Electronic & Cyber Defense
Vol. 13, No. 3, 2025 (Serial No. 51). pp. 1-17
ISSN: 2322-4347, E-ISSN: 2980-8979

Computer Networks traffic classification model based on DBScan clustering and gamma classification

Seyede Zohreh Majidian[1], Shiva TaghipourEivazi[2]*, Bahman Arasteh[3], Ali Ghaffari[4]
[1] PhD Student. Department of Computer Engineering, Aras international Branch, Islamic Azad University, Tabriz, Iran. Email: *std.z.majidian@iaut.ac.ir*
[2] Assistant Professor. Department of Computer Engineering, Ta.C.Islamic Azad university, Tabriz, Iran (Correspondence: *taghipour@iau.ac.ir*)
[3] Associate Professor. Department of Computer Engineering, Ta.C.Islamic Azad university, Tabriz, Iran. Email: *b.arasteh@iaut.ac.ir*
[4] Professor. Department of Computer Engineering, Ta.C.Islamic Azad university, Tabriz, Iran. Email: *a.ghaffari@iaut.ac.ir*

ARTICLE INFO

Article history:

Article Type: Research paper

Receive: 08 June 2025
Revise: 09 August 2025
Accept: 09 September 2025
Available online: 03 October 2025

Keywords:
Traffic classification
machine learning
DBScan clustering
gamma classification

ABSTRACT

Traffic classification is a crucial network monitoring process with wide applications in security, quality of service, and network management. With the increasing complexity and variety of network traffic, new challenges arise, including the lack of labeled training data. To address this challenge, this paper presents a traffic classification mechanism that combines unsupervised and semi-supervised machine learning algorithms. This mechanism uses a limited set of labeled training data to improve classification accuracy. The proposed method represents each traffic flow as a feature vector containing the statistical characteristics of that flow. The number of features generated for each sample is reduced using principal component analysis. DBScan clustering is employed to determine the correct traffic type for each untagged traffic stream. Finally, the gamma classifier model is used to separate the new traffic flows. The efficiency of the proposed method has been evaluated using real data sets. The results show that the proposed method can classify traffic flows with an average accuracy of 95.12%, representing at least a 7.03% improvement over previous approaches.

Cite this article: Majidian, S. Z., TaghipourEivazi, S., Arasteh, B., & Ghaffari, A. (2025). Computer NetworksTraffic Classification Model Based on DBScan Clustering and Gamma Classification. *Electronic and Cyber Defense*, *13*(3), 1-17. 2025.

DOI: https://dor.isc.ac/dor/20.1001.1.23224347.1404.13.3.1.0

Publisher: Imam Hossein University

پدافند الکترونیکی و سایبری
سال سیزدهم، شماره ۳، پاییز ۱۴۰۴، ص ۱-۱۷
شاپا الکترونیکی: ۴۳۴۷-۲۳۲۲ شاپا چاپی: ۸۹۷۹-۲۹۸۰

مدل طبقه‌بندی ترافیک شبکه‌های کامپیوتری مبتنی بر خوشه‌بندی DBScan و طبقه بند گاما

سیده زهره مجیدیان [۱]، شیوا تقی پور عیوضی*[۲]، بهمن آراسته[۳]، علی غفاری[۴]

[۱] دانشجوی دکتری، گروه مهندسی کامپیوتر، واحد بین‌المللی ارس، دانشگاه آزاد اسلامی، تبریز، ایران. *std.z.majidian@iaut.ac.ir*

[۲] استادیار، گروه مهندسی کامپیوتر، واحد تبریز، دانشگاه آزاد اسلامی، تبریز، ایران. (نویسنده مسئول: *taghipour@iau.ac.ir*)

[۳] دانشیار، گروه مهندسی کامپیوتر، واحد تبریز، دانشگاه آزاد اسلامی، تبریز، ایران. *b.arasteh@iaut.ac.ir*

[۴] استاد،گروه مهندسی کامپیوتر، واحد تبریز، دانشگاه آزاد اسلامی، تبریز، ایران. *a.ghaffari@iaut.ac.ir*

