Scientific Journal of Electronical & Cyber Defence
Vol. 10, No. 2, Summer 2022, Serial No. 38

A New Hybrid Approach for Traffic Identification and Classification in Wireless Networks

M. Bazooband, H. Bahramghiri*

* Assistant Professor, Electrical and Computer Faculty, Malik Ashtar University of Technology, Tehran, Iran

(Received: 19/04/2021, Accepted: 29/09/2021)

ABSTRACT

Using the ad hoc approach is one of the desirable options for configuration of wireless networks because of the features such as distributed management between nodes, facilitating their entry and exit into the network and the possibility of better mobility. This scheme leads to the dynamic behavior of the traffic generated by applications in such networks, which affects the issue of network management and traffic control between nodes. Identifying and classifying the network traffic can help to deal with these challenges in wireless networks. Because conventional traffic detection and classification methods are not able to provide proper performance with such traffic, applied machine-learning-based methods can improve the detection and classification performance. As the precision required to find a specific network traffic implies a high probability of detection, and the elimination of wrong decisions needs the false alarm rate reduction, in this paper a new hybrid method, based on the combination of machine learning methods is introduced to increase the accuracy and efficiency of identifying and classifying traffic in ad hoc wireless networks based on purposeful combination of various machine learning methods. The results show that in addition to improving the evaluation criteria of traffic classification, the proposed method increases the detection probability and reduces the false alarm rate, in comparison to the cases where only a single machine learning method is used .

Keywords: Wireless Ad hoc Networks, Network Traffic, Probability of Detection, False Alarm Rate, Hybrid Approach in Machine Learning.

* Corresponding Author Email: Bahramgiri@mut.ac.ir

This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Publisher: Imam Hussein University © Authors

نشریه علمی "پدافند الکترونیکی و سایبری"
سال دهم، شماره ۲، تابستان ۱۴۰۱، ص ۴۱-۳۱

علمی-پژوهشی

ارائه روش ترکیبی برای شناسایی و طبقه‌بندی ترافیک در شبکه‌های بی‌سیم

مریم بازوبند[۱]، حسین بهرامگیری[۲*]

۱- کارشناسی ارشد مخابرات سیستم، ۲- استادیار، دانشکده برق و کامپیوتر، دانشگاه صنعتی مالک اشتر، تهران، ایران

(دریافت: ۱۴۰۰/۰۱/۳۰، پذیرش: ۱۴۰۰/۰۷/۰۷)

چکیده

استفاده از رویکرد اقتضایی با بهره‌گیری از ویژگی‌هایی از جمله مدیریت توزیع یافته بین گره‌ها، تسهیل در امر ورود و خروج آن‌ها به شبکه و امکان تحرک بهتر، یکی از گزینه‌های مطلوب جهت پیکربندی شبکه‌های بی‌سیم می‌باشد. همین امر موجب تولید ترافیک با رفتار پوی توسط نرم‌افزارهای کاربردی در چنین شبکه‌هایی می‌شود که مسئله مدیریت شبکه و کنترل ترافیک بین گره‌های را تحت تأثیر خود قرار می‌دهد. شناسایی و طبقه‌بندی ترافیک جاری در شبکه می‌تواند کمک شایانی به این چالش در شبکه‌های بی‌سیم کند. از آنجا که روش‌های مرسوم شناسایی و طبقه‌بندی ترافیک قادر به ارائه عملکرد مناسب با چنین ترافیک‌هایی نیستند بنابراین استفاده از روش‌های مبتنی بر یادگیری ماشین می‌تواند برای بهبود طبقه‌بندی ترافیک به‌کارگرفته شوند. از آنجا که حساسیت بالا جهت یافتن ترافیک‌هایی خاص نیازمند افزایش احتمال آشکارسازی و عدم ارائه تصمیم اشتباه نیازمند کاهش هشدار غلط در سامانه می‌باشد،بنابراین در این مقاله روشی جدید جهت افزایش دقت و بهره‌وری در شناسایی وطبقه‌بندی ترافیک در شبکه‌های بی‌سیم اقتضایی ارائه می‌شود که مبتنی بر ترکیب هدفمند روش‌های یادگیری ماشین می‌باشد. نتایج نشان می‌دهند که روش ارائه شده علاوه بر بهبود معیارهای ارزیابی طبقه‌بندی کننده ترافیک موجب افزایش احتمال آشکارسازی و کاهش نرخ هشدار غلط به نسبت به‌کارگیری روش‌های یادگیری ماشین به‌صورت یکتا می‌باشد.

کلیدواژه‌ها: شبکه‌های اقتضایی بی‌سیم، ترافیک شبکه، احتمال آشکارسازی، نرخ هشدار غلط، رویکرد ترکیبی در یادگیری ماشین

۱- مقدمه

شبکه‌های مخابراتی بی‌سیم همواره یکی از گزینه‌های مطلوب جهت شبکه‌سازی در حوزه‌های تجاری و نظامی بوده‌اند. از این رو جهت تحرک مؤثر گره‌ها، سازوکارهای ورود و خروج به شبکه و مدیریت غیر متمرکز شبکه که موجب افزایش انعطاف عملیاتی در شبکه می‌شوند از رویکرد پیکربندی اقتضایی استفاده می‌شود. مسئله مدیریت منابع حیاتی در شبکه همچون پهنای باند، منابع اطلاعاتی و پردازشی و کنترل دسترسی گره‌ها به چنین منابعی اهمیت فراوانی دارد. از این رو شناخت و تسلط بر ترافیک جاری بر بستر شبکه کمک شایانی به مدیریت و کنترل آن می‌کند. از گذشته روش‌های اصلی شناسایی و طبقه‌بندی ترافیک مبتنی بر شناسایی شماره پورت، تحلیل بار ترافیکی[۱] بسته‌ها و تحلیل آماری بوده است [۱]. روش شناسایی شماره پورت، مبتنی بر یافتن شماره پورت مبدأ و مقصد بسته‌های ترافیکی و دسته‌بندی بر اساس آن‌ها بود. از این رو با توجه به امکان استفاده از پورت-های مجازی و پویا توسط نرم‌افزارهای جدید این روش دارای دقت کافی نیست [۱]. در روش تحلیل بار ترافیکی با استناد به الگوی بیت‌های اطلاعاتی در قسمت بدنه بار ترافیکی بسته‌های در حال تبادل و مقایسه آن با الگوی ذخیره شده از هر نوع ترافیک در یک پایگاه داده، ابتدا آن ترافیک شناسایی و سپس در یکی از دسته‌های مشابه طبقه‌بندی می‌شود. اما این روش نیازمند به افزایش قدرت پردازشی و ظرفیت ذخیره‌سازی و به‌روزرسانی دائمی پایگاه داده آن از الگوی ترافیکی انواع پروتکل‌ها و نرم‌افزار-ها می‌باشد. از طرفی با رمزنگاری بار ترافیکی نیز استفاده از این روش در عمل کارایی نخواهد داشت [۲]. تحلیل کمیت‌های آماری همچون میانگین و انحراف معیار مرتبط با ویژگی‌های ترافیکی بسته‌ها فارغ از شماره پورت و یا تحلیل الگوریتم بدنه ارزشمند بار ترافیکی، توانایی کار بهتری در مواجه با ترافیک‌های پویا دارند. از همین رو روش‌های مبتنی بر یادگیری ماشین در طی سال‌های اخیر توانستند عملکرد مؤثرتری را در شناسایی و طبقه‌بندی ترافیک شبکه نشان دهند. در این روش‌ها، مبتنی بر مشاهدات پیشین از ترافیک جاری در شبکه به نام مجموعه داده که شامل نمونه‌هایی از ترافیک و نوع آن‌ها به نام برچسب است، با

* رایانامه نویسنده مسئول: Bahramgiri@mut.ac.ir

[1] Payload

* این مقاله یک مقاله با دسترسی آزاد است که تحت شرایط و ضوابط مجوز Creative Commons Attribution (CC BY) توزیع شده است.
© نویسندگان
ناشر: دانشگاه جامع امام حسین (ع)

آموزش سامانه، مدلی ریاضی حاصل می‌شود تا در زمان اعمال به شبکه و اخذ نمونه از ترافیک جاری در آن، با استناد به نتایج آن مدل ترافیک مورد نظر طبقه‌بندی شود. روش‌های مبتنی بر یادگیری ماشین به دو دسته کلی نظارتی[1] و غیر نظارتی[2] تقسیم می‌شوند. در شیوه نظارتی مجموعه داده اعمال شده به سامانه دارای برچسب می‌باشد. یعنی مقادیر ورودی به‌عنوان ویژگی‌ها و نتیجه خروجی به‌عنوان برچسب تحت یک مجموعه داده به الگوریتم اعمال می‌شود. اما در روش‌های غیر نظارتی، داده‌ها برچسب ندارند و باید از روابط میان آن‌ها برای ساخت مدلی مناسب استفاده کرد.

