را ایجاد می‌کند که بر یک بستر امن و کنترل‌شده بتوان از طریق SSh و یا Walleye تنظیمات و گزارش‌های مدیریتی را انجام داد. هانی‌نت آزمایشگاهی مجازی ایجادشده برای دوره‌ای از تاریخ ۵ شهریور ۱۳۹۴ (۲۷ آگوست ۲۰۱۵) تا ۵ آذر ۱۳۹۴ (۲۶ نوامبر ۲۰۱۵) به مدت ۹۱ روز در دو مکان پتروشیمی و مرکز تحقیقاتی آنلاین بود. در این مدت در مجموع تقریباً حدود ۲۱۶۰۲ اتصال به هانی‌نت صورت گرفت که از این تعداد حدود ۲۰۲۳۷ غیر حمله و حدود ۲۳۶۴ حمله تشخیص داده شده است. این اتصالات بر اساس سه پروتکل TCP، UDP و ICMP می‌باشد. اولین حمله ۹ روز بعد از راه اندازی هانی‌نت‌ها شناسایی گردیده است.

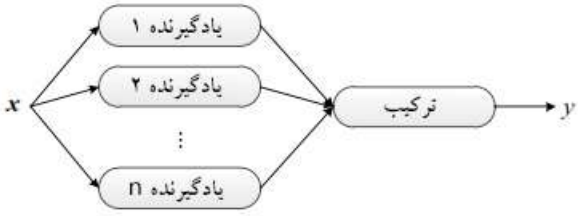


شکل (۱): هانی‌نت آزمایشگاهی مجازی پیشنهادی.

۴- یادگیری جمعی

یادگیری جمعی فرآیندی است که در آن مدل‌های متعدد از قبیل طبقه‌بندها یا خبره‌ها به‌صورت راهبردی تولید و یا ترکیب می‌شوند تا به یک مشکل خاص هوش محاسباتی پاسخ دهند. در یادگیری ماشین، روش‌های یادگیری از چندین الگوریتم یادگیری استفاده می‌کنند تا عمکرد بهتری در پیش‌بینی نسبت به هر کدام از الگوریتم‌های یادگیری تشکیل‌دهنده داشته باشند [۲۴-۲۲].

روش‌های جمعی از دهه ۱۹۹۰ تبدیل به الگوی اصلی یادگیری شدند که پیشرفت زیادی با دو اثر پیشرو داشت. یکی از آن‌ها تجربی [۲۵] است که در آن، این نتیجه به‌دست آمد که پیش‌بینی‌های انجام‌شده از طریق ترکیب مجموعه‌ای از طبقه‌بندها اغلب دقیق‌تر از پیش‌بینی‌های انجام‌شده توسط بهترین طبقه‌بند منفرد می‌باشد. اثر دیگر نظری است [۲۶] و در آن اثبات شد که یادگیرندگان ضعیف را می‌توان به یادگیرندگان قوی توسعه داد. شکل (۲) نشان‌دهنده یک معماری روش اول روش جمعی می‌باشد.



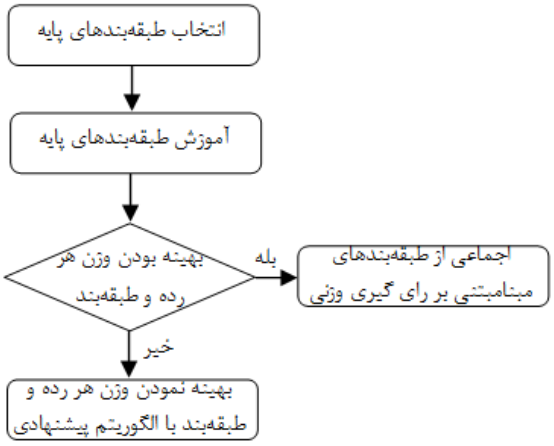
شکل (۲): معماری متداول روش جمعی [۲۷].

فرآیند یادگیری جمعی دارای سه مرحله است: ۱) تولید جمعی، ۲) انتخاب جمعی و ۳) ادغام جمعی. تولید جمعی در صورتی همگن است که یک الگوریتم قیاسی یکسان برای تولید تمامی طبقه‌بندهای جمعی به کار رود. در غیر این‌صورت گفته می‌شود که تولید جمعی ناهمگن می‌باشد. در مرحله تولید جمعی، مجموعه‌ای از طبقه‌بندهای پایه مختلف ایجاد می‌شود. این عمل را می‌توان از طریق (۱) پارامترهای آغازگر مختلف طبقه‌بندهای پایه و یا (۲) زیر مجموعه‌های مختلف فضای ویژگی

(سطح ویژگی) یا (۳) زیر مجموعه‌های مختلف داده‌ها (سطح داده‌ای) برای آموزش طبقه‌بندهای پایه انجام داد. انتخاب جمعی نیازمند انتخاب طبقه‌بندها از میان مجموعه طبقه‌بندهای مختلف می‌باشد. ادغام جمعی شامل ترکیب پیش‌بینی‌های مجموعه‌ای از طبقه‌بندهای پایه است [۲۳].

۵- روش پیشنهادی

روش یادگیری جمعی پیشنهادی، استفاده از رای‌گیری وزنی براساس الگوریتم ژنتیک می‌باشد که به اختصار GAWVote نامیده می‌شود. در روش پیشنهادی، طبقه‌بندهای مبنا به‌طور تصادفی به‌وسیله الگوریتم آدابوست تولید می‌شوند، سپس وزن‌های مناسب تمام طبقه‌بندهای مبنا، بسته به اطمینان پیش‌بینی از طریق الگوریتم ژنتیک مشخص می‌شوند. یکی از اساسی‌ترین نوآوری‌های روش پیشنهادی این مقاله این است که، هم رده و هم رده‌بند از نظر وزن‌دهی توسط الگوریتم ژنتیک بهینه خواهند شد. کل روش کار در شکل (۳) به‌طور خلاصه بیان شده است.



شکل (۳): روندنمای فرآیند کار برای روش پیشنهادی.

۵-۱- انتخاب و آموزش طبقه‌بندهای مبنا

استفاده از مجموعه طبقه‌بندها، پذیرش گسترده‌ای را در یادگیری ماشینی و جامعه آماری به‌علت بهبود قابل توجه در صحت به‌دست آورده است. طبقه‌بندهای فردی باید تا حد ممکن متنوع باشند. در تکنیک‌های جمعی معروف، مانند بگینگ و بوستینگ، این تنوع‌ها با دست‌کاری نمونه‌های آزمایش به‌منظور ایجاد فرضیه‌های چندگانه به‌دست می‌آیند. در روش پیشنهادی چند طبقه‌بند مبنا با استفاده از الگوریتم آدابوست برای یادگیری انتخاب می‌نماییم، شامل C4.5، بیز ساده، شبکه‌های بیز، نزدیکترین همسایه (K-NN) و .... الگوریتم آدابوست از کل مجموعه داده به‌منظور آموزش هر طبقه‌بند استفاده می‌کند، اما بعد از هر بار آموزش، بیشتر بر روی داده‌های سخت تمرکز می‌کند تا به درستی رده‌بندی شوند. این روش تمرکز بیشتر بر روی نمونه‌هایی است که قبلاً به‌طور صحیح رده‌بندی نشده‌اند. در ابتدا تمام رکوردها وزن یکسانی می‌گیرند و وزن‌ها در هر تکرار افزایش پیدا خواهند کرد. وزن نمونه‌هایی که به اشتباه طبقه‌بندی شده‌اند افزایش خواهد یافت در حالی که آن دسته از نمونه‌هایی که به درستی رده‌بندی شده‌اند وزنشان کاهش خواهد یافت. سپس وزن دیگری به‌صورت مجزا به هر دسته‌کننده با توجه به‌دقت کلی آن اختصاص داده می‌شود که بعداً در مرحله آزمایش مورد استفاده قرار می‌گیرد. دسته‌کننده‌های دقیق از ضریب اطمینان بالاتری برخوردار خواهند بود. در نهایت هنگام ارائه یک نمونه جدید هر دسته‌کننده یک وزن پیشنهاد می‌دهد و برچسب رده با رأی اکثریت انتخاب خواهد شد. شبه کد این روش به‌صورت زیر می باشد:

Input: Training data set $D=\{(x_1,y_1),...,(x_n,y_n)\}$,
Learning algorithm L
Size of ensemble T

Process: $d_1(i)=\frac{1}{n}$ *for* $i=1...n$ -- Initialize distribution

for $t=1...T$

$C_t=L(D,d_t)$ -- Build classifier C_t with D and distribution d_t

$\varepsilon_t=\sum_{i:C_t(x_i)\neq y_i}^{n} d_t(i)$ -- Calculate training error for C_t

$\omega_t=\frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$ -- Calculate weight of classifier C_t

$d_{t+1}(i)=d_t(i)\times e^{\omega_t y_i C_t(x_i)}$ -- Calculate new weights of instances in training set

Outputs: $H(x)=sign\left(\sum_{t=1}^{T}\omega_t C_t(x)\right)$

۵-۲- وزن‌دهی هر طبقه‌بند و هر رده با استفاده از الگوریتم ژنتیک

فرض کنید هر کروموزوم از الگوریتم ژنتیک را به‌عنوان یک بردار وزن در نظر بگیریم. هر کروموزوم دارای یک مقدار برازش می‌باشد که به‌عنوان وزن هر رده و هر رده‌بند در نظر گرفته می‌شود. حال فرض کنید پارامترهای استفاده شده در روابطی که در ادامه تشریح خواهند شد به‌صورت زیر باشند:

n: تعداد رده‌بندها

c: تعداد رده‌ها

m: تعداد داده‌ها

فرض کنید $P_{i,j}$ به‌عنوان یک متغیر واسط جهت تبدیل کروموزم به ماتریس تصمیم در نظر گرفته شود که از رابطه (۱) محاسبه می‌شود (در این رابطه ch_k کروموزوم k-ام را نشان می‌دهد):

$$P_{i,j} = ch_k \qquad (۱)$$

وزن $w_{i,j}$ از رابطه (۲) محاسبه می‌شود ($W_{i,j}$ بردار وزنی است):

$$w_{i,j} = \frac{P_{i,j}}{\sum_{k=1}^{n} ch_k} \qquad (۲)$$

مقدار برازش یک کروموزم در روش پیشنهادی از رابطه زیر به‌دست می‌آید:

$$ch_i = \sum_{k=1}^{m}\sum_{j=1}^{c}\left|\sum_{i=1}^{n} w_{i,j} c_i^j(x_k) - t^j(x_k)\right| \qquad (۳)$$

در رابطه (۳) $t^j(x_k)$ ترانهاده داده‌های متعلق به هر رده می‌باشد که در این رابطه اگر x_k متعلق به c_j باشد مقدار ۱ و در غیر این‌صورت مقدار ۰ را می‌دهد (c_i^j) یعنی خروجی رده‌بند iام بر روی داده jام.