چکیده

طبقه‌بندی ترافیک یکی از مهم‌ترین فرآیندهای نظارت بر شبکه است که کاربردهای گسترده‌ای در حوزه‌های امنیت، کیفیت خدمات و مدیریت شبکه دارد. با افزایش پیچیدگی و تنوع ترافیک شبکه، چالش‌های جدیدی ازجمله کمبود داده‌های آموزشی برچسب به وجود می‌آید. به‌منظور رفع این چالش در این مقاله، سازوکار طبقه‌بندی ترافیک با ترکیب الگوریتم‌های یادگیری ماشین بدون نظارت و نیمه نظارتی ارائه می‌شود. این سازوکار از مجموعه محدودی از داده‌های آموزشی برچسب‌گذاری شده برای بهبود دقت طبقه‌بندی استفاده می‌کند. روش پیشنهادی، هر جریان ترافیک را به‌عنوان یک بردار ویژگی توصیف می‌کند که شامل ویژگی‌های آماری آن جریان است. تعداد ویژگی‌های ایجادشده برای هر نمونه نیز با استفاده از تحلیل مؤلفه‌های اصلی کاهش می‌یابد. خوشه‌بندی DBScan برای تعیین نوع ترافیک صحیح برای هر جریان ترافیک بدون برچسب استفاده می‌شود. درنهایت، از مدل طبقه‌بندی گاما برای تفکیک جریان‌های ترافیک جدید استفاده می‌شود. کارایی روش پیشنهادی با استفاده از مجموعه داده‌های واقعی ارزیابی‌شده است. نتایج نشان می‌دهد که روش پیشنهادی قادر به طبقه‌بندی جریان‌های ترافیکی با دقت متوسط ۹۵,۱۲ درصد است که حداقل ۷,۰۳ درصد بهبود را نسبت به رویکردهای قبلی نشان می‌دهد.

مشخصات مقاله

تاریخچه مقاله:

نوع مقاله: علمی- پژوهشی

دریافت: ۱۴۰۴/۰۳/۱۸

بازنگری: ۱۴۰۴/۰۵/۱۳

پذیرش: ۱۴۰۴/۰۶/۱۶

ارائه آنلاین: ۱۴۰۴/۰۷/۱۶

کلیدواژه‌ها:

طبقه‌بندی ترافیک

یادگیری ماشین

خوشه‌بندی DBScan

طبقه‌بندی گاما

استناد: مجیدیان، سیده زهره، تقی پور عیوضی، شیوا، آراسته، بهمن، و غفاری، علی. (۱۴۰۴). مدل طبقه‌بندی ترافیک شبکه‌های کامپیوتری مبتنی بر خوشه‌بندی DBScan و طبقه بند گاما. *پدافند الکترونیکی و سایبری*، ۱۳(۳)، ص ۱-۱۷. https://dor.isc.ac/dor/20.1001.1.23224347.1404.13.3.1.0

ناشر: دانشگاه جامع امام حسین(ع).