تحقیقات متعددی در زمینه طبقه‌بندی و شناسایی ترافیک شبکه‌های سیمی و بی‌سیم صورت گرفته است. در مرجع [۱] مقایسه‌ای بین روش غیر نظارتی خوشه‌بندی که الگوریتم آن بر اساس بیشترین شباهت است، با روش دسته‌بندی کننده Naïve Bayes ارائه شده است. روش خوشه‌بندی دقت ۹۱ درصدی داشته و حدود ۹ درصد از روش Naïve Bayes بهتر عمل کرده است. روش‌های مبتنی بر یادگیری ماشین توانایی بسیار بهتری در شناسایی و طبقه‌بندی ترافیک‌های ناشناخته دارند [۱]. ترافیک اینترنت به‌صورت زمان حقیقی در حال تولید و جابه‌جایی است و نیازمند روشی با سرعت مناسب برای شناسایی و طبقه-بندی آن است. این مسئله، در مرجع [۲] با درخت تصمیم (DT[3]) که روشی نظارتی می‌باشد بررسی شده است. الگوریتم C4.5 در این کار مورد استفاده قرار گرفته و پاسخ مناسبی داشته است [۲]. در کار تحقیقاتی [۳] از طبقه‌بندی ترافیک به‌عنوان یک مسئله مهم در امنیت و مدیریت شبکه یاد می‌شود. روابط ریاضی برای افزایش بهره‌وری روش‌های نظارتی به ویژه در زمانی که مجموعه داده محدود می‌باشد مورد بررسی قرار گرفته است. در کار تحقیقاتی [۴] پس از بررسی دلایل نیاز به شناسایی و طبقه‌بندی ترافیک با دیدگاهی آموزشی، روش‌های یادگیری ماشین در این زمینه بررسی، تحلیل و ذکر شده است که در مباحث اندازه‌گیری معیارهای کیفیت سرویس شبکه نیز، مبحث طبقه‌بندی ترافیک بسیار کمک کننده است. در مرجع [۵]، کاربرد شبکه عصبی امتزاجی (CNN[4]) در شناسایی ترافیک‌های ناشناس در شبکه و سپس طبقه‌بندی سریع و با دقت آن‌ها بررسی شده که دقت ۸۶/۰۵ درصدی در این کار حاصل شده است. کار تحقیقاتی [۶] توسعه شبکه‌ها و کاربرد وسیع اینترنت را از مهم‌ترین دلایل افزایش پیچیدگی ترافیک دانسته است که مدیریت آن را بسیار سخت کرده و بر اساس آن روش‌های قدیمی همچون روش‌های مبتنی بر پورت دیگر کارایی ندارند [۶]. در این کار تحقیقاتی روشی بر مبنای شناسایی الگوی سرآیند پروتکل‌ها ارائه شده است که حتی با وجود رمزنگاری در سایر بخش‌های بسته، سرآیندها رمز نمی‌شوند. این روال معمولاً در شبکه‌های محلی کوچک، قابل اعتماد و دارای کارایی مناسبی است. در کار تحقیقاتی دیگری ضمن بیان اهمیت شناسایی و طبقه‌بندی ترافیک در مدیریت شبکه، روش‌های یادگیری ماشین بردار پشتیبان[5] و نزدیک‌ترین همسایه[6] با دو رویکرد متفاوت مقایسه شده‌اند [۷]. با رویکرد بهره‌وری در مدیریت شبکه‌های مبتنی بر IP، در کار تحقیقاتی [۸]، نقش شناسایی و طبقه‌بندی ترافیک بررسی و روش‌های مرسوم و نوین همچون انواع روش‌های مبتنی بر یادگیری ماشین مرور شده است. تهیه مجموعه داده، منظم-سازی آن و انتخاب ویژگی‌ها بر حسب نوع نرم‌افزارها، می‌تواند در بهره‌وری روش‌های یادگیری ماشین اثرگذار باشد [۸]. شبکه عصبی احتمالاتی[7] یک روش جدید در مباحث ریاضی توابع محاسباتی شبکه عصبی است که در مرجع [۹] معرفی شده است. تنها اندازه بسته‌ها به همراه چهار مشخصه میانگین، انحراف معیار، کمینه و بیشینه آن‌ها به‌عنوان ویژگی انتخاب شده‌اند. با وجود همین تعداد کم ویژگی، اما حضور شبکه عصبی باعث افزایش دقت در مقایسه با سایر روش‌ها شده است. مسئله یادگیری عمیق و کاربرد آن در شناسایی ترافیک، شناسایی پروتکل و شناسایی حمالات سایبری بررسی شده است [۱۰]. یادگیری عمیق نسخه پیچیده‌تر و سنگین‌تری از شبکه عصبی است که قدرت پردازشی بسیار بالاتر برای کار با مجموعه داده‌های بزرگ را ارائه می‌دهد [۱۱]. با بهره‌گیری از رویکرد غیر نظارتی، کارایی خوشه‌بندی در مقابل روش‌های نظارتی بررسی شده است که سرعت یادگیری این روش در زمانی که مجموعه داده بسیار حجیم است و ترکیب نمونه‌ها دارای رفتار غیر خطی عجیبی است، می‌تواند به بهبود عملکرد امر شناسایی و طبقه‌بندی ترافیک کمک کند [۱۲]. رادیو نرم‌افزار[8] ویژگی بسیار نوینی در پیکربندی ارتباطات بی‌سیم است که به وسیله آن می‌توان در هر لحظه موارد مورد درخواست کاربر را به شبکه اعمال نمود. در کار تحقیقاتی دیگری، در شبکه‌های حسگر بی‌سیم که تعداد گره‌ها و حجم ترافیک زیاد است، از روش غیر نظارتی خوشه بندی استفاده شده است [۱۳]. با تمرکز بر روی شبکه‌های مبتنی بر IP، در کار تحقیقاتی [۱۴] روش نزدیک‌ترین همسایه انتخاب شده و به بررسی اثر انتخاب ویژگی

[1] Supervised
[2] Unsupervised
[3] Decision Tree (DT)
[4] Convolutional Neural Network (CNN)
[5] Supported Vector Machine (SVM)
[6] Nearest Neighbor
[7] Probabilistic Neural Network (PNN)
[8] Software Defined Radio

در نتایج پردازش پرداخته شده است و نتایج نشانگر آن هستند که با انتخاب‌های گوناگون، دقت تا ۹۰ درصد افزایش یافته است. در حوزه شبکه‌های بی‌سیم با توپولوژی مش، بررسی اثر روش درخت تصمیم با الگوریتم C4.5 مورد ارزیابی قرار گرفته است و سرعت و دقت در این کار در کنار ویژگی‌های انتخاب شده برای مجموعه داده، به‌عنوان پارامترهایی اثرگذار در انتخاب روش مبتنی بر یادگیری ماشین ذکر شده‌اند [۱۵]. روش شبکه عصبی امتزاجی سه بعدی نیز در کار دیگری بررسی شده است که این روش در شبکه‌های بی‌سیم با تبدیل هر جریان ترافیکی به تصویری سه بعدی، آن داده را به شبکه CNN اعمال و نسبت به شناسایی و طبقه‌بندی ترافیک اقدام می‌کند [۱۶]. برای شناسایی ترافیک از دیدگاه دسته‌بندی بدافزارها و یا حملات سایبری، از شبکه‌های CNN مورد استفاده قرار گرفته است [۱۷] که نکته مهم آن است که ترافیک به این شبکه‌ها اعمال شده و خود شبکه عصبی CNN نسبت به استخراج ویژگی با بهره بالا و شناسایی و طبقه‌بندی ترافیک اقدام می‌کند. ترکیب روش خوشه‌بندی با انتشار برچسب، موضوع تحقیقاتی دیگری بود که بر مبنای روشی نیمه نظارتی انجام شده است که در آن هم از خوشه‌بندی و هم از خصوصیات برچسب انتشار یافته بین نمونه‌ها بهره برده شده است [۱۸].