شکل ۴ کروموزوم‌ها در الگوریتم ما را نشان می‌دهد:

ch_1	ch_2	.....	ch_c	ch_{c+1}	.....	ch_{c+c}	ch_{c+c+c}	.....	$ch_{c+c+c+\cdots+c}$

شکل (۴): شکل کروموزوم‌ها در الگوریتم پیشنهادی.

حال با استفاده از الگوریتم (۱) روش پیشنهادی برای وزن‌دهی هر رده‌بند و هر رده توسط الگوریتم ژنتیک تشریح می‌شود:

الگوریتم ۱: الگوریتم GA برای انتخاب مدل ***GAWVote***.

ورودی: پارامترهای کنترل GA: عامل جهش F، نرخ تقاطع CR، و تعداد جمعیت N.

(۱)ارزش‌دهی اولیه: تولید جمعیت تصادفی توزیعی به‌طور یکنواخت N فرد به‌طوری که:

$$X_G = \{X_{1,G}, X_{2,G}, \ldots, X_{i,G}, \ldots, X_{N,G}\} \qquad (۴)$$

(۲) $\mathbf{T}_{i,G} = [x_{i,G}(1), x_{i,G}(2), \ldots, t_{i,G}(j), t_{i,G}(n)]$ بردار نمایش‌دهنده وزن‌های n $(w_1, w_2, \ldots, w_i, \ldots, w_n)$ رده‌بند پایه است.

(۳) $\mathbf{H}_{j,G} = [y_{j,G}(1), y_{j,G}(2), \ldots, h_{j,G}(j), ht_{j,G}(c)]$ بردار نمایش‌دهنده وزن‌های c $(w_1, w_2, \ldots, w_j, \ldots, w_c)$ رده از رده‌بندهای پایه است.

(۴) تکرارکننده تولید را $G = 0$ قرار بده.

(۵) تا زمانی که معیار توقف ارضاء نشده است کارهای زیر را انجام بده.

(۶) بردار جهش $V_{i,G}$ و $V_{j,G}$ را برای هر رده‌بند و هر رده با استفاده ازمرحله (۱) محاسبه کن.

(۷) برای $(\mathbf{i}=0; \mathbf{i} < n; i{+}{+})$ انجام بده.

(۸) برای $(\mathbf{j}=0; \mathbf{j} < c; j{+}{+})$ انجام بده.

(۹) فرد آزمایشی $\mathbf{T}_{i,G}$ را با استفاده از مرحله (۶) تصمیم گیری کن.

(۱۰) فرد آزمایشی $\mathbf{H}_{j,G}$ را با استفاده از مرحله (۶) تصمیم گیری کن.

(۱۱) پایان حلقه ***for***

(۱۲) برازش بردار $T_{i,G}$ و $X_{i,G}$ را با استفاده از اعتبارسنجی متقابل ۱۰ تایی، و به‌هنگام‌سازی بردار $X_{i,G+1}$ نسل بعد

(۱۳) پایان حلقه ***for***

(۱۴) تکرارکننده نسل G = G + 1 را به‌هنگام‌سازی کن.

(۱۵) پایان حلقه ***while***

خروجی: وزن‌های بهینه $(w_{1,1}, w_{2,2}, \ldots, w_{i,j}, \ldots, w_{n,c})$ برای ***GAWVote***.

۵-۳- مدل مبتنی بر GA برای انتخاب پارامترها

دراین بخش، تمرکز برای انتخاب پارامتر برای ***GAWVote*** پیشنهادی است. پارامترهایی که باید در ***GAWVote*** بهینه شوند وزن‌های رده‌بندهای مبنا و وزن‌های هر رده از رده‌بندهای مبنا در یک مجموعه می‌باشند. تنظیمات مختلف پارامترها تاثیر زیادی بر عملکرد ***GAWVote*** دارد. الگوریتم ژنتیک برای جستجوی وزن‌های بهینه را انتخاب می‌نماییم.

GA دارای یک جمعیت اولیه تصادفی از کاندیدهای راه حل می‌باشد که سپس با استفاده از عملیات تکاملی بهبود می‌یابد. به‌طور کلی، از حداکثر تکرارهای از پیش تعریف‌شده $\mathbf{M}_{max}$ برای تعیین معیار توقف $(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_j, \ldots, \mathbf{w}_D)$ استفاده شده است. سایر پارامترهای کنترل برای N عبارتند از: عامل مقایس‌بندی جهش $F \in (0,1)$، نرخ تقاطع $\mathbf{CR} \in (0,1)$ و تعداد جمعیت N. پروسه انتخاب پارامترهای مبتنی بر ***GA*** برای ***GAWVote*** در الگوریتم ۱ با توضیحات زیر ارائه شده است:

ارزش‌دهی اولیه-ارزش‌دهی اولیه جمعیت N فرد:

$\mathbf{X}_G=\{\mathbf{X}_{1,G},\mathbf{X}_{2,G},\ldots,\mathbf{X}_{i,G},\ldots,\mathbf{X}_{N,G}\}$. یک فرد به‌عنوان یک بردار D بعدی تعریف شده است: $\mathbf{X}_{i,G}=[\mathbf{x}_{i,G}(1),\mathbf{x}_{i,G}(2),\ldots,\mathbf{x}_{i,G}(\mathbf{j}),\ldots,\mathbf{x}_{i,G}(\mathbf{n})]$ است، که نشان‌دهنده وزن‌های $(\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_j,\ldots,\mathbf{w}_n)$ دسته‌بندی کننده‌های مبنا می‌باشند و n اندازه دسته‌بندی‌کننده‌های مبنا است. هر فرد از طریق توزیع یکنواخت در دامنه [۰،۱] تولید می‌شود. ارزیابی برازش ***GAWVote*** را با استفاده از هر بردار فردی آموزش داده می‌شود، و سپس صحت اعتبارسنجی متقابل ۱۰ تایی متناظر به‌عنوان تابع برازش ارزیابی می‌شود. با توجه به تعداد رده‌های C و n طبقه‌بندهای مبنا برای رای‌گیری، طبقه پیش‌بینی $\mathbf{c}_k$ رأی‌گیری وزنی برای هر نمونه k به‌صورت زیر بیان می‌شود:

$$c_k = \arg_j \max \sum_{i=1}^{D} (\Delta_{ji} \times w_{i,j}) \qquad (۵)$$

که Δ_{ji} متغیر دودویی است. اگر طبقه‌بند مبنا i ام نمونه k را به رده j ام طبقه‌بندی نماید، آنگاه $\Delta ji=1$ در غیر این‌صورت، $\Delta_{ji}=0$. $w_{i,j}$ وزن طبقه‌بند مبنا i ام (رده j ام) در یک مجموعه است که توسط الگوریتم ***GA*** در الگوریتم ۱ بهینه می‌شود. به این ترتیب، دقت تعریف می‌شود:

$$Acc=\frac{\sum_k\{1|ck\text{ است}\ k\text{ طبقه واقعی نمونه}\}}{\text{اندازه نمونه های آزمایش}}\times 100\% \qquad (۶)$$

بعد از به‌دست آوردن بهترین نمونه از طریق الگوریتم ژنتیک یعنی، وزن بهینه $(\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_j,\ldots,\mathbf{w}_D)$ نویسندگان این مقاله دسته‌بندی‌کننده جمعی را برای دسته‌بندی مجموعه داده‌های آزمایش تولید می‌نمایند.

۵-۴- پیش پردازش داده‌ها

تهیه مجموعه داده‌های واقعی با استفاده از هانی‌پات‌ها و هانی‌نت‌ها می‌تواند آسان باشد. به‌منظور به‌دست آوردن یک دید بهتر نسبت به حملات رایانه‌ای و ایجاد مجموعه داده‌ای برچسب‌گذاری‌شده از چنین حملاتی، با به‌کارگیری هانی‌پات‌ها و هانی‌نت به‌عنوان یک ابزار جمع‌آوری داده‌ها برای ارزیابی روش‌های توسعه‌یافته در این تحقیق بهره گرفته شده است. از حدود تقریباً ۲۱۶۰۲ اتصال به هانی‌نت حدود ۲۰۲۳۷ غیر حمله و حدود ۲۳۶۴ حمله تشخیص داده شده است. این اتصالات بر اساس سه پروتکل TCP، UDP و ICMP می‌باشد که ویژگی‌های مجموعه داده تلقی می‌گردند. با توجه به اینکه هدف تشخیص حمله بودن و یا حمله نبودن داده‌ها می‌باشد در این تحقیق مجموعه داده را به دو رده برچسب حمله و غیر حمله تقسیم شده و بر این اساس مدل آموزش داده شد. مجموعه داده دیگری که در این تحقیق جهت آموزش و آزمایش مدل از آن‌ها بهره برده شده، مجموعه داده DARPA99 می‌باشد. از میان مجموعه داده DARPA99، داده‌های دو روز از آن را که حدوداً ۱۴۱۰۲۶ رکورد می‌باشد مورد استفاده قرار گرفت. در این مجموعه داده نیز دو رده برچسب حمله و غیر حمله در نظر گرفته شد که ۱ نشان‌دهنده غیر حمله بودن و ۲ نشان‌دهنده حمله بودن داده‌ها تعریف گردیده است. مجموعه داده سوم شامل مجموعه داده NSL-KDD می‌باشد. این مجموعه داده رکوردهای انتخابی از مجموعه داده KDD می‌باشند که شامل ۴۱ ویژگی و ۱۲۵۹۷۳ رکورد با برچسب‌های حمله و نرمال هستند.