۱. مقدمه

طبقه‌بندی ترافیک شبکه کاربردهای زیادی ازجمله کنترل ترافیک، امنیت شبکه و تحلیل شبکه دارد[۱]. یکی از مشکلات مهم در زمینه طبقه‌بندی ترافیک شبکه، کمبود داده‌های برچسب‌گذاری شده آموزشی است. برای آموزش مدل‌های طبقه‌بندی ترافیک شبکه، نیاز به داده‌های زیادی است که هرکدام با برچسب صحیح مشخص‌شده باشند. جمع‌آوری چنین داده‌هایی بسیار دشوار و زمان‌بر است. راه‌حل‌های مختلفی برای حل این مشکل ارائه‌شده است. استفاده از روش‌های یادگیری تقویتی[۲] یکی از راه‌حل‌ها است که در این روش مدل به‌صورت آزمایش و خطا یاد می‌گیرد. این روش می‌تواند برای آموزش مدل‌های طبقه‌بندی ترافیک شبکه بدون نیاز به داده‌های برچسب‌گذاری شده استفاده شود. روش دیگر، به‌کارگیری روش-های یادگیری بدون نظارت یا نیمه نظارتی است. در یادگیری بدون نظارت (نیمه نظارتی)، الگو (مدل) با استفاده از داده‌های بدون برچسب (یا نمونه‌های محدود دارای برچسب) یادگیری را انجام می‌دهد. این روش می‌تواند برای شناسایی الگوهای موجود در داده‌ها استفاده شود [۳]. از طرفی، لزوم بررسی شماره درگاه، داده‌های کاربر یا رفتار میزبان در شبکه ازجمله محدودیت‌هایی هستند که روش‌های پیشین برای طبقه‌بندی ترافیک شبکه با آن مواجه هستند و دسترسی به این اطلاعات در شبکه‌های عمومی منجر به نقض حریم خصوصی کاربران می‌شود. هدف این پژوهش، ارائه روشی جدید برای طبقه‌بندی جریان‌های ترافیک شبکه با استفاده از روش‌های یادگیری ماشین و بدون نیاز به بررسی شماره درگاه، داده‌های کاربر یا رفتار میزبان است. این الگوریتم، از ترکیبی از روش‌های یادگیری ماشین[۴]، [۵] بدون نظارت و نظارت‌شده استفاده می‌کند. در مرحله اول، از تحلیل مؤلفه‌های اصلی برای کاهش ابعاد ویژگی‌های آماری جریان استفاده می‌شود. این کار، حجم داده‌های موردنیاز برای آموزش مدل را کاهش می‌دهد و آن را برای سناریوهای بلادرنگ قابل‌اجرا می‌کند. در مرحله دوم، از خوشه‌بندی بدون نظارت توسط DBScan[1] [۶] برای تخمین برچسب نمونه‌های آموزشی بدون برچسب استفاده می‌شود. این کار، دقت طبقه‌بندی را در صورت کمبود داده‌های برچسب‌گذاری بهبود می‌بخشد. درنهایت، طبقه‌بندی جریان توسط یک مدل طبقه‌بندی گاما انجام می‌شود. این مدل، دقت طبقه‌بندی را در داده‌های ترافیک واقعی بهبود می‌بخشد. رویکرد پیشنهادشده برای طبقه‌بندی ترافیک شبکه در این مقاله، نوآوری‌های متعددی را ارائه می‌دهد که به‌طور مؤثری چالش‌های دنیای واقعی در این حوزه را برطرف می‌نماید. ترکیب نوآورانه ارائه‌شده از راهبردهای یادگیری بدون نظارت و با نظارت، این امکان بهره‌مندی از حداکثر پتانسیل داده‌های برچسب‌گذاری شده و بدون برچسب را می‌دهد. به‌ویژه، استفاده از الگوریتم خوشه‌بندی DBScan برای کشف الگوهای پنهان در داده‌های بدون برچسب، روشی جدید و کارآمد برای شناسایی انواع مختلف ترافیک ارائه می‌دهد که در تحقیقات پیشین این حوزه موردتوجه قرار نگرفته است. همچنین، طبقه‌بندی گاما به‌عنوان یک مدل یادگیری با نظارت، توانایی بالایی در تشخیص تفاوت‌های ظریف بین انواع مختلف ترافیک را دارد. این ترکیب منحصربه‌فرد، به‌ویژه در شرایطی که داده‌های برچسب‌گذاری شده محدود است، عملکرد مطلوبی از خود نشان می‌دهد و می‌تواند چالش‌های ناشی از شباهت زیاد بین برخی از انواع ترافیک را برطرف نماید. درمجموع، این روش، یک رویکرد قوی و انعطاف‌پذیر برای طبقه‌بندی جریان‌های ترافیک شبکه است که می‌تواند در سناریوهای دنیای واقعی مورداستفاده قرار گیرد.

ادامه این مقاله به شرح زیر سازمان‌دهی شده است: بخش ۲ شامل بررسی تحقیقات مرتبط است. جزئیات روش پیشنهادی در بخش ۳ توضیح داده‌شده است. بخش چهارم به نتایج و بحث‌ها اختصاص دارد. بخش ۵ شامل نتیجه‌گیری و ارائه طرحی کلی برای کارهای آینده است.