در این مقاله، مسئله شناسایی و طبقه‌بندی ترافیک در سامانه‌ای که قابلیت اجرای چهار الگوریتم KNN، SVM، Naïve Bayes و DT دارد مورد توجه قرار داده شده و به ارائه روش ترکیبی جهت افزایش احتمال آشکارسازی (PD[1]) و کاهش نرخ هشدار غلط (FAR[2]) برچسب‌های ترافیکی مورد نظر پرداخته شده است. در ادامه، کارایی این روش بر اساس مجموعه داده USTC-TFC2016 و مقایسه آن با چهار الگوریتم مذکور ارائه شده و کارایی آن را با شبیه سازی، بررسی شده است. هر کدام از این دو روش ترکیبی بنا بر نیاز می‌توانند در سامانه‌های تشخیص مورد استفاده قرار گیرند.

ساختار ادامه این مقاله بدین ترتیب است: در فصل دوم ضمن بیان مسئله با تعریف ترافیک، مروری بر روش‌های اساسی مبتنی بر یادگیری ماشین خواهد داشت. در فصل سوم الگوریتم ارائه شده در این مقاله با رویکرد ترکیبی بیان شده است و در فصل چهارم ضمن معرفی معیارهای ارزیابی به تحلیل نتایج خواهد پرداخت. نکات ارزشمند و جمع‌بندی کار نیز در فصل پنجم ارائه خواهد شد.

[1] Probability of Detection (PD)
[2] False Alarm Rate (FAR)

۲- روش‌های مورد نظر مبتنی بر یادگیری ماشین

در این بخش ضمن تعریف جریان ترافیکی در شبکه مروری بر روش‌های مورد استفاده مبتنی بر یادگیری ماشین که در این مقاله مورد استفاده قرارگرفته‌اند، خواهد داشت.

۲-۱- ترافیک در شبکه

یک شبکه مخابراتی وظیفه ایجاد بستری در جهت ارسال داده-های ارزشمند تولیدی کاربران از نقطه‌ای به نقطه یا نقاط دیگر را بر عهده دارد. این داده‌های ارزشمند که به‌صورت بیت‌های اطلاعاتی در ساختار پروتکل‌های ارتباطی تهیه می‌شوند را زمانی می‌توان به‌عنوان یک جریان ترافیکی برشمرد که پنج ویژگی زیر در آن مشخص باشد [۲]:

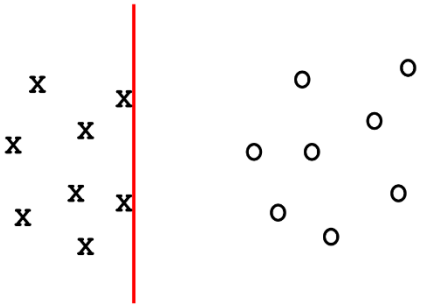
۱. شماره پورت مبدأ
۲. شماره پورت مقصد
۳. آدرس (IP[3]) مبدأ
۴. آدرس (IP) مقصد
۵. پروتکل لایه کاربرد

بنابراین می‌توان هر جریان ترافیکی را شامل بیت‌های اطلاعاتی تولید شده توسط نرم‌افزار لایه کاربرد به همراه اطلاعات سرآیند[4] لایه چهار و سه دانست که خود می‌توانند شامل تعدادی بسته[5] باشند. بنابراین می‌بایست برای تهیه مجموعه داده جهت اجرای الگوریتم‌های یادگیری ماشین از ویژگی‌های مبتنی بر اطلاعات بسته‌های در حال تبادل از مبدأ مشخص به مقصد یا مقاصد مشخص بهره برد. اجرای هر روش از یادگیری ماشین را می‌توان بر سه فاز تقسیم کرد که به ترتیب فاز آموزش، ارزیابی و آزمایش می‌باشند. در هر فاز بسته به اهمیت آن‌ها، تعدادی نمونه از محیط مورد نظر که برای ما ترافیک یک شبکه بی‌سیم است به سامانه اعمال می‌شود. در فاز آموزش با اعمال نمونه‌های آموزشی مدل ریاضی تشکیل می‌شود و با بررسی تمامی نمونه‌ها، سعی بر کاهش خطا در قبال تشخیص برچسب نمونه‌ها است. در فاز ارزیابی با اعمال نمونه‌هایی جدید، میزان دقت مدل به‌دست آمده بررسی شده و اساساً امکان استفاده، رد شدن یا اصلاح مجدد آن روش بررسی می‌شود. دقت حیاتی برای طراحان سامانه، عملکرد مدل به‌دست آمده در فاز آزمایش می‌باشد که می‌بایست تا حد امکان بالا باشد.

[3] Internet Protocol (IP)
[4] Header
[5] Packet

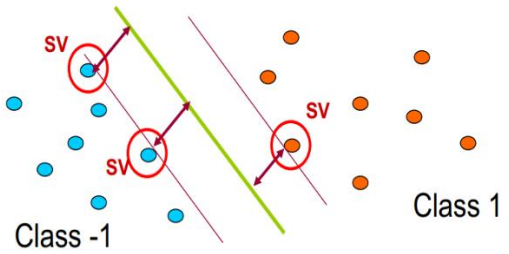
۲-۲- روش ماشین بردار پشتیبان (SVM)

ساده‌ترین مدل دسته‌بندی باینری با توجه به یک مجموعه داده را می‌توان پرسپترون دانست که هر خطی را که بتواند میان دو دسته داده مرز تصمیم‌گیری ایجاد کند را به‌عنوان مدل نهایی ارائه می‌کند. اما این خط می‌تواند به یک مجموعه داده بسیار نزدیک و از دسته داده دیگر دور باشد و این باعث عدم عادلانه بودن نواحی تصمیم‌گیری می‌شود که اثر خود را در فاز آزمایش با جابه‌جایی نمونه‌های جدید اطراف نمونه‌های آزمایشی و ایجاد خطا در نتیجه خروجی می‌گذارد. شکل (۱) نمونه‌ای از مدل خطی پرسپترون را نمایش می‌دهد [۱۸].



شکل (۱): جداسازی دو دسته داده بر اساس قاعده پرسپترون

با فرض جداپذیر بودن خطی مجموعه داده، ماشین بردار پشتیبان با یافتن نمونه‌های حاشیه‌ای هر دسته داده و یافتن خطی که به‌صورت عادلانه میان آن‌ها حداکثر حاشیه را ایجاد کند، بهترین خط جدا کننده را ارائه می‌دهد. شکل (۲) نمونه‌ای از خط بهینه با حداکثر حاشیه را که توسط ماشین بردار پشتیبان به‌دست آمده است نشان می‌دهد [۱۸].



شکل (۲): یافتن خط بهینه بر اساس نمونه‌های بردار پشتیبان

در صورتی که نمونه‌های مجموعه داده جداپذیر خطی نباشند با استفاده از توابع کرنل مدل ساده خطی به فضای چند بعدی منتقل شده و با یافتن یک ابر صفحه نسبت به جداسازی دسته‌های داده اقدام می‌کند. در واقع تابع کرنل ضرایب داخلی نمونه‌های آن مجموعه داده در فضای چند بعدی را می‌یابد و با تعریف یک تابع به‌طور مستقیم ابرصفحه در آن فضا را به‌عنوان یک مدل برای جداسازی به ما ارائه می‌کند. به‌عنوان مثال تابع کرنل چند جمله‌ای را با رابطه (۱) می‌توان بیان کرد [۱۸].

$$k(x_i, x_j) = (x_i.x_j + 1)^d \qquad (۱)$$

دقت شود که پارامترهای درونی کرنل‌ها از نوع ابر پارامتر هستند که توسط فرد خبره تعیین می‌شوند. باید به این نکته نیز اشاره شود که تعیین نوع کرنل هم یک ابر پارامتر است. چرا که با انتخاب نامناسب کرنل هم بار محاسباتی افزایش می‌یابد و هم اینکه در یک کاربرد خاص، آن تابع کرنل به خوبی عمل نمی‌کند.