۶- روش‌های ارزیابی روش پیشنهادی

برای ارزیابی عملکرد مدل‌های یادگیری و همچنین شناخت مدلی که بهترین عملکرد را دارد نیاز به امتیازبندی مدل‌ها می‌باشد. معیارهای آماری دقت، یادآوری و معیار F نیازمند درک قوی از دو نوع پیش‌بینی خطا، مثبت‌های کاذب و منفی‌های کاذب می‌باشند. یکی دیگر از روش‌ها برای رتبه‌بندی روش‌ها استفاده از آزمون آماری می‌باشد. از معیار ناحیه زیر نمودار ROC برای این‌که نشان داده شود کدام روش بهترین عملکرد را بر اساس مجموعه داده‌های مختلف دارد نیز استفاده شد.

معیار F یا امتیاز F عموماً برای مواردی به کار گرفته می‌شود که در آن‌ها هدف بازیابی عناصر وابسته به یک رده خاص به‌صورت صحیح باشد به‌گونه‌ای که عناصر زیادی از سایر رده‌ها در آن نباشد. این معیار شامل پارامترهای p (دقت) و r (یادآوری) برای آزمون جهت محاسبه امتیاز می‌باشد، که در آن p تعداد نتایج صحیح مثبت تقسیم بر تعداد تمام نتایج مثبت است و r نیز تعداد نتایج صحیح مثبت تقسیم بر جمع منفی کاذب و صحیح مثبت می‌باشد [۳۴].

$$p=\frac{TP}{TP+FP} \qquad r=\text{sensitivity}=\frac{TP}{TP+FN} \qquad (۷)$$

$$F_1=2\times\frac{precision\times recall}{precision+recall} \qquad (۸)$$

ماتریس درهم ریختگی به ماتریسی گفته می‌شود که در آن عملکرد الگوریتم‌های مربوطه را نشان می‌دهند. در این تحقیق در مسأله تشخیص حملات، چهار دسته مثبت و منفی یا حمله و غیر حمله وجود دارد که همان‌طور که در جدول (۱) قابل مشاهده است هر ستون از ماتریس، نمونه‌ای از مقدار پیش‌بینی‌شده را نشان می‌دهد.

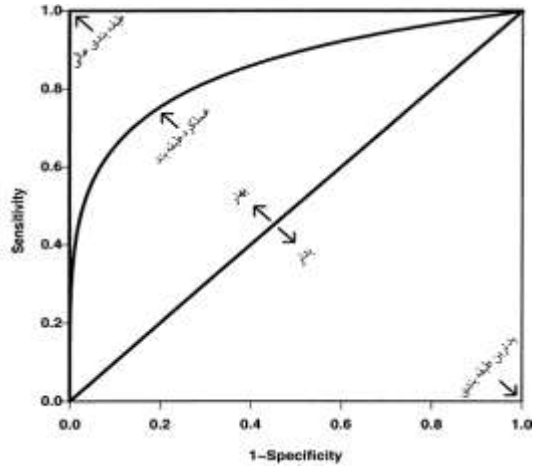
جدول (۱): ماتریس درهم ریختگی

نوع دسته		تشخیص داده شده	
		مثبت	منفی
واقعی	مثبت	TP	FN
	منفی	FP	TN

منحنی ROC یک نمودار گرافیکی بین نرخ مثبت صحیح (TPR) یا حساسیت (یا یادآوری) و نرخ مثبت کاذب (FPR) یا specificity-۱ می‌باشد. در حقیقت منحنی ROC عملکرد یک سیستم طبقه‌بند دودویی را نشان می‌دهد. معیار AUC که سطح زیر نمودار ROC را نشان می‌دهد می‌تواند نقش تعیین‌کننده‌ای در معرفی طبقه‌بند برتر ایفا کند.

$$TPR=\frac{TP}{TP+FN} \qquad FPR=\frac{FP}{FP+TN} \qquad \text{specificity}=\frac{TN}{TN+FP} \qquad (۹)$$

شکل (۵) نمودار ROC را نمایش می‌دهد. نقاط روی منحنی عملکرد طبقه‌بند را نشان می‌دهند. بهترین مدل، مدلی می‌باشد که دارای نرخ مثبت صحیح برابر ۱ و نرخ مثبت کاذب ۰ می‌باشد یعنی نقطه (۱،۰) و بدترین سناریو زمانی می‌باشد که نرخ صحیح مثبت ۰ و نرخ مثبت کاذب ۱ باشد یعنی نقطه (۰،۱).



شکل (۵): نمای کلی نمودار ROC [۳۴].

نمودار DET یک نمودار گرافیکی از نرخ خطا برای سیستم‌های طبقه‌بندی دودویی است که شامل نرخ مثبت کاذب (FPR) در برابر نرخ منفی کاذب (FNR) می‌باشد. در این نمودار هرچه عملکرد سیستم بهتر باشد منحنی به مبدا نزدیکتر خواهد بود.

آزمونی که در این تحقیق از آن استفاده شد، آزمون T زوجی می‌باشد. زمانی از این آزمون استفاده می‌شود که نمونه یکسانی در دو وضعیت متفاوت مورد آزمون قرار گیرد.

بیشنه احتمال ارتکاب خطای نوع اول با α نشان داده و سطح معنی‌دار نیز گفته می‌شود. سطح معنی‌دار در علم آمار به $P-value$ (مقدار احتمال) معروف است. در تفسیر نتایج هر آزمونی، اعم از پارامتری و ناپارامتری، هر چه مقدار سطح معنی‌دار کوچک‌تر باشد، رد فرض صفر ساده‌تر است. آلفا (α) سطح خطایی است که محقق در نظر می‌گیرد (مقدار آلفا معمولاً ۰/۰۵ یا ۵ درصد است). و بر همین اساس مقدار P از رابطه زیر محاسبه می‌شود.

$$P=1-\frac{\alpha}{2} \qquad (۱۰)$$

چون در این تحقیق ۱۳ روش مختلف که هرکدام ۳۰ بار اجرا گردیده‌اند با هم مقایسه خواهند شد، بنابراین ۲۹ = ۱− ۳۰ درجه آزادی خواهد بود. برای محاسبه پارامترها رابطه زیر را خواهیم داشت [۳۴]:

$$t=\frac{\bar{d}}{\sqrt{\frac{s^2}{n}}} \qquad (۱۱)$$

در معادله بالا $\bar{d}$ برابر با میانگین اختلاف بین دو نمونه، s^2 معادل واریانس نمونه، n برابر تعداد نمونه و t نیز آزمون T زوجی با درجه آزادی $n-1$ می‌باشد. یک فرمول جایگزین برای این آزمون به‌صورت زیر می‌باشد [۳۴].

$$t=\frac{\sum d}{\sqrt{\frac{n\left(\sum d^2\right)-\left(\sum d\right)^2}{n-1}}} \qquad (۱۲)$$

۷- تجزیه و تحلیل داده‌ها و مقایسه روش‌ها

نتایج حاصل از سه مجموعه داده DARPA99، هانی‌پات (RealTraffic) و NSL-KDD برای الگوریتم پیشنهادی و دوازده روش دیگر مورد ارزیابی و مقایسه قرار گرفته است. نتایج به‌دست آمده میانگین ۳۰ بار اجرا می‌باشد. این روش‌ها به‌ترتیب عبارتند از طبقه‌بند fisher، طبقه‌بند درجه دوم، Uncorrelated normal based quadratic Bayes classifier که این روش به محاسبه طبقه‌بند درجه دوم بین رده‌ها در مجموعه داده A با فرض تراکم عادی/معمولی با ویژگی‌های ناهمبسته می‌پردازد، طبقه‌بند درخت تصمیم آماری، طبقه‌بند درخت تصمیم، طبقه‌بند بیز ساده هر رده و هر ویژگی را به‌طور جداگانه تخمین می‌زند. فرض در بیز ساده این است که ویژگی‌ها به‌طور شرطی از هم مستقل هستند، طبقه‌بند BagEnsNaiveBC بر اساس طبقه‌بند بگینگ و با استفاده از اجماع پانزده‌تایی از طبقه‌بند بیز ساده ایجاد می‌شود، طبقه‌بند BagEnsWeakC بر اساس طبقه‌بند بگینگ و با استفاده از اجماع پانزده‌تایی از طبقه‌بند Weak ایجاد می‌شود، طبقه‌بند BagEnsDT بر اساس اجماع بگینگ و درخت تصمیم با استفاده از K نزدیک‌ترین همسایه می‌باشد، روش بعد اجماع بوستینگ با استفاده از ADABOOST که شامل یک اجماع پانزده‌تایی از

طبقه‌بند بیز ساده می‌باشد، طبقه‌بند BoostEnsDT بر اساس اجماع بوستینگ و درخت تصمیم با استفاده از طبقه‌بند ADABOOST می‌باشد. در حقیقت محاسبه ترکیب طبقه‌بندها بر اساس ADABOOST می‌باشد. در مجموع ورژن‌های N وزنی که از مجموعه آموزشی A تولید می‌شوند (A مجموعه داده و N تعداد طبقه‌بندهای آموزش داده شده می‌باشد) تکرار می‌گردند و برای آموزش یک طبقه‌بند مشخص استفاده می‌شود. وزن‌ها استفاده شده برای احتمالات اشیا در مجموعه آموزشی انتخاب شده بر اساس قانون ADABOOST به روز رسانی می‌شوند و روش آخر الگوریتم ماشین بردار پشتیبان می باشد. در جدول (۲) روش‌های مختلف بر اساس خطای معیار فیشر مقایسه گردیده‌اند.