۲. مروری بر رویکردهای پیشین

تجزیه‌وتحلیل ترافیک ترکیبی از روش‌هایی است که به دنبال تعیین پیکربندی‌های نادرست، ناهنجاری‌ها، الگوها و روابط در داده‌های ترافیکی شبکه است[۷] و هدف آن تعیین نوع ترافیک یا نام برنامه کاربردی است. این راهبردها می‌توانند ترافیک را در کلاس‌های از پیش تعریف‌شده مانند غیرعادی یا عادی دسته‌بندی کنند. دقت روش‌ها برای طبقه‌بندی ترافیک به نحوه ردیابی ترافیک بستگی دارد که این امر خود به نمونه‌های زمینه، یعنی برنامه‌ها یا پروتکل‌ها با برچسب‌های صحیح متکی است. جمع‌آوری داده‌های معتبر از جریان ترافیک ضروری است زیرا اساس آموزش و ارزیابی عملکرد است[۸]. امروزه طبقه‌بندی ترافیک به یک مسئله چالش‌برانگیز تبدیل‌شده است و دلیل آن ظهور روش‌های جدید مانند رمزنگاری و کپسوله‌سازی است که عملکرد راهبردهای طبقه‌بندی مرسوم را کاهش می‌دهد[۹]. محققان در [۱۰] بر انتخاب ویژگی‌های بهینه در مسئله طبقه‌بندی ترافیک رمزگذاری شده تمرکز کرده‌اند و رویکردی هدفمند برای این منظور ارائه کرده‌اند. در این روش، رویکردهای انتخاب ویژگی با جزئیات برای مجموعه داده‌های مختلف تجزیه‌وتحلیل شده است. درروش ارائه‌شده در مقاله [۱۱] مدل ماشین بردار پشتیبان حساس به هزینه برای حل این مشکل معرفی‌شده است. رویکرد ارائه‌شده ترکیبی از ماشین بردار پشتیبان چند کلاسه و روش یادگیری فعال است که از وزن دهی پویا برای بهبود عملکرد ماشین بردار پشتیبان استفاده می‌کند. دقت طبقه‌بندی و عملکرد الگوریتم ارائه‌شده با استفاده از دو مجموعه داده MOORE_SET و NOC_SET بررسی‌شده است.