۲-۳- روش نزدیک‌ترین همسایه (KNN)

یکی از روش‌های اجرای الگوریتم‌های یادگیری ماشین، استفاده از روش‌های مبتنی بر یادگیری نمونه یا (IBL[1]) است [۱۳]. در این سازوکار در واقع مدلی بر مبنای تمام نمونه‌ها به‌دست نخواهد آمد تا پس از یادگیری در فاز آزمایش از آن استفاده شود. بلکه در هر بار بررسی نمونه‌های جدید، همان لحظه با نمونه‌های داده آموزشی که هیچ پردازشی روی آن‌ها صورت نگرفته (در فاز آموزش کاری انجام نشده) مقایسه شده است و با شباهت با آن نمونه‌ها تصمیم‌گیری برای تعیین برچسب نمونه جدید انجام می‌شود. بنابراین به پایگاه داده نسبتاً بزرگی منطبق با حجم داده‌های آموزشی نیاز دارید که نمونه‌ها را در آن ذخیره شود. روش K-NN[2] ساده‌ترین و متداول‌ترین روش بر پایه یادگیری نمونه است. در این روش فرض می‌شود که تمام نمونه‌ها نقاطی در فضای n بعدی حقیقی هستند و همسایه‌ها بر مبنای فواصل اقلیدسی استاندارد تعیین می‌شوند. منظور از k، تعداد همسایه‌های درنظر گرفته شده است. اگر یک مثال دلخواه را به‌صورت یک بردار ویژگی به‌صورت $< a_1(x), a_2(x), ..., a_n(x) >$ نمایش داده شود، فاصله اقلیدسی بین دو نمونه به‌صورت x_i و x_j رابطه (۲) تعریف می‌شود [۱، ۱۶ و-۱۸].

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2} \qquad (۲)$$

برای یک تابع هدف گسسته f، الگوریتم K-NN به‌صورت رابطه (۳) است [۱، ۱۶ و-۱۸].

$$f : \Re^n \rightarrow V \text{ , where V is the finite set } \{v_1, ..., v_s\} \qquad (۳)$$

یک لیست به نام لیست آموزش[3] برای نمونه‌های آموزشی ایجاد می‌شود که هر مثال آموزشی $< x, f(x) >$ به لیست

[1] Instance Base Learning (IBL)
[2] K- Nearest Neighbor (KNN)
[3] Training List

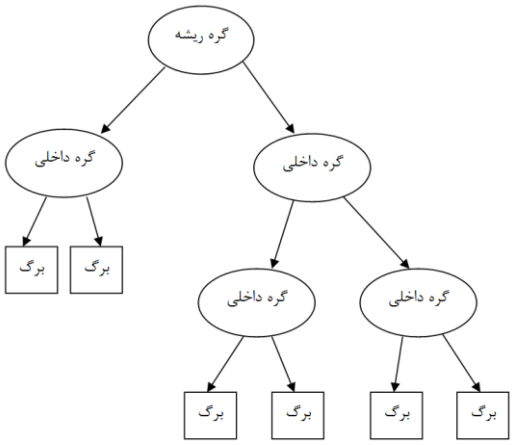
آموزش اضافه می‌شود. الگوریتم دسته‌بندی نیز در فاز آزمایش به این صورت خواهد بود که برای نمونه مورد بررسی x_q به تعداد K نزدیک‌ترین نمونه از نمونه‌های آموزشی همسایه به آن را با $x_1,...,x_k$ نمایش داده می‌شود. سپس مقدار رابطه (۴) را محاسبه نموده و برمی‌گردانید [۱۸].

$$\hat{f}(x_q) \leftarrow \arg\max_{v \in V} \sum_{i=1}^{k} \delta(v, f(x_i)) \quad (۴)$$

Where: $\delta(a,b) = \begin{cases} 1 & a = b \\ 0 & otherwise \end{cases}$

۲-۴- روش درخت تصمیم (DT)

درخت تصمیم یک روش مبتنی بر یادگیری ماشین است که مناسب کار با مجموعه داده‌هایی است که حاوی زوج ویژگی مقدار[1] هستند. یعنی مقادیر ویژگی‌ها تنها عدد نیست و از پارامترهای توصیفی نیز استفاده می‌شود. بنابراین این روش قابل استفاده در تحلیل‌های توصیفی همانند ساختار else و if می‌باشد. شکل (۳) ساختار روش درخت تصمیم را نشان می‌دهد. ویژگی‌ها در گره‌های میانی و برچسب‌ها در گره‌های انتهایی به نام برگ جای می‌گیرند. هر ویژگی بر اساس موارد توصیفی یا عددی خود، نمونه‌های مربوط به خود را دسته بندی کرده و این گام رو به جلو تا رسیدن به نمونه‌های هم برچسب و ختم شدن به برگ پیش می‌رود. برای تشکیل درخت تصمیم الگوریتم‌های متنوعی همچون ID3 و C4.5 وجود دارد [۲ و ۱۴].



شکل (۳): ساختار روش درخت تصمیم

معیار انتخاب ویژگی مناسب برای قرارگیری در ریشه، داشتن کمترین ابهام منطبق بر قاعده آنتروپی برای همان ویژگی و سپس محاسبه بهره اطلاعاتی منجر از قرارگیری آن ویژگی در گره ریشه و سایر ویژگی‌ها در گره‌های میانی می‌باشد. رابطه (۵ و ۶) به ترتیب آنتروپی و بهره اطلاعاتی را برای هر ویژگی نشان می‌دهند [۲، ۱۴ و ۱۸].

$$Entropy \quad E = -P \times \log_2 P \quad (۵)$$

در جایی که P نسبت نمونه‌های متناسب با آن ویژگی بر کل نمونه‌ها می‌باشد، مفهوم احتمال را خواهد داشت.

$$G(root, feature) = Entropy(root) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (۶)$$

در جایی که $Entropy(root)$ آنتروپی ویژگی گره ریشه معادل رابطه (۶) و S_v نمونه‌های مربوط به سایر ویژگی‌ها از مجموعه A و S نیز کل نمونه‌ها می‌باشند. $Entropy(S_v)$ نیز آنتروپی نمونه‌های سایر ویژگی‌ها است.

۲-۵- روش Naïve Bayes (NB)

به‌طور ساده روش بیز روشی برای دسته‌بندی پدیده‌ها، بر پایه احتمال وقوع یا عدم وقوع یک پدیده است. اگر برای فضای نمونه مفروضی بتوان چنان افرازی انتخاب شود که با دانستن اینکه کدام یک از پیشامدهای افراز شده رخ داده است، بخش مهمی از عدم اطمینان کاهش می‌یابد. فرض می‌شود $B_1,...,B_k$ افراز برای فضای نمونه‌ای S تشکیل دهند. طوری که به ازای هر $j = 1,...,k$ داشته باشید $P(B_j) > 0$ و فرض کنید A پیشامدی با فرض $P(A) > 0$ باشد، در این صورت به ازای $i = 1,...,k$ رابطه (۷) دارید [۱ و ۱۸].

$$P(B_i \mid A) = P(B_i) \frac{P(A \mid B_i)}{\sum_{j=1}^{k} P(B_j) P(A \mid B_j)} \quad (۷)$$

باید توجه داشت که نیازی به محاسبه کامل $P(C_i \mid A)$ نیست، چون در همه دسته‌ها مقدار مخرج مشترک خواهد بود. پس می‌توان از $P(A \mid C_i)P(C_i)$ برای هر دسته استفاده نمایید که در اینجا A مجموعه ویژگی‌ها و C_i هر یک از دسته‌ها است. پس با استفاده از قضیه بیز با هدف دسته‌بندی می‌توان رابطه (۸) بیز را به‌صورت زیر بازنویسی کرد:

$$C_{MAP} = \arg\max_{C_i \in C} \frac{P(a_1, a_2, ..., a_n \mid C_i) P(C_i)}{\sum_{j=1}^{k} P(C_j) P(A \mid C_j)} = \arg\max_{C_i \in C} P(a_1, a_2, ..., a_n \mid C_i) P(C_i) \quad (۸)$$