نتایج حاصل نشان می‌دهد که الگوریتم پیشنهادی بعد از روش DTC برای هر سه مجموعه داده نسبت به سایر روش‌ها از خطای فیشر مناسب‌تری برخوردار می‌باشد.

جدول (۳) درصد دقت انتخاب طبقه‌بندهای مختلفی را نشان می‌دهد. برای مجموعه داده DARPA99 روش SVM و روش الگوریتم پیشنهادی از دقت بیشتری نسبت به سایر روش‌ها برخوردار می‌باشند. برای مجموعه داده هانی‌پات‌ها روش الگوریتم پیشنهادی و درخت تصمیم بهترین روش‌ها می‌باشند. برای مجموعه داده NSL-KDD روش پیشنهادی، طبقه بند درخت تصمیم و StatsDTC بیشترین دقت را دارند.

جدول(۲): درصد معیار فیشر(خطا) روش‌های گوناگون

	DARPA99	Real Traffic	NSL-KDD
Proposed	۰/۰۰۴۵ ± ۰/۰۰۱۹	۷/۱۴۴۶ ± ۰/۰۷۰۵	۰/۲۵۸۰ ± ۰/۰۹۹۷
Fisher	۰/۶۰۶۵ ± ۰/۰۰۴۹	۱۱/۷۴۸۳ ± ۰/۰۵۷۸	۷/۴۷۲۶ ۰/۰۱۰۸ ±
Quadratic	۱۸/۱۲۶۸ ± ۱۷/۶۴۳۰	۱۱/۸۰۶۹ ± ۰/۰۳۲۰	۲۹/۰۱۴۲ ± ۶/۷۶۳۷
UDC	۴/۶۸۸۲ ± ۰/۰۰۵۹	۱۳/۳۴۰۷ ± ۰/۰۶۲۹	۱۷/۴۳۱۳ ± ۳/۱۴۷۰
StatsDTC	۰/۰۰۶۷ ± ۰/۰۰۱۶	۹/۱۴۰۹ ± ۰/۱۵۲۱	۰/۲۱۴۹ ± ۰/۰۰۹۹
DTC	۰/۰۰۴۲ ± ۰/۰۰۱۲	۷/۱۳۱۵ ± ۰/۰۶۴۶	۰/۲۱۲۲ ± ۰/۰۰۹۷
NaiveBC	۳/۸۹۲۰ ± ۰/۰۱۸۲	۱۱/۵۰۴۵ ± ۰/۱۸۵۹	۶/۲۴۶۹ ± ۰/۰۸۲۲
BagEnsNaiveBC	۳/۹۱۷۷ ± ۰/۰۲۶۱	۱۱/۸۹۰۱ ± ۰/۱۹۵۸	۶/۲۷۸۶ ± ۰/۱۲۰۶
BagEnsWeakC	۲/۵۵۱۲ ± ۰/۶۱۵۱	۵۰/۰۰۰۰ ± ۰/۰۰۰۰	۴۶/۷۰۲۴ ± ۶/۴۹۴۳
BagEnsDT	۰/۹۹۱۶ ± ۰/۱۰۲۳	۲۱/۸۲۰۸ ± ۰/۴۴۵۵	۴/۹۶۲۶ ± ۰/۲۶۸۴
ADABOOSTC	۰/۷۳۴۲ ± ۰/۲۷۹۴	۱۳/۵۳۹۸ ± ۰/۳۲۹۸	۴/۶۹۸۶ ± ۰/۳۶۳۴
BoostEnsDT	۰/۱۸۳۹ ± ۰/۱۰۴۲	۱۵/۴۶۰۴ ± ۱/۳۲۹۵	۱/۲۲۷۵ ± ۰/۱۵۴۸
SVM	۰/۰۰۱۴ ± ۰/۰۰۰۰	۱۲/۸۵۳۲ ± ۰/۵۸۷۵	۶/۲۶۸۳ ± ۰/۳۹۱۵

جدول (۳): درصد دقت انتخاب روش‌های گوناگون

	DARPA99	Real Traffic	NSL-KDD
Proposed	۹۹/۹۹۵۵ ± ۰/۰۰۱۹	۹۲/۸۵۱۰ ± ۰/۰۳۶۵	۹۹/۷۴۲۰ ± ۰/۰۹۹۸
Fisher	۹۹/۴۱۴۳ ± ۰/۰۰۳۶	۸۹/۷۵۶۹ ۰/۰۲۲۵±	۹۲/۵۱۹۰ ± ۰/۰۱۰۹
Quadratic	۷۳/۴۲۱۶ ± ۲۳/۷۷۴۸	۸۷/۸۳۱۸ ۰/۰۴۱۳±	۶۷/۱۴۱۶ ± ۸/۱۸۹۴
UDC	۹۱/۵۸۴۰ ± ۰/۰۱۰۹	۸۸/۱۹۳۱ ۰/۰۵۴۴±	۸۱/۵۴۸۱ ± ۳/۰۵۱۰
StatsDTC	۹۹/۹۹۳۳ ± ۰/۰۰۱۶	۹۱/۴۶۳۲ ۰/۱۳۷۷±	۹۹/۷۸۵۱ ± ۰/۰۰۹۹
DTC	۹۹/۹۹۵۸ ± ۰/۰۰۱۲	۹۲/۸۵۶۰ ± ۰/۰۴۳۲	۹۹/۷۸۷۸ ± ۰/۰۰۹۷
NaiveBC	۹۳/۰۵۱۲ ± ۰/۰۳۳۹	۸۹/۸۷۳۴ ۰/۰۸۹۰±	۹۳/۷۵۰۶ ± ۰/۰۸۲۹
BagEnsNaiveBC	۹۳/۰۰۳۵ ± ۰/۰۴۸۵	۸۹/۶۶۹۰ ۰/۱۰۶۹±	۹۳/۷۱۸۲ ± ۰/۱۲۱۴
BagEnsWeakC	۹۵/۹۵۹۰ ± ۱/۲۴۹۵	۸۹/۵۴۰۲ ± ۷/۲۲۶۸	۵۳/۴۸۴۶ ± ۰/۰۷۸۴
BagEnsDT	۹۸/۴۹۴۰ ± ۰/۱۸۴۳	۷۲/۲۰۰۱ ۰/۷۲۸۹±	۹۵/۰۲۹۲ ± ۰/۲۷۰۷
ADABOOSTC	۹۸/۹۳۲۷ ± ۰/۴۷۴۰	۸۴/۹۳۰۶ ± ۰/۵۲۳۴	۹۵/۲۹۹۸ ± ۰/۳۶۲۷
BoostEnsDT	۹۹/۷۸۵۵ ± ۰/۱۳۷۱	۸۰/۹۹۳۹ ۲/۲۷۲۰±	۹۸/۷۷۱۹ ± ۰/۱۵۵۰
SVM	۹۹/۹۹۸۵ ± ۰/۰۰۰۰	۸۵/۹۴۱۳ ۰/۹۰۵۹±	۹۳/۵۴۷۷ ± ۰/۳۵۱۷

جدول (۴) مقایسه‌ای از رتبه معیار فیشر برای روش‌های گوناگون را بر اساس آزمون آماری نشان می‌دهد. همان‌طور که قابل مشاهده است الگوریتم پیشنهادی بر روی مجموعه داده DARPA99 رتبه دوم و برای مجموعه داده‌های RealTraffic و NSL-KDD دارای رتبه یک می‌باشد.

جدول (۵) درصد ناحیه زیر نمودار ROC را نشان می‌دهد. هرچه این درصد بیشتر باشد بیانگر این است که روش مورد نظر عملکرد بهتری را دارد. الگوریتم پیشنهادی برای هر سه مجموعه داده دارای درصد قابل قبولی می‌باشد. همانطور که در جدول (۶) نمایش داه شده است الگوریتم پیشنهادی برای مجموعه DARPA99 و هانی‌پات رتبه دوم و برای مجموعه داده NSL-KDD رتبه اول را دارد. جدول (۷) مقایسه‌ای از معیار زمان اجرا برای روش‌های گوناگون را نشان می‌دهد. روش پیشنهادی دارای زمان بیشتری نسبت به سایر روش‌ها می‌باشد.

جدول(۴): مقایسه رتبه معیار فیشر روش‌های گوناگون با آزمون آماری

	DARPA99	Real Traffic	NSL-KDD
Proposed	۲	۱	۱
Fisher	۶	۵	۱۰
Quadratic	۱۳	۶	۱۲
UDC	۱۲	۹	۱۱
StatsDTC	۴	۳	۲
DTC	۳	۲	۳
NaiveBC	۱۰	۴	۷
BagEnsNaiveBC	۱۱	۷	۸
BagEnsWeakC	۹	۱۳	۱۳
BagEnsDT	۸	۱۲	۶
ADABOOSTC	۷	۱۰	۵
BoostEnsDT	۵	۱۱	۴
SVM	۱	۸	۹

جدول (۵): درصد ناحیه زیر نمودار ROC روش‌های گوناگون

	DARPA99	Real Traffic	NSL-KDD
Proposed	۹۹/۹۵۰۲	۸۰/۲۶۷۴	۹۹/۹۸۵۶
Fisher	۹۸/۶۱۵۱	۷۴/۴۳۳۵	۹۷/۹۷۹۰
Quadratic	۶۸/۴۴۳۰	۷۳/۰۷۰۸	۷۳/۶۰۷۴
UDC	۸۹/۱۷۶۸	۶۵/۲۰۲۸	۷۹/۳۵۱۹
StatsDTC	۹۹/۹۴۵۶	۷۹/۶۰۰۸	۹۹/۸۷۱۵
DTC	۹۹/۹۱۲۷	۷۹/۹۸۷۹	۹۹/۹۶۶۷
NaiveBC	۹۴/۴۰۹۱	۶۲/۰۴۳۸	۸۱/۲۲۴۹
BagEnsNaiveBC	۹۴/۴۸۵۱	۶۲/۰۰۷۶	۸۱/۴۱۹۹
BagEnsWeakC	۷۱/۸۱۰۹	۵۰/۰۰۴۳	۵۰/۲۸۴۰
BagEnsDT	۹۹/۶۵۲۲	۶۵/۰۳۵۷	۹۷/۸۲۶۴
ADABOOSTC	۸۷/۵۰۴۰	۷۴/۹۴۱۰	۹۸/۶۰۶۸
BoostEnsDT	۹۹/۸۹۱۴	۸۰/۸۶۷۶	۹۹/۸۵۸۳
SVM	۹۹/۹۹۹۴	۷۳/۴۱۸۵	۹۸/۸۹۳۵