[1] Attribute Value Pair

حال با استفاده از داده‌های آموزشی سعی می‌شود دو جمله معادله بالا را تخمین زده شود. فرض روش طبقه‌بندی ساده بیز بر اساس این ساده‌سازی است که مقادیر صفات با داشتن مقادیر تابع هدف از یکدیگر مستقل شرطی می‌باشند. به عبارت دیگر، این فرض بیانگر این است که به شرط مشاهده خروجی تابع هدف احتمال مشاهده صفات $a_1, a_2, ..., a_n$ برابر ضرب احتمالات هر صفت به‌طور جداگانه می‌باشد. اگر این مفهوم را جایگزین معادله بالا کنید، رابطه (۹) روش طبقه‌بندی ساده بیزی را نتیجه می-دهد.

$$C_{NB} = \arg\max_{c_i \in C} P(a_1, a_2, ..., a_n \mid C_i) P(C_i) = \arg\max_{c_i \in C} P(C_i) \prod_j P(a_j \mid C_j) \quad (۹)$$

که در آن، NB خروجی طبقه‌بندی ساده بیزی برای تابع هدف می‌باشد. توجه شود که تعداد جملات $P(a_j \mid C_J)$ که در این روش باید محاسبه شوند برابر تعداد ضرب در تعداد دسته‌های خروجی برای تابع هدف می‌باشد که این مقدار از تعداد جملات بسیار کمتر است. نتیجه اینکه یادگیری ساده بیزی سعی در تخمین مقادیر مختلف $P(a_j \mid C_J)$ و $P(C_i)$ با استفاده از میزان تکرار آن‌ها در داده‌های آموزشی دارد.

بنابراین هر یک از روش‌های مبتنی بر یادگیری ماشین که منطبق بر مراجع علمی مورد بررسی قرار گرفتند را می‌توان به‌صورت یکتا در سامانه مدیریت ترافیک به‌عنوان طبقه‌بندی کننده ترافیک قرار داد. اما از آنجا که هر یک از این روش‌ها مانند هر الگوریتمی دارای نقاط ضعف و قوت مشخصی هستند، انتظار می‌رود برای رسیدن به کارایی بهتر، جدای از تلاشی که برای اصلاح‌سازی‌های درونی آن‌ها در کارهای انجام شده صورت گرفته است، ترکیب هدفمند آن‌ها رویکردی است که در این مقاله پیشنهاد و کارایی آن بررسی خواهد شد.

۳- رویکرد پیشنهادی ترکیبی

برای بررسی و ارزیابی صحت عملکرد یک مدل مبتنی بر یادگیری ماشین جدولی مشابه جدول (۱) با نام مشاهدات ارزیابی تعریف می‌شود که در ازای هر نمونه از مجموعه داده، نتیجه تصمیم‌گیری مدل نسبت به واقعیت برچسب آن نمونه مقایسه می‌شود. در فاز آزمایش بنابر توجه طراح سامانه به عملکرد امنیتی، مدیریتی و یا دسترسی کاربران به پهنای باند، نیاز مبرم به افزایش میزان شناسایی بیشتر آن نمونه ترافیکی وجود دارد تا حتی اگر یک نمونه که در واقعیت برچسبی مشابه آن نمونه را داراست هم با همان برچسب پیش‌بینی شود. با توجه به جدول (۱) معیار اصلی ما سطری است که در واقعیت برچسب مورد نظر ما وجود داشته و حالا سعی داریم تا آشکار شدن آن را توسط سامانه بررسی کنیم. بنابراین نسبت آشکار شدن درست به تمامی مشاهدات سامانه نسبت به آن برچسب را نرخ یا میزان احتمال آشکارسازی (PD) می‌دانیم که رابطه (۱۰) این معیار را نشان می‌دهد.

جدول (۱): مشاهدات ارزیابی

	مدل آن کمیت را تشخیص داد	مدل آن کمیت را تشخیص نداد
کمیت مورد نظر واقعا وجود دارد	True Positive (TP)	False Negative (FN)
کمیت مورد نظر واقعا وجود ندارد	False Positive (FP)	True Negative (TN)

$$PD = \frac{TP}{TP + FN} \quad (۱۰)$$

بنابراین با وجود TP در صورت رابطه (۱۰) می‌توان سه نکته را بیان کرد:

۱. با افزایش TP احتمال آشکارسازی افزایش می‌یابد.
۲. با کاهش FN احتمال آشکارسازی افزایش می‌یابد و چنانچه FN معادل صفر باشد نرخ آشکارسازی ۱۰۰٪ خواهد بود.
۳. چنانچه معیار تصمیم‌گیری ما به نحوی باشد که نظر چندین مدل در آن سهیم باشد می‌توان شاهد افزایش احتمال آشکارسازی بود.

با توجه به نکته سوم، چنانچه بیش از یک مدل در تصمیم‌گیری نهایی سامانه شرکت داشته باشند و اگر یکی از آن‌ها بیان کند که برچسب مورد نظر ما را برای آن نمونه تشخیص داده است، نظر کل سامانه نسبت به آن نمونه مبنی بر تشخیص دادن آن برچسب خواهد بود. بنابراین به نوعی تمامی مدل‌های به‌کار برده شده با یکدیگر به‌صورت OR منطقی ترکیب خواهند شد. شکل (۴) الگوریتم مورد نظر را بر مبنای برداری از برچسب‌های اصلی نمونه‌های آزمایشی و برداری مبتنی بر تصمیم‌گیری دو مدل یادگیری ماشین را نشان می‌دهد. هشدار غلط[1] به معنای آشکار سازی یک کمیت به‌صورت اشتباه می‌باشد. یعنی کمیتی در واقعیت وجود ندارد اما مدل به اشتباه آن را آشکار می‌کند. با

[1] False Alarm

توجه به جدول (۱) این تعریف معادل FP می‌باشد. اما برای اینکه بتوانیم احتمال یا نرخی برای این کمیت به نام نرخ هشدار غلط (FAR) تعریف کنیم نیازمند رابطه‌ای هستیم که در این کار تحقیقاتی آنرا تعریف می‌کنیم.

Algorithm 1: Hybrid Approach for increasing Probability of Detection

Input: test_vector as real labels, two vectors named predicted_vetctor for each ML models as their predicted labels.
Output: Probability of Detection (PD) value

```
1.define model_vector as output of Hybrid classifier.
2. For each index of two predicted_vectors:
3.        if one of them detected the desired label:
4.              model_vector[index]=1
5.        else:
6.              model_vector[index]=0
7. For each index of test_vector and model_vector:
8.        search for true labels reporting:
9.              Observing TP
10.       search for labels that do not reported:
11.             Observing FN
12.Calculating PD
13.Return PD
```

شکل (۴): شبه کد الگوریتم ترکیبی افزایش احتمال آشکارسازی

با توجه به جدول (۱) از آنجا که اساساً مبنای کار ما در FAR عدم وجود برچسب مورد نظر در آن نمونه است، بنابراین سطر دوم برای ما اهمیت دارد. بنابراین می‌توان اینگونه بیان کرد که نرخ هشدار غلط برابر رابطه (۱۱) است.

$$FAR = \frac{FP}{FP + TN} \qquad (۱۱)$$

با توجه به رابطه ارائه شده برای FAR، می‌توان اینگونه نتیجه گرفت که اثر تشخیص‌های اشتباه به‌عنوان FP در بالابردن FAR بسیار مشهود می‌باشد. از آنجا که هدف ما کاهش FAR است، قرارگیری FP در صورت کسر می‌تواند اثر مستقیمی در کاهش آن داشته باشد. بنابراین کاهش FP بسیار مؤثر خواهد بود. برای اینکه بتوانیم تا حد امکان از بروز تصمیم و آشکارسازی اشتباه یک برچسب جلوگیری کنیم، می‌توان از ترکیب بیش از یک مدل و استفاده از خروجی آن‌ها به‌صورت AND منطقی استفاده کرد. بدان معنا که اگر تمامی مدل‌های مورد استفاده ما در رویکرد ترکیبی دلالت بر تشخیص یک برچسب داشتند خروجی کلی سامانه ما برای آن نمونه مورد نظر اعلام تشخیص برچسب باشد. با این رویکرد اگر یک مدل که دچار خطا شده است برچسبی را به اشتباه تشخیص دهد اما سایر مدل‌ها این تشخیص را ندهند، خروجی سامانه پاسخی مبنی بر شناسایی آن برچسب بروز نخواهد داد. با اینکار به نوعی احتمال آشکارسازی را کاهش می‌دهیم اما با کنترل بهتر خطاهای مدل‌ها، می‌توان انتظار کاهش FP به معنای هشدار غلط و سپس کاهش FAR را داشت. بر همین مبنا با توجه به شکل (۵) می‌توان الگوریتم رویکرد ترکیبی جهت کاهش نرخ هشدار غلط و محاسبه آن را مبتنی بر رابطه (۱۱) مشاهده کرد. مشابه شکل (۴)، در شکل (۵) نیز تصمیم‌گیری با مقایسه دو بردار برچسب‌های پیش‌بینی شده توسط همکاری دو مدل بر مبنای AND منطقی مشاهده می‌شود. بنابراین انتظار داریم تا با به‌کارگیری الگوریتم‌های ارائه شده در این بخش، شاهد بهبود احتمال آشکارسازی و نرخ هشدار غلط مبتنی بر روابط (۱۰ و ۱۱) باشیم. این استدلال‌ها در بخش آتی در کنار نتایج به‌کارگیری از مدل‌های مبتنی بر یادگیری ماشین به‌صورت یکتا بررسی و تحلیل خواهند شد.

Algorithm 2: Hybrid Approach for Decreasing False Alarm Rate

Input: test_vector as real labels, two vectors named predicted_vetctor for each ML models as their predicted labels.
Output: False Alarm Rate (FAR) value

```
1.define model_vector as output of Hybrid classifier.
2. For each index of two predicted_vectors:
3.        if both of them detected the desired label:
4.              model_vector[index]=1
5.        else:
6.              model_vector[index]=0
7. For each index of test_vector and model_vector:
8.        search for false labels reporting:
9.              Observing FP
10.       search for labels that truly do not reported:
11.             Observing TN
12.Calculating FAR
13.Return FAR
```

شکل (۵): شبه‌کد الگوریتم ترکیبی برای کاهش هشدار غلط

۴- شبیه‌سازی و تحلیل نتایج

برای شبیه‌سازی و اجرای روش‌ها و الگوریتم‌های بیان شده در بخش دوم و سوم، بستر سخت‌افزاری و نرم‌افزاری مطابق با جدول (۲) مورد استفاده قرار گرفت که مبتنی بر زبان برنامه نویسی پایتن می‌باشد.

مجموعه داده مورد استفاده در این مقاله تحقیقاتی مجموعه داده USTC-TFC2016 است که یک مجموعه از فایل‌های ضبط شده ترافیکی شبکه‌ای بی‌سیم بر مبنای نرم‌افزار وایرشارک می‌باشد که تمامی فایل‌های آن با فرمت pcap. ذخیره می‌باشند.

جدول (۲): اطلاعات سخت‌افزاری و نرم‌افزاری پیاده‌سازی

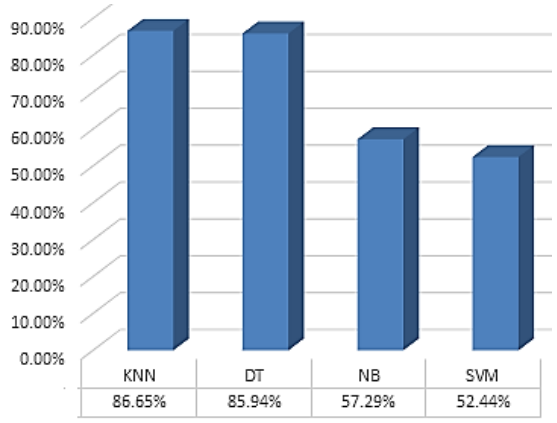
اطلاعات نرم‌افزاری	
نرم‌افزار پیاده‌سازی	Anaconda 3
محیط توسعه کدنویسی	Jupyther Notebook
نسخه زبان پایتن	3.6.0
اطلاعات سخت‌افزاری	
پردازنده	Intel Corei7- 4500U
حافظه	6G

این فایل‌ها در چند مرحله تبدیل به مجموعه داده قابل کار با روش‌های یادگیری ماشین شدند که شامل ۱۰۰۰ نمونه با پنج ویژگی میانگین دوره زمان بین ارسال بسته‌ها، میانگین، حداقل و حداکثر طول بسته‌ها و تعداد آن‌ها برای هر جریان ترافیکی است.

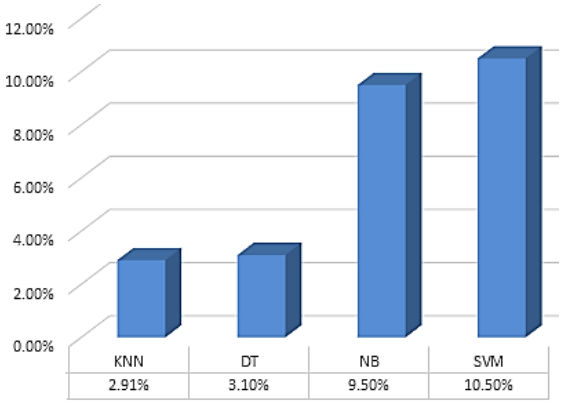
ابتدا با توجه به انواع تنظیمات و پارامترهایی که برای هر یک از این مدل‌ها به‌صورت جداگانه مورد ارزیابی قرار داده شد، بهترین حالت به لحاظ دقت برای هر یک از مدل‌ها استخراج و سپس با یکدیگر مقایسه شدند. منطبق بر همین مسیر برای مدل KNN با پنج همسایه (K=5) بهترین پاسخ را به‌دست آورده شد. برای مدل SVM با تابع کرنل چند جمله‌ای درجه شش بهترین پاسخ حاصل شد. لازم به ذکر است که نرمال‌سازی Standard Scaler برای تمامی موارد و مدل‌ها منجر به پاسخ‌های بهینه بود که تمامی این موارد در فاز ارزیابی، تحلیل و بررسی شدند.

۴-۱- تحلیل نتایج به‌کارگیری روش‌های یادگیری ماشین به‌صورت یکتا

با توجه به روابط (۱۰ و ۱۱) که هدف اصلی ما از بررسی عملکرد مدل‌های معرفی شده در بخش دوم و ارائه رویکرد پیشنهادی در بخش سوم برای افزایش کیفیت آن‌ها بودند، بنابریان این دو رابطه را به‌عنوان معیارهای ارزیابی درنظر می‌گیریم. در گام نخست ابتدا هر یک از مدل‌ها به‌صورت یکتا را به وسیله مجموعه داده آموزش می‌دهیم و پس از فاز ارزیابی، در فاز آزمایش میزان احتمال آشکارسازی و نرخ هشدار غلط را استخراج کردیم. شکل‌های (۶ و ۷) به ترتیب مقایسه احتمال آشکارسازی و نرخ هشدار غلط را برای مدل‌های مورد نظر نشان می‌دهند. همان‌طور که پیش‌تر نیز بیان شد، هدف افزایش میزان احتمال آشکارسازی و کاهش نرخ هشدار غلط می‌باشد. بنابراین انتظار می‌رود تا در صورت به‌کارگیری از رویکرد ترکیبی ارائه شده در این مقاله شاهد بهبود دو معیار PD و FAR باشیم.