جدول (۶): رتبه‌بندی بر اساس ناحیه زیر نمودار ROC روش‌های گوناگون

	DARPA99	Real Traffic	NSL-KDD
Proposed	۲	۲	۱
Fisher	۷	۶	۷
Quadratic	۱۳	۸	۱۲
UDC	۱۰	۹	۱۱
StatsDTC	۳	۴	۳
DTC	۴	۳	۲
NaiveBC	۹	۱۱	۱۰
BagEnsNaiveBC	۸	۱۲	۹
BagEnsWeakC	۱۲	۱۳	۱۳
BagEnsDT	۶	۱۰	۸
ADABOOSTC	۱۱	۵	۶
BoostEnsDT	۵	۱	۴
SVM	۱	۷	۵

جدول (۷): معیار زمان روش‌های گوناگون (بر حسب ثانیه)

	DARPA99	Real Traffic	NSL-KDD
Proposed	۸۰۴/۸۳۸۲	۱۸۹/۳۳۱۷	۱۷۷۵/۳
Fisher	۱/۳۹۷۹	۰/۲۴۷۳	۱/۷۸۸۸
Quadratic	۱/۲۷۱۸	۰/۲۴۶۵	۱/۴۲۱۷
UDC	۱/۲۹۹۰	۰/۳۲۷۰	۱/۴۶۲۴
StatsDTC	۱/۶۸۴۱	۰/۴۸۷۸	۵/۱۱۴۳
DTC	۳۲/۲۸۸۲	۷/۷۵۹۷	۱۱۶/۱۵۵۷
NaiveBC	۱۰/۶۸۶۲	۱/۷۴۵۰	۱۶/۰۸۴۴
BagEnsNaiveBC	۱۲/۶۷۵۱	۲/۵۵۹۳	۱۸/۰۶۱۸
BagEnsWeakC	۸/۶۰۱۰	۴/۳۲۵۰	۹/۶۳۰۴
BagEnsDT	۸/۰۸۰۸	۱/۳۰۲۲	۷/۵۴۰۱
ADABOOSTC	۴/۸۸۹۴	۱/۶۲۳۸	۷/۱۸۷۴
BoostEnsDT	۷/۶۸۰۹	۴۳/۶۰۲۷	۲۳۸/۹۱۰۲
SVM	۱۷/۱۳۲۷	۴۳/۹۸۲۰	۱۷/۰۴۰۰

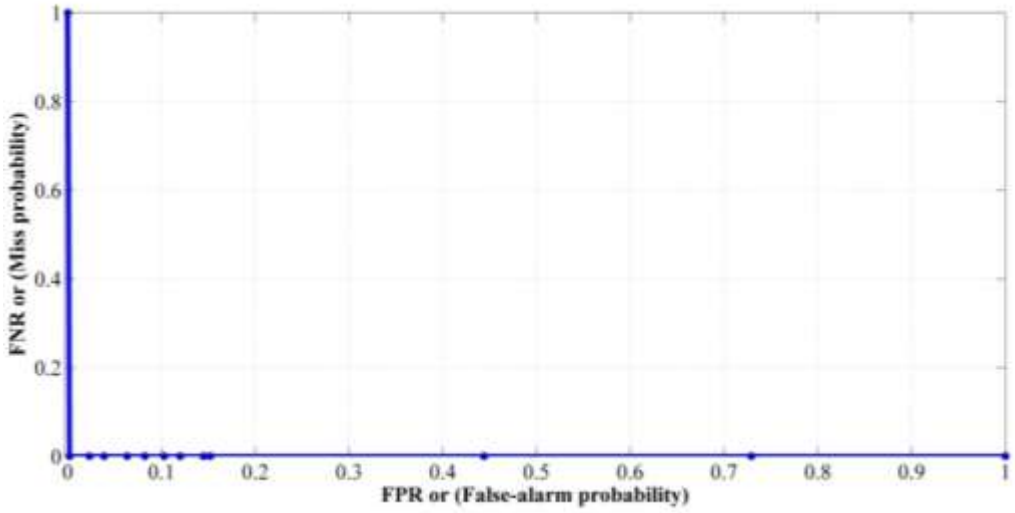
شکل‌های (۸-۶) خروجی روش پیشنهادی را بر اساس نمودار DET و برای سه مجموعه داده نشان می‌دهند. هرچه نرخ منفی کاذب کم‌تر می شود نرخ مثبت کاذب افزایش می‌یابد که بیانگر نسبت عکس این دو نسبت به هم می‌باشد.

شکل‌های (۱۱-۹) خروجی سایر روش‌ها را بر اساس نمودار DET و برای هر سه مجموعه داده نشان می‌دهند.

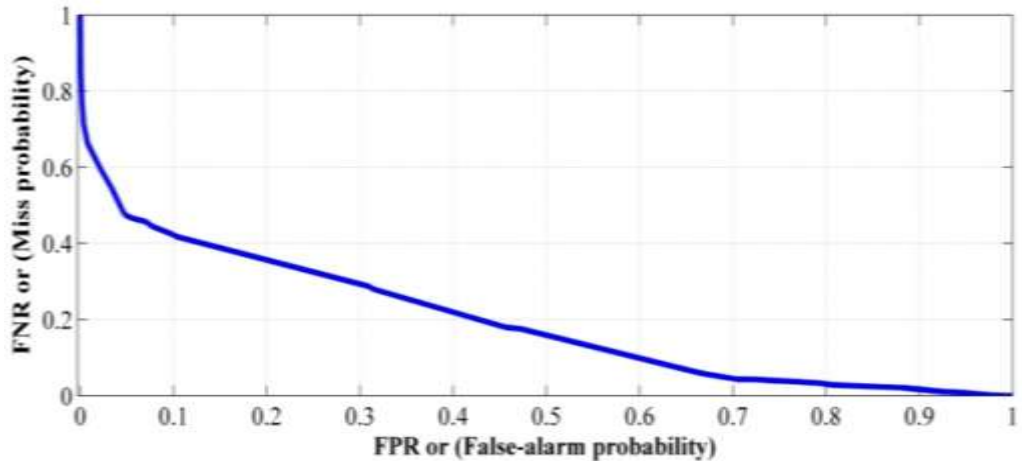
شکل‌های (۱۴-۱۲) نیز ناحیه زیر نمودار ROC بر اساس نرخ نمونه برداری از مجموعه داده‌ها را نشان می‌دهند.

جدول (۸) نتایج به‌دست آمده روش‌های مختلف بر اساس ماتریس در هم‌ریختگی را نشان می‌دهد. جدول (۹) خلاصه مقایسه نتایج به‌دست آمده با نتایج جدیدترین تحقیقات در زمینه تشخیص حملات و نفوذ را نشان می‌دهد. قابل توجه می‌باشد که هر روش نقاط قوت و نقاط ضعفی دارد که امکان مقایسه و

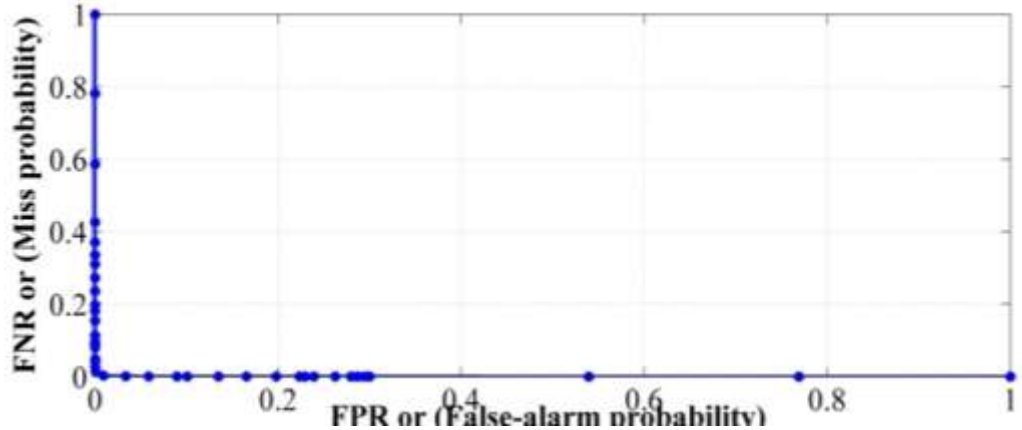
انتخاب بهترین روش را دچار مشکل می‌نماید.