شکل (۶): مقایسه احتمال آشکارسازی روش‌های یادگیری ماشین به‌صورت یکتا



شکل (۷): مقایسه میزان نرخ هشدار غلط روش‌های یادگیری ماشین به‌صورت یکتا

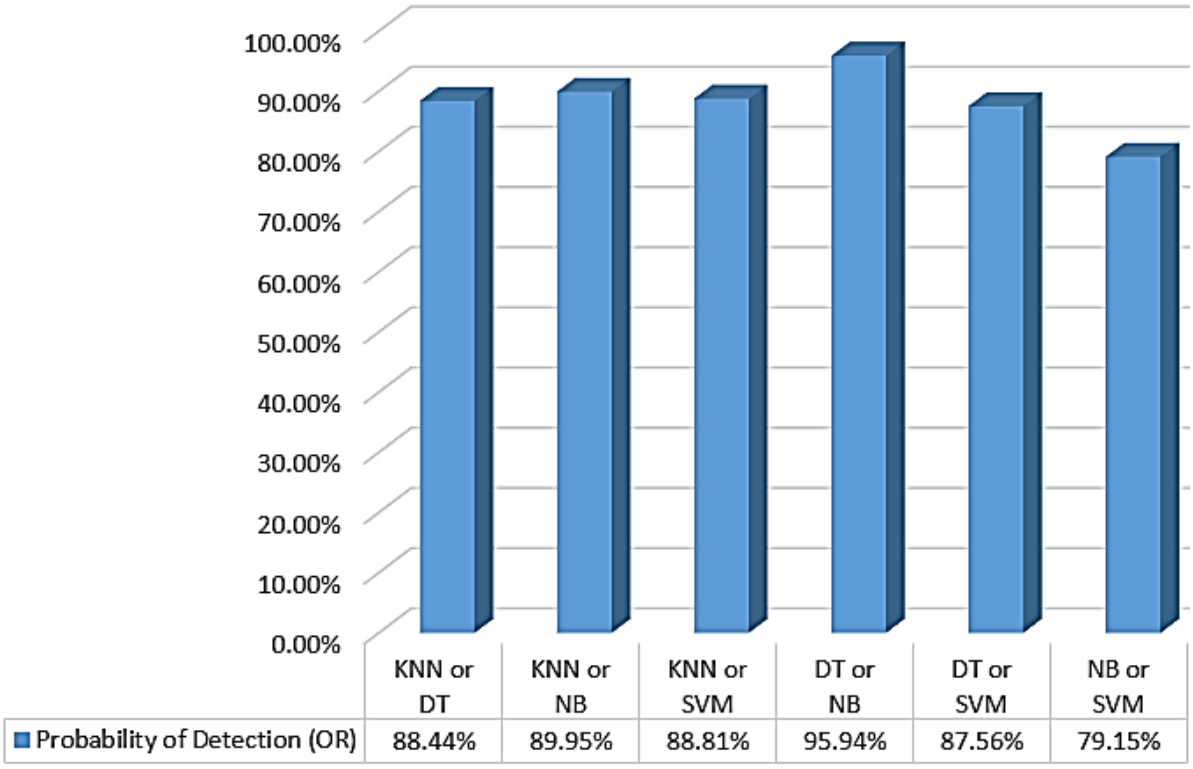
۴-۲- تحلیل کارایی رویکرد پیشنهادی ترکیبی

همان‌طور که در بخش سوم بیان شد، جهت افزایش میزان حساسیت به آشکارسازی برچسب‌های مطلوب ترافیکی و همچنین کاهش اعلام تصمیمات اشتباه سامانه طبقه‌بندی کننده، رویکرد ترکیب هدفمند مدل‌های یادگیری ماشین را مطرح کردیم. مبتنی بر شکل‌های (۴ و ۵) الگوریتم‌های ترکیبی مدل-های مورد نظر را به‌صورت دو به دو با هر یک از روش‌های یادگیری ماشین به‌کار گرفتیم که نتایج دو شکل (۸ و ۹) به ترتیب اثر مثبت رویکرد پیشنهادی این مقاله را در افزایش احتمال آشکارسازی و کاهش نرخ هشدار غلط نشان می‌دهند. از آنجا که در فاز آزمایش این روش‌ها (به‌صورت یکتا) در شکل‌های (۶ و ۷) نیز مشاهده شد، دو روش KNN و DT بهترین عملکرد را ارائه داده بودند. بنابراین مطابق با انتظار نتایج دو شکل (۸ و ۹) نیز نشان می‌دهند که ترکیب روش‌های قوی‌تر با روش‌های دیگر می‌تواند موجب افزایش بهره‌وری همه مدل‌های یادگیری ماشین

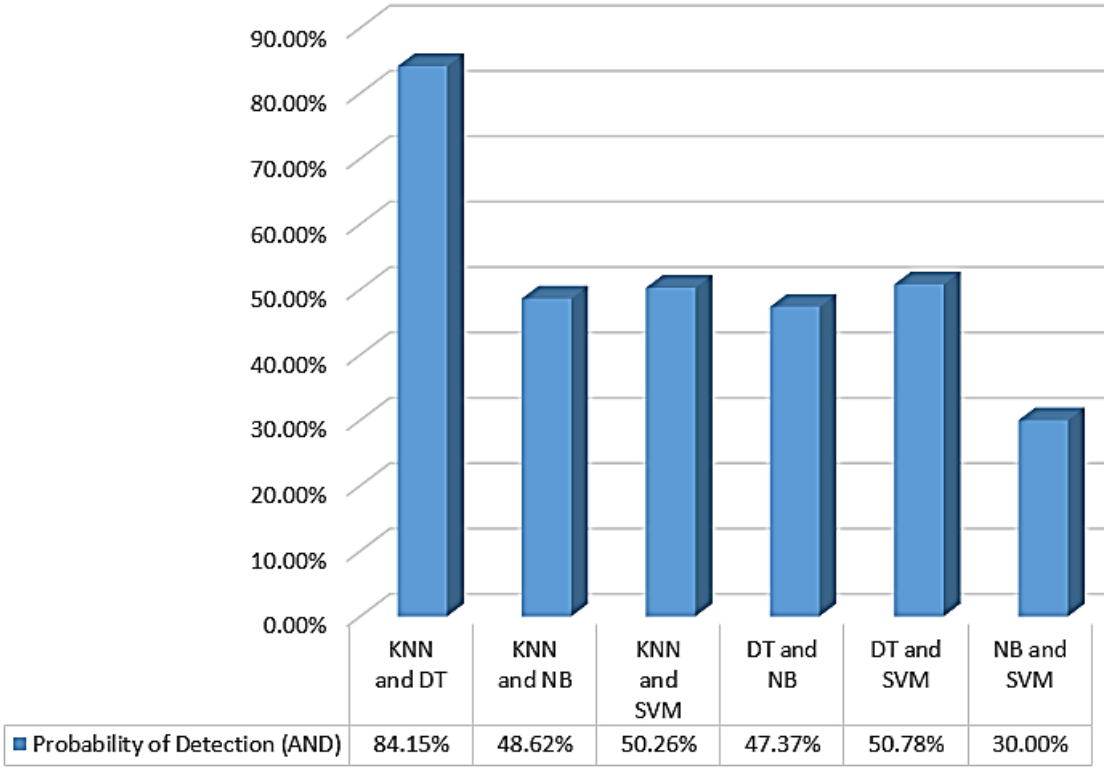
از جمله SVM و NB باشد. بنابراین می‌توان بیان کرد که به‌کارگیری رویکرد پیشنهادی این مقاله می‌تواند گام مثبتی در بهبود عملکرد روش‌های مبتنی بر یادگیری ماشین در شناسایی و طبقه‌بندی ترافیک در شبکه‌های بی‌سیم باشد.

همان‌طور که در شکل (۸) نیز مشاهده می‌شود به‌کارگیری الگوریتم AND می‌تواند به بهبود احتمال آشکارسازی کمک شایانی نماید که این امر اثر بسیار نامطلوبی در نرخ هشدار غلط ندارد. این بدان معنا است که چنانچه در یک سامانه هر دو معیار PD و FAR پارامترهای کارایی سامانه طبقه‌بندی کننده باشند و هدف افزایش PD باشد، به‌کارگیری الگوریتم OR نباید موجب افزایش بیش از حد نرخ هشدار غلط شود. مطابق شکل (۸) نتیجه به‌کارگیری رویکرد ترکیبی OR از مجموع نرخ هشدار غلط یکتای دو روش کمتر است. بنابراین این نکته نشان می‌دهد که چنانچه مطلوب افزایش PD باشد، با استفاده از رویکرد ترکیبی مبتنی بر الگوریتم OR این امر به نحوی حاصل می‌شود که اثر مخرب زیادی در نرخ هشدار غلط به وجود نمی‌آید. بنابراین مبادله‌ای بین PD و FAR مشاهده می‌شود.