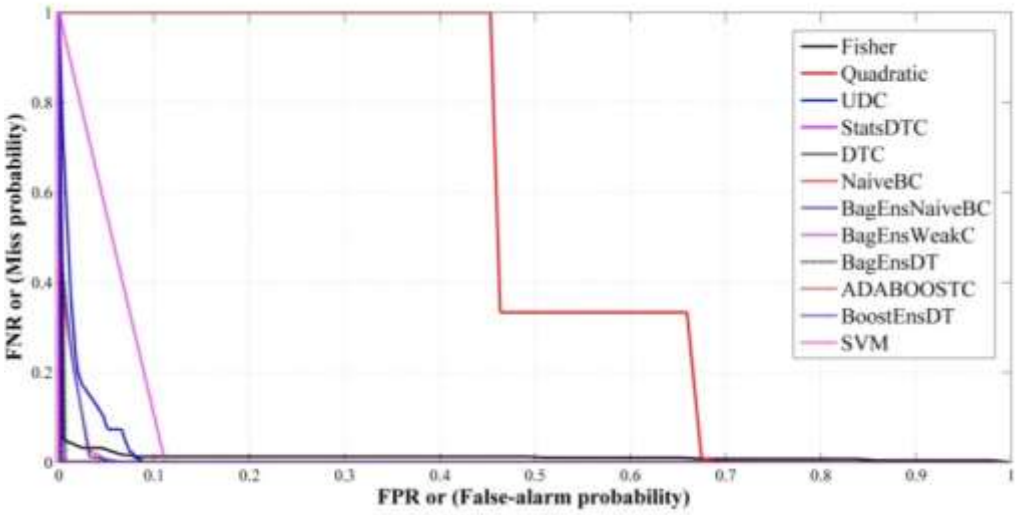
شکل (۶): DET روش پیشنهادی (مجموعه داده DARPA99)



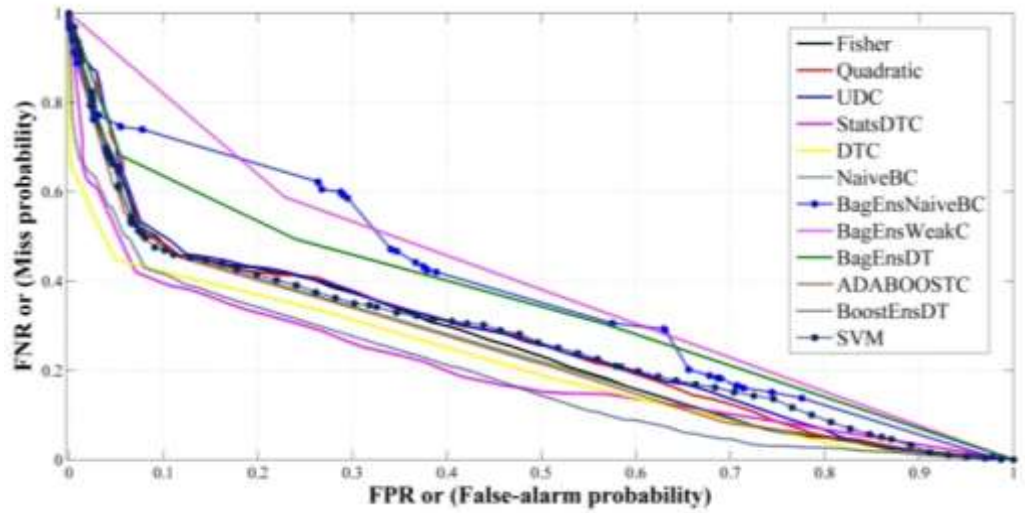
شکل (۷): DET روش پیشنهادی (مجموعه داده Real Traffic)



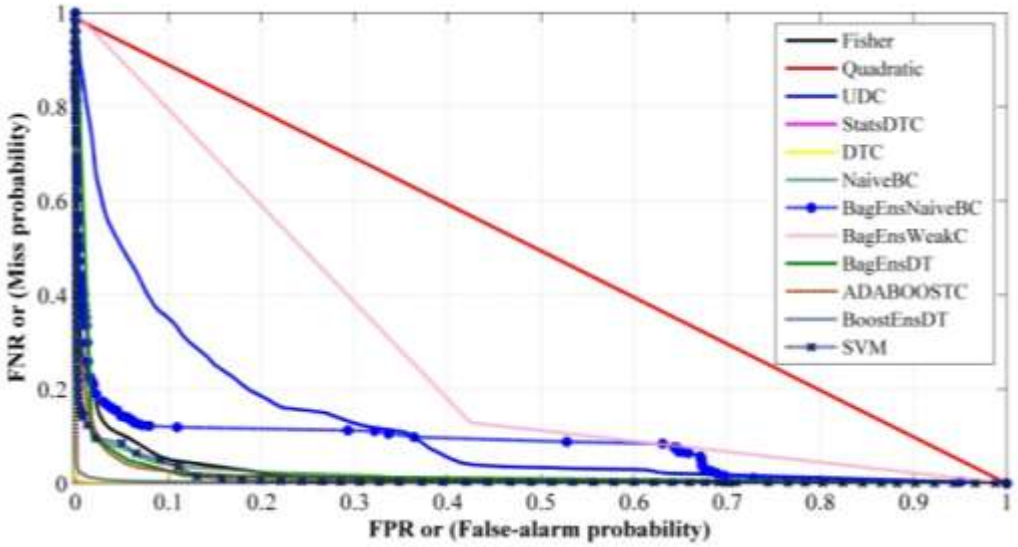
شکل (۸): DET روش پیشنهادی (مجموعه داده NSL-KDD)



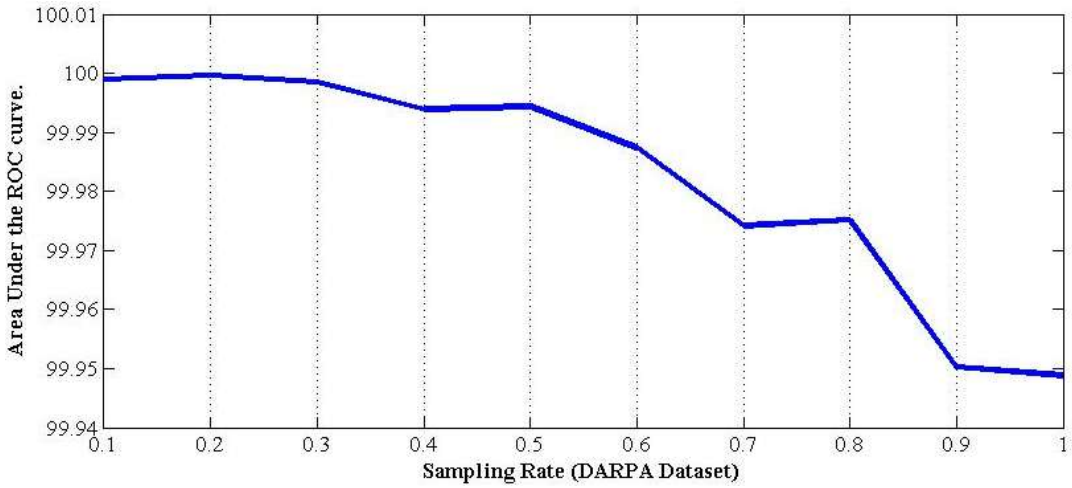
شکل (۹): DET مربوط به دوازده روش دیگر (مجموعه داده DARPA99)



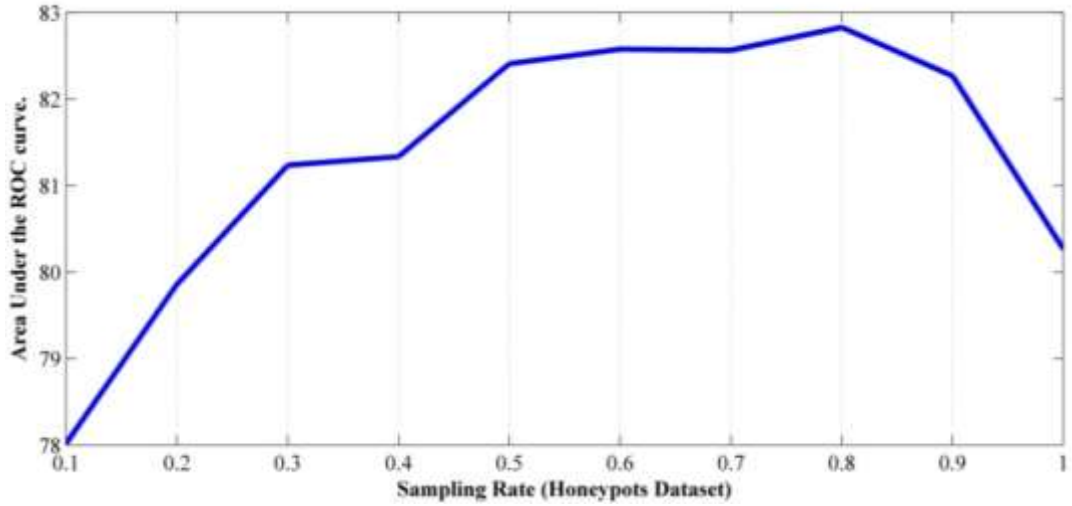
شکل (۱۰): DET مربوط به دوازده روش دیگر (مجموعه داده Real Traffic)



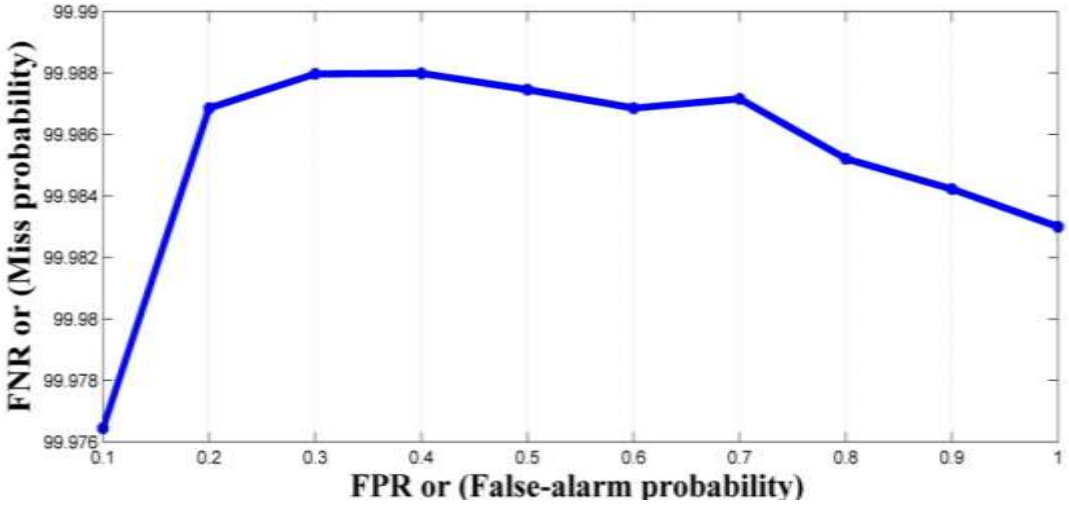
شکل(۱۱): DET مربوط به دوازده روش دیگر (مجموعه داده NSL-KDD)