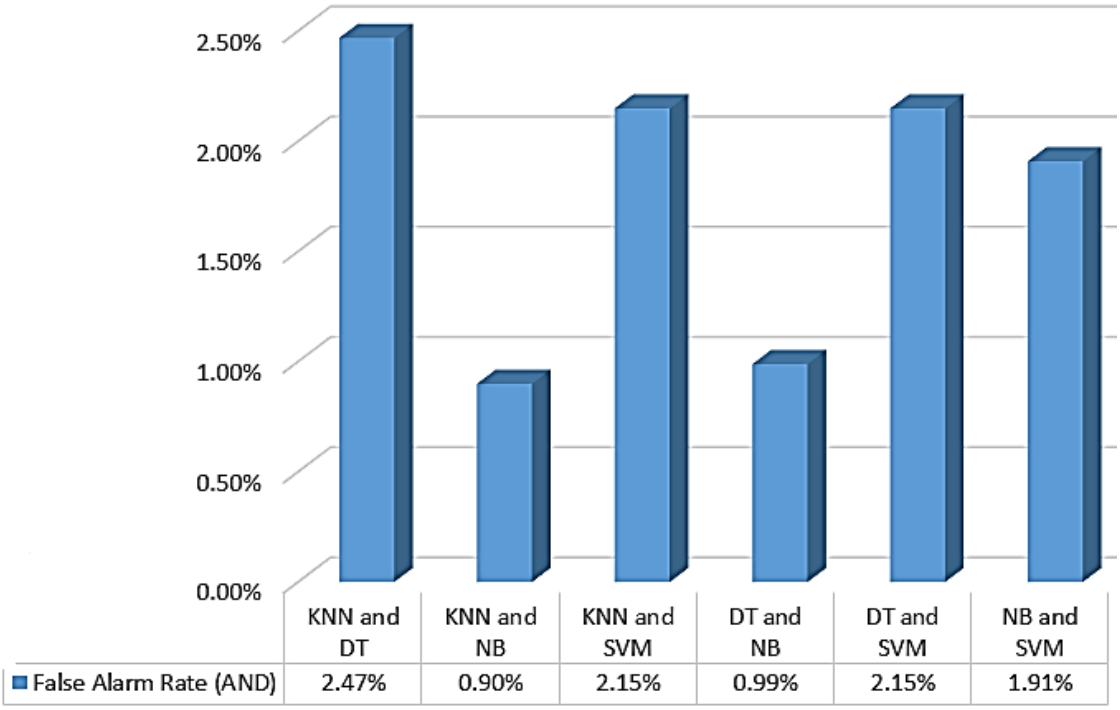
منطبق بر شکل (۹) در صورتی که هدف اصلی از بررسی کارکرد سامانه طبقه‌بندی کننده بهبود نرخ هشدار غلط باشد، با استفاده از رویکرد ترکیبی مبتنی بر الگوریتم AND می‌توان به نحوی به این مهم دست یافت که کاهش شدید احتمال آشکارسازی را نیز شاهد نباشیم. این بدان معنا است که نرخ هشدار غلط با استفاده از الگوریتم AND به نسبت به‌کارگیری یکتای روش‌های مورد نظر کمتر است. از طرفی نتیجه احتمال آشکارسازی الگوریتم AND نسبت به کمترین PD روش‌های شرکت کننده در رویکرد ترکیبی کمتر است. با توجه به شکل (۹-الف) می‌توان نتیجه گرفت که ترکیب روش‌های قوی‌تری همچون DT و KNN می‌تواند از کاهش PD جلوگیری کند. در سمت مقابل به دلیل آنکه روش‌های SVM و NB به‌صورت یکتا نیز کم است، بنابراین از ترکیب آن‌ها نمی‌توان انتظار نتیجه مطلوبی در حفظ PD داشت.



شکل (۸): مقایسه اعمال رویکرد ترکیبی OR: (الف) احتمال آشکارسازی و (ب) نرخ هشدار غلط



(الف)



(ب)

شکل (۹): مقایسه اعمال رویکرد ترکیبی AND: (الف) احتمال آشکارسازی و (ب) نرخ هشدار غلط

۵- نتیجه‌گیری

در این مقاله با توجه به گسترش حوزه‌های کاربردی شبکه‌های بی‌سیم اقتضایی و نیز مسئله مدیریت چنین شبکه‌هایی، روش‌های مبتنی بر یادگیری ماشین به‌عنوان مدل‌هایی بدون دخالت انسان و با دقت بالاتر نسبت به روش‌های مرسوم جهت شناسایی و طبقه‌بندی ترافیک بحث و بررسی شدند. دو معیار احتمال آشکارسازی و نرخ هشدار غلط تعریف شد که رویکرد ترکیبی پیشنهادی در این مقاله جهت کمک به بهبود آن‌ها معرفی شد. نتایج نشان دادند که ترکیب هدفمند مدل‌های مبتنی بر یادگیری ماشین اثر بسیار محسوس و مثبتی در این معیارهای ارزیابی دارند و موجب بهبود عملکرد سامانه طبقه‌بندی ترافیک شبکه می‌شوند. از سویی دیگر به‌کارگیری از این روش پیشنهادی می‌تواند نکات مثبت مدل‌ها را در جهت افزایش دقت و کارایی سامانه با یکدیگر ترکیب کرده و از میزان اثر نقاط ضعف یکدیگر جلوگیری کنند.

۶- مراجع

[1] J. Erman, A. Mahanti, and M. Arlitt, "Qrp05-4: Internet Traffic Identification Using Machine Learning," *IEEE Globecom*, IEEE, 2006.

[2] L. T. Hu and LiJun Zhang, "Real-time Internet Traffic Identification Based on Decision Tree," *World Automation Congress*, IEEE, 2012.

[3] J. Zhang, C. Chen, Y. Xiang, W. Zhou and A. V. Vasilakos, "An Effective Network Traffic Classification Method with Unknown Flow Detection," *IEEE Transactions on Network and Service Management, vol.* 10, no.2, pp. 133-147, 2013.

[4] A. C. Callado, "A Survey on Internet Traffic Identification," *IEEE Communications Surveys and Tutorials, vol.* 11, no. 3, pp. 37-52, 2009.

[5] R. Ma and S. Qin, "Identification of Unknown Protocol Traffic Based on Deep Learning," *3rd IEEE International Conference on Computer and Communications (ICCC)*, IEEE, 2017.

[6] Sh. Zuozhi, Y. Yue, and M. Yunlang, "The Research of Protocol Identification Based on Traffic Analysis," *10th International Conference on Intelligent Computation Technology and Automation (ICICTA)*, IEEE, 2017.

[7] L. Kong, G. Huang, K. Wu, Q. Tang and S. Ye, "Comparison of Internet Traffic Identification on Machine Learning Methods," *International Conference on Big Data and Artificial Intelligence (BDAI)*, IEEE, 2018.

[8] S. Kokila, A. Sathish, and R. Shankar, "Comparative Analysis of Internet Traffic Identification Methods," *Proceedings of the UGC Sponsored National Conference on Advanced Networking and Applications*, 2015.

[9] L. Peng, "Early Stage Internet Traffic Identification Using Probabilistic Neural Networks," *International Journal of Computer and Communication Engineering, vol.* 4, no. 6, pp. 417-425, 2015).

[10] Zh. Wang, "The Applications of Deep Learning on Traffic Identification," *BlackHat USA* 24, 2015.

[11] P. Singhal, R. Mathur, and H. Vyas, "Network Traffic Classification Based on Unsupervised Approach," *International Journal of Computer Applications*, 975, 8887, 2013.

[12] R. Thupae, B. Isong, N. Gasela and A. M. Abu-Mahfouz, "Machine Learning Techniques for Traffic Identification and Classifiacation in SDWSN: A Survey," *IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society*, IEEE, 2018.

[13] Y. Liu, W. Li, and Y. Ch. Li, "Network Traffic Classification Using K-means Clustering," *Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2007)*, IEEE, 2007.

[14] Ch. GU and Sh. Zhang, "Online Wireless Mesh Network Traffic Classification Using Machine Learning," Journal of Computational Information Systems, 2011.

[15] J. Ran and Y. Chen, "Three Dimensional Convolutional Neural Network Based Traffic Classification for Wireless Communications," pp. 624-627, 2018,

[16] W. Wang, M. Zhu, X. Zeng, X. Ye, and Y. Sheng, "Malware Traffic Classification Using Convolutional Neural Network for Representation Learning," ICOIN, IEEE, 2017.

[17] F. Noorbehbahani and S. Mansoori, "A New Semi-Supervised Method for Network Traffic Classification Based on X-Means Clustering and Label Propagation," 8th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, pp. 120-125, 2018.

[18] C. M. Bishop, "Pattern Recognition and Machine Learning (Information Science and Statistics)," Springer, 2007. ISBN: 0387310738https://github.com/echowei/DeepTraffic/tree/master/1.malware_traffic_classification/1.DataSet(USTC-TFC2016)/Benigen.