شکل (۱۲): ناحیه زیر نمودار ROC (دقت) بر اساس نرخ نمونه‌برداری مجموعه داده DARPA99



شکل (۱۳): ناحیه زیر نمودار ROC (دقت) بر اساس نرخ نمونه برداری مجموعه داده Real Traffic



شکل (۱۴): ناحیه زیر نمودار ROC (دقت) بر اساس نرخ نمونه برداری مجموعه داده NSL-KDD

جدول (۸): نتایج به‌دست آمده بر اساس ماتریس درهم‌ریختگی

	DARPA99				Real Traffic				NSL-KDD			
	FP	TP	FN	TN	FP	TP	FN	TN	FP	TP	FN	TN
Proposed	۳	۹۷۰	۳	۱۴۰۰۵۰	۷۴	۸۱۵	۱۵۴۹	۲۰۱۶۳	۱۷۵	۵۸۱۶۱	۴۶۹	۶۷۱۶۸
Fisher	۱۳	۱۷۳	۸۰۰	۱۴۰۰۴۰	۱۳۹	۱۹۶	۲۱۶۸	۲۰۰۹۸	۵۱۷۲	۵۴۳۸۲	۴۲۴۸	۶۲۱۷۱
Quadratic	۳۰۰۵۳	۱۵۸	۸۱۵	۱۱۰۰۰۰	۱۵۳۲	۱۱۵۵	۱۲۰۹	۱۸۷۰۵	۴۰۶۳	۴۲۲۵۵	۱۶۳۷۵	۶۳۲۸۰
UDC	۱۱۸۶۲	۹۶۵	۸	۱۲۸۱۹۱	۷۰۵	۴۱۶	۱۹۴۸	۱۹۵۳۲	۲۳۹۰۷	۵۰۳۷۸	۸۲۵۲	۴۳۴۳۶
StatsDTC	۸	۹۷۲	۱	۱۴۰۰۴۵	۴۵۲	۸۹۷	۱۴۶۷	۱۹۷۸۵	۱۵۸	۵۸۵۰۳	۱۲۷	۶۷۱۸۵
DTC	۱	۹۷۰	۳	۱۴۰۰۵۲	۴۸	۷۹۸	۱۵۶۶	۲۰۱۸۹	۱۱۸	۵۸۴۹۱	۱۳۹	۶۷۲۲۵
NaiveBC	۹۷۷۸	۹۷۳	۰	۱۳۰۲۷۵	۱۷۳	۲۶۵	۲۰۹۹	۲۰۰۶۴	۴۰۹۵	۵۴۹۰۸	۳۷۲۲	۶۳۲۴۸
BagEnsNaiveBC	۹۸۴۲	۹۷۳	۰	۱۳۰۲۱۱	۱۶۸	۲۵۶	۲۱۰۸	۲۰۰۶۹	۴۲۲۲	۵۴۸۹۳	۳۷۳۷	۶۳۱۲۱
BagEnsWeakC	۱۴۰۸۶	۹۴۴	۲۹	۱۲۵۹۶۷	۱۹۶۱	۲۵۸	۲۱۰۶	۱۸۲۷۶	۱۱۱۷۷	۳۷۸۰۹	۲۰۸۲۱	۵۶۱۶۶
BagEnsDT	۱۹۷۵	۹۷۳	۰	۱۳۸۰۷۸	۴۶۷۳	۱۱۵۷	۱۲۰۷	۱۵۵۶۴	۳۰۸۷	۵۵۹۸۸	۲۶۴۲	۶۴۲۵۶
ADABOOSTC	۱۴۳۴	۹۷۲	۱	۱۳۸۶۱۹	۲۱۱۲	۱۲۵۲	۱۱۱۲	۱۸۱۲۵	۲۷۸۳	۵۴۹۹۷	۳۶۳۳	۶۴۵۶۰
BoostEnsDT	۱۲۸	۹۷۳	۰	۱۳۹۹۲۵	۲۹۵۳	۱۴۷۳	۸۹۱	۱۷۲۸۴	۶۵۲	۵۷۸۰۷	۸۲۳	۶۶۶۹۱
SVM	۰	۹۷۱	۲	۱۴۰۰۵۳	۲۰۳۹	۱۲۷۴	۱۰۹۰	۱۸۱۹۸	۷۷۵	۵۰۶۷۰	۷۹۶۰	۶۶۵۶۸
Min	۰	۱۵۸	۰	۱۱۰۰۰۰	۴۸	۱۹۶	۸۹۱	۱۵۵۶۴	۱۱۸	۳۷۸۰۹	۱۲۷	۴۳۴۳۶
Max	۱۴۰۸۶	۹۷۳	۸۱۵	۱۴۰۰۵۳	۴۶۷۳	۱۴۷۳	۲۱۶۸	۲۰۱۸۹	۲۳۹۰۷	۵۸۵۰۳	۲۰۸۲۱	۶۷۲۲۵

جدول (۹): خلاصه مقایسه نتایج به‌دست آمده روش پیشنهادی با نتایج روش‌های پیشنهادی سایر تحقیقات

نویسنده (ها)	سال	تکنیک (ها)	اجماع	مجموعه داده (ها)	پیش پرادزش	معیار(ها)	دقت (%)		
							KDD	DARPA	Real Traffic
Rezazadeh et al.	۲۰۱۶	Fisher, Quadratic, UDC, SDT, DT, NaiveB, Weak, SVM	AdaBoost, Boosting, Bagging, Weighted majority voting chosen by Proposed GA	DARPA, Real traffic, NSL-KDD	Data mining based	True and false positive rate, true and false negative rate, precision, ,recall F-measure, ROC curve, Accuracy, execution time and T-test	۹۹/۷۴	۹۹/۹۹	۹۲/۸۵
Aburomman and Ibne Reaz [۱۶]	۲۰۱۶	SVM, k-NN	Weighted majority voting chosen by PSO	KDD' 99	None	Classification accuracy	۹۲/۵۹	۹۲/۸۳ *	۸۶/۲۰ *
Singh et al. [۱۷]	۲۰۱۵	OS-ELM	Feature aggregation stage	NSL-KDD, Real traffic	Filtered, consistency, CFS subset evaluation and DBSCAN	Accuracy, true and false positive rate, true and false negative rate, F1-score, precision and execution time.	۶۶/۹۸	۹۱/۹۸*	۳۷/۹۶
Sreenath and Udhayan [۱۱]	۲۰۱۵	Naive bayes, Decision Stump, Hoeffding Tree, ADTree	Bagging	NSL-KDD	None	Accuracy	۹۷/۸۵	۹۸/۰۹*	۹۱/۰۸*
Gaikwad and Thool [۱۲]	۲۰۱۵	RepTree, Naive bayes, Random Tress, C4.5/J48, Decision Stump,	Bagging, AdaBoost	NSL-KDD	BIRCH hierarchical clustering	execution time, Accuracy, false positive rate	۸۱/۲۹	۸۱/۵۰*	۸۶/۵۷*
Masarat et al. [۱۵]	۲۰۱۴	J48 trees, fuzzy	Fuzzy Ensemble	KDD' 99	Gain Ratio Evaluation	Accuracy, Cost	۹۳/۰۰	۹۳/۲۳*	۸۶/۵۷*
Elbasiony et al. [۱۸]	۲۰۱۳	Weighted variant of K-means	Random Forest	KDD' 99	None	ROC curve	۹۸/۱۵	۹۸/۳۹*	۹۱/۳۶*
Aslahi-Shahri et al. [۲۸]	۲۰۱۵	SVM, GA	Hybrid machine learning	KDD' 99	None	True and false positive rate, ROC curve, F-measure, Recall, Accuracy	۹۷/۲۰	۹۷/۴۴*	۹۰/۴۸*
Rastegari et al. [۲۹]	۲۰۱۵	GA	None	NSL-KDD	Data mining based	True positive and true negative rate, accuracy	۹۸/۴۰۰۰	۹۸/۶۵*	۹۱/۶۰*
PROPOSED	-	GAWote	AdaBoost, Boosting, Bagging, Weighted majority voting chosen by Proposed GA	DARPA, Real traffic, NSL-KDD	Data mining based	True and false positive rate, true and false negative rate, precision, ,recall F-measure, ROC curve, Accuracy, execution time and T-test	۹۹/۷	۹۹/۹۹	۹۲/۸۵

۸- نتیجه‌گیری

در این تحقیق به پیاده‌سازی یک هانی‌نت آزمایشگاهی بر پایه مجازی‌سازی پرداخته شد تا بتوان بخشی از داده‌های مورد نیاز برای آزمایش الگوریتم پیشنهادی جمع‌آوری شود. در طول انجام کارهای تحقیقاتی هر دو هانی‌نت آزمایشگاهی طراحی‌شده در دو مکان متفاوت به‌دفعات مورد حمله قرار گرفت. در حقیقت ایجاد بستری برای تعامل با مهاجمین و تجزیه‌وتحلیل فعالیت‌های آن‌ها درک عمیقی به ما جهت شناخت حملات داده است و کمک می‌کند که بتوان الگوهای حملات را استخراج نموده و الگوریتم پیشنهادی را با کارایی و دقت بیشتری طراحی نمود. انتخاب فن‌های مناسب برای یادگیری از بین جریان‌های مخرب مهم‌ترین بخش از تحقیق می‌باشد. با توجه به نامتوازن بودن داده‌ها و هم‌چنین مشکلاتی که در زمینه وزن‌دهی به‌وجود می‌آمد، قابلیت شناسایی جریان‌های ناشناخته توسط مدل یادگیری جمعی بر اساس رأی‌گیری وزنی پیشنهادی مبتنی بر الگوریتم ژنتیک به‌عنوان راه‌حل مناسب و کارا با دقت بالا عمل نمود که یکی از بهترین انتخاب‌ها در این زمینه نسبت به روش‌های دیگر می‌باشد. به‌طور متوسط دقت الگوریتم پیشنهادی برای سه مجموعه داده با ویژگی‌های متفاوت و نامتوازن بودن داده‌ها ۹۷/۵۳٪ می‌باشد. نویسندگان این مقاله امیدوارند که نتایج تحقیقات به فرآیند افزایش امنیت و ارتقا فنون امنیتی جهت شناسایی حملات ناشناخته به‌طور مؤثر کمک نماید. اگرچه این تحقیق نشان می‌دهد که الگوریتم‌های جمعی می‌توانند یکی از بهترین و کارآترین روش‌ها برای شناسایی حملات ناشناخته باشند، ولی در مطالعات آتی می‌بایست بر روی افزایش دقت و همچنین کاهش زمان اجرا در طراحی این‌گونه الگوریتم‌ها تمرکز نمود.

۹- مراجع

[1] R. Talabis, "Honeynet learning: discovering IT security," SIGCSE Bull, vol. 38, pp. 110-114, 2006.

[2] S. Hido and H. Kashima, "Roughly balanced bagging for imbalanced data," Statistical Analysis and Data Mining: the ASA Data Science, vol. 5, pp. 143-152, 2009.

[3] P. Kang and S. Cho, "Ensemble of under-sampled SVMs for data imbalance problems," Lecture Notes in Computer Science, vol. 4232, pp. 837–846, 2006.

[4] I. Corona, G. Giacinto, C. Mazzariello, F. Roli, and C. Sansone, "Information fusion for computer security: state of the art and open issues," Information Fusion, vol. 10, no. 4, pp. 274–284, 2009.

[5] J. Z. Kolter and M. A. Maloof, "Learning to detect and classify malicious executables in the wild," Journal of Machine Learning Research, vol. 7, pp. 2721–2744, 2006.

[6] M. G. Schultz, E. Eskin, E. Zadok, and S. J. Stolfo, "Data mining methods for detection of new malicious executables," In Proceedings of the IEEE Symposiumon Security and Privacy, pp. 38–49, 2001.

[7] G. Giacinto, R. Perdisci, M. D. Rio, and F. Roli, "Intrusion detection in computer networks by a modular ensemble of one-class classifiers," Information Fusion, vol. 9, no. 1, pp. 69–82, 2008.

[8] G. Giacinto, F. Roli, and L. Didaci, "Fusion of multiple classifiers for intrusion detection in computer networks," Pattern Recognition Letters, vol. 24, no. 12, pp. 1795–1803, 2003.

[9] B. Zhang, J. Yin, S. Wang, and X. Yan, "Research on virus detection technique based on ensemble neural network and SVM," Advanced Intelligent Computing Theories and Methodologies, vol. 137, pp. 24-33, 2014.

[10] S. M. AbdElrahman and A. Abraham, "Intrusion detection using error correcting output code based ensemble," International Conference Hybrid Intelligent Systems (HIS), IEEE, pp. 181-186, 2014.

[11] M. Sreenath and J. Udhayan, "Intrusion detection system using Bagging Ensemble Selection," International Conference Engineering and Technology (ICETECH), pp. 1-4, 2015.

[12] D. P. Gaikwad and R. C. Thool, "Intrusion detection System using bagging ensemble method of machine learning," International Conference Computing Communication Control and Automation (ICCUBEA), pp. 291-295, 2015.

[13] P. Sornsuwit and S. Jaiyen, "Intrusion detection model based on ensemble learning for U2R and R2L attacks," International Conference on Information Technology and Electrical Engineering (ICITEE), pp. 354-359, 2015.

[14] S. Masarat, H. Taheri, and S. Sharifian, "A novel framework, based on fuzzy ensemble of classifiers for intrusion detection systems," International Conference Computer and Knowledge Engineering (ICCKE), pp. 165-170, 2014.

[15] M. Milliken, Y. Bi, L. Galway, and G. Hawe, "Ensemble learning utilising feature pairings for intrusion detection," World Congress on Internet Security (WorldCIS), pp. 24-31, 2015.

[16] A. A. Aburomman and M. B. IbneReaz, "A novel SVM-kNN-PSO ensemble method for intrusion detection system," Applied Soft Computing, vol. 38, pp. 360–372, 2016.

[17] R. Singha, H. Kumarb, and R. K. Singlac, "An intrusion detection system using network traffic profiling and online sequential extreme learning machine," Expert Systems with Applications, vol. 42, pp. 8609–8624, 2015.

[18] R. M. Elbasiony, E. A. Sallam, T. E. Eltobely, and M. M. Fahmy, "A hybrid network intrusion detection framework based on random forests and weighted k-means," Ain Shams Eng. J, vol. 4, pp. 753–762, 2013.

[19] D. Watson and J. Riden, "The honeynet project," Technical Report, 2006.

[20] Members of the Honeynet Project, "Know Your Enemy: Learning about Security Threats," 2nd edn. Addison-Wesley, Boston, 2004.

[21] N. Provos and T. Holz, "Virtual honeypots: from botnet tracking to intrusion detection," 1st edn. Addison-Wesley Professional, Boston, 2007.

[22] L. Rokach, "Ensemble-based classifiers,' Artificial Intelligence Review, vol. 33, pp. 1–39, 2010.

[23] R. Polikar, "Ensemble based systems in decision making," Circuits and Systems Magazine, vol. 6, pp. 21–45, 2006.

[24] D. Opitz and R. Maclin, "Popular ensemble methods: An empirical study," Journal of Artificial Intelligence Research, vol. 11, pp. 169–198, 1999.

[25] L. K. Hansen and P. Salamon, "Neural network ensembles," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 10, pp. 993–1001, 1990.

[26] R. E. Schapire, "The strength of weak learnability," Machine Learning, vol. 5, no. 2, pp. 197–227, 1990.

[27] Z. H. Zhou, "Ensemble methods: foundations and algorithms. machine learning & pattern recognition series," Chapman & Hall/CRC, Boca Raton FL, 2012.

[28] B. M. Aslahi-Shahri, R. Rahmani, M. Chizari, A. Maralani, M. Eslami, M.J. Golkar, and A. Ebrahimi, "A hybrid method consisting of GA and SVM for intrusion detection system," Neural Computing and Applications, pp. 1–8, 2015.

[29] S. Rastegari, P. Hingston, and C. P. Lam, "Evolving statistical rulesets for network intrusion detection," Applied Soft Computing, vol. 33, pp. 348–359, 2015.

[30] R. Ghorbani and H. Abrishami, "Using Stereo Vision to Provide a Vision-Based Augmented Reality System," Tabriz Journal of Electrical Eng., vol. 42, no. 1, 2012. (In Persian)

[31] S. Abdollahzadeh, M. A. Balafar, and L. Mohammad Khanli, "|Using Clustering and Markov Model in Predicting Web Users' Next Request," Tabriz Journal of Electrical Eng., vol. 45, no. 3, 2014. (In Persian)

[32] F. H. Abbasi, R. J. Harris, S. Marsland, and G. Moretti, "An Exemplar-Based Learning Approach for Detection and Classification of Malicious Network Streams in Honeynets," Security and Communication Networks, vol. 7, no. 2, pp. 352-364, 2014.

[33] H. Parvin, S. Ansari, and S. Parvin, "Proposing a New Method for Non-Relative Imbalanced Dataset," Soft Computing Models in Industrial and Environmental Applications, vol. 188, pp. 297-306, 2013.

[34] H. Parvin, B. Minaei, H. Alinejad-Rokny, and W. Punch, "Data weighing mechanisms for clustering ensembles," Computers and Electrical Engineering, vol. 5. no. 39, pp. 1433–1450, 2013.

[35] R. K. Shahzad and N. Lavesson, "Comparative Analysis of Voting Schemes for Ensemble-based Malware Detection," Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, vol. 4, no. 1, pp. 98-117, 2013.

[36] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," IEEE Communications Surveys and Tutorials, vol. 18, no. 2, pp. 1153–1176, 2016.

[37] A. A. Aburomman and M. B. Ibne Reaz, "A novel SVM-Knnpso ensemble method for intrusion detection system," Applied Soft Computing Journal, vol. 38, pp. 360–372, 2016.

[38] M. Li, S. Pan, Y. Zhang, and X. Cai, "Classifying networked text data with positive and unlabeled examples," Pattern Recognition Letters, vol. 77, pp. 1–7, 2016.

[39] S. Parsa and M. Zeinipour, "Botnet Detection with Flow Behavior Analysis Approach," Journal of Electrical & Cyber Defence, vol. 5, no. 4, pp. 35-50, 2017. (In Persian)

Journal of Electronical & Cyber Defence
Vol. 6, No. 2, 2018, Serial No. 22

Detection of Unknown Malicious Network Streams using Ensemble Learning

H. Parvin, V. Rezaei, S. Nejatian*, R. Omidvar, M. Yasrebi

*Mamasani Branch, Islamic Azad University

(Received: 15/07/2017, Accepted: 23/09/2017)

ABSTRACT

Security is a significant issue in this world and is given several dimensions by varying circumstances. Among different security areas, cyber security can be claimed to have one of the most important places in new circumstances of this world. In this study, two virtual honeynets were designed in two different laboratories to help us study unknown attacks. Other scientific datasets were also used for this purpose. Imbalanced data always cause problems for network datasets and reduce the efficiency for the prediction of minority classes. To cope with this problem, ensemble learning methods were applied in order to detect network attacks and most specifically, unknown attacks, while taking advantage of different techniques and action model learning. It was found that ensemble learning method was suitable for describing the security problems because activities done on computer systems can be viewed at multiple levels of abstraction and information can be collected from multiple data sources. Statistical analysis was used as the research method in order to measure the reliability and validity of findings. Here, we applied statistical techniques and tests to show that the algorithm designed by the proposed weighted voting and based on the genetic algorithm has a better performance than other twelve classifiers.

Keywords: Honeynet, Unknown Attacks, Ensemble Learning, Imbalanced Data, Weighted Voting, Statistical Tests

* Corresponding Author Email: nejatian@iauyasooj.ac.ir