[1] Density-Based Spatial Clustering of Applications with Noise

نتایج بیانگر این است که الگوریتم ارائه‌شده منجر به کاهش هزینه محاسبات و نیز افزایش دقت طبقه‌بندی‌شده است. روش ارائه‌شده در مقاله [۱۲] باهدف طبقه‌بندی ترافیک رمزگذاری شده، از یک شبکه عصبی عمیق با ساختار موازی در شبکه تقویت‌شده است و پس از هر لایه کانولوشن یک شبکه میکرو را برای بهبود مدل‌سازی محلی اتخاذ می‌کند. این طبقه‌بندی کننده مدل یادگیری عمیق مبتنی بر معماری شبکه عصبی کانولوشن است. هر جزء کانولوشن به‌عنوان یک شبکه میکرو برای بهبود مدل‌سازی محلی عمل می‌کند. همچنین این مدل لایه‌های کاملاً متصل را با لایه‌های ادغام میانگین سرا سری جایگزین کرده است که در ساده‌سازی مدل مؤثر است و تعداد پارامترهای مدل را به میزان قابل‌توجهی کاهش می‌دهد. نتایج آزمایش‌های روی مجموعه داده‌های ترافیک رمزگذاری شده ISCX VPN-nonVPN بیانگر این است که مدل پیشنهادی تعادل بهتری بین دقت طبقه‌بندی و پیچیدگی مدل را داراست. نویسندگان در مقاله [۱۳]، از روش‌های یادگیری گروهی برای طبقه‌بندی ترافیک شبکه ایمن استفاده کرده‌اند. دو مدل مجموعه‌ای تقویت گرادیان و درخت تصمیم‌گیری برای طبقه‌بندی ترافیک در اتصالات استفاده‌شده است. ترکیبی از ویژگی‌های طیفی و زمانی ترافیک شبکه برای طبقه‌بندی در [۱۴] استفاده‌شده است. در این روش، شبکه‌های رمزگذار خودکار مکرر با شبکه عصبی کانولوشن بازگشتی ترکیب می‌شوند تا رابطه بین ویژگی‌های مختلف شبکه و ویژگی‌های طیفی را توصیف کنند. همچنین دنباله این ویژگی‌ها به‌عنوان ویژگی‌های زمانی در نظر گرفته می‌شود که باهم ترکیب‌شده و برای طبقه‌بندی استفاده می‌شوند. روش ارائه‌شده که از روی‌هم قرار دادن رمزگذار خودکار با یک شبکه عصبی کاملاً متصل شکل‌گرفته است، به بهبود ۲۸ درصدی در معیار دقت نسبت به رویکردهای مطالعه شده قبلی مبتنی بر یادگیری ماشین دست‌یافته است. یک مطالعه تجربی برای طبقه‌بندی ترافیک شبکه اینترنت اشیا با استفاده از یادگیری ماشین در مقاله [۱۵] انجام‌شده است. این تحقیق از مجموعه داده‌ای از ترافیک شبکه دستگاه‌های اینترنت اشیا در یک بازه زمانی ۲۰ روزه استفاده کرده و دقت الگوریتم‌های مختلف یادگیری ماشین را در طبقه‌بندی این داده‌ها گزارش کرده است. در [۱۶] از سازوکار یادگیری انتقال عمیق برای طبقه‌بندی ترافیک در شبکه‌های اینترنت اشیا نسل پنجم استفاده‌شده است. این رویکرد، طبقه‌بندی را با تنظیم دقیق شبکه عصبی و سازوکار انتقال وزنی پیکربندی می‌کند. بر اساس این راه‌حل، کل داده‌های آموزشی موردنیاز نخواهد بود و تنها با استفاده از ۱۰ درصد از داده‌ها، می‌توان به‌دقت طبقه‌بندی مشابه دست‌یافت. در رویکرد ارائه‌شده در مقاله [۱۷] روش‌های مختلف یادگیری ماشین ازجمله ماشین بردار پشتیبان و شبکه‌های بیزین ساده در شبکه‌های تعریف‌شده نرم‌افزاری برای طبقه‌بندی ترافیک داده‌ها ادغام‌شده‌اند و طبق نتایج حاصل‌شده، شبکه‌های بیزین می‌توانند با دقت متوسط ۹۶,۷۹ درصد از سایر طبقه‌بندی کننده‌ها بهتر عمل کند. رویکرد مورداستفاده در این مقاله به‌کارگیری یادگیری عمیق به‌منظور طبقه‌بندی ترافیک بر اساس ویژگی‌های استخراج‌شده خودکار است که با استفاده از طبقه‌بندی Distiller سامانه قادر به حل مشکل ناهمگونی و غلبه بر محدودیت‌های یادگیری چندوجهی می‌شود. در مقابل مزایای این رویکرد، می‌توان به ترافیک چندوجهی (رمزگذاری شده) اشاره نمود که منجر به حمل اطلاعاتی در شبکه می‌شود که باعث اختلال در شبکه خواهد شد. طبقه‌بندی ترافیک رمزگذاری شده دستگاه‌های تلفن همراه در [۱۸] ارائه‌شده است. این رویکرد شامل چهار مرحله قطعه‌بندی جریان، استخراج ویژگی‌ها، فیلتر کردن ویژگی‌ها و طبقه‌بندی ویژگی‌ها است که با استفاده از شبکه عصبی کانولوشن عمیق انجام می‌شود. در ادامه در مقاله [۱۹] نیز مدلی برای طبقه‌بندی ترافیک رمزگذاری شده توسط یادگیری عمیق چندوظیفه‌ای چندوجهی پیشنهادشده است که می‌تواند ارتباط بین داده‌های ورودی و کلاس‌های هدف را بدون تقسیم مسئله به مشکلات فرعی مانند انتخاب یا استخراج ویژگی‌ها، مدل‌سازی کند. مزایای این رویکرد ارزیابی استحکام طبقه‌بندی کننده ترافیک شبکه، مبتنی بر یادگیری عمیق در برابر ترافیک متخاصم است. از طرفی وجود ترافیک متخاصم در طبقه‌بندی کننده‌های ترافیک شبکه مبتنی بر یادگیری عمیق با استفاده از روش‌های تولید آشفتگی متخاصم جهانی، منجر به‌پیش بینی اشتباه می‌شود. روش‌های یادگیری عمیق همچنین می‌توانند سازوکار پیش‌پردازش را خودکار کنند [۲۰]. تحقیق انجام‌شده در [۲۱]، کاربرد روش‌های مبتنی بر یادگیری عمیق را برای طبقه‌بندی ترافیک بررسی کرده است.

۳. ارائه روش پیشنهادی

در این بخش مراحل روش پیشنهادی شرح داده‌شده است. در رویکرد پیشنهادی، ترکیب روش‌های بدون نظارت و نیمه نظارتی به‌منظور طبقه‌بندی ترافیک استفاده‌شده است. برای این هدف، مجموعه محدودی از جریان‌های ترافیکی برچسب‌گذاری شده برای بهبود عملکرد طبقه‌بندی استفاده می‌شود.

رویکرد پیشنهادی شامل پنج مرحله اصلی است:

۱- توصیف ویژگی‌های جریان‌های ترافیکی بر چسب‌دار و بدون برچسب

۲- استخراج ویژگی‌ها توسط تجزیه‌وتحلیل مؤلفه‌های اصلی[1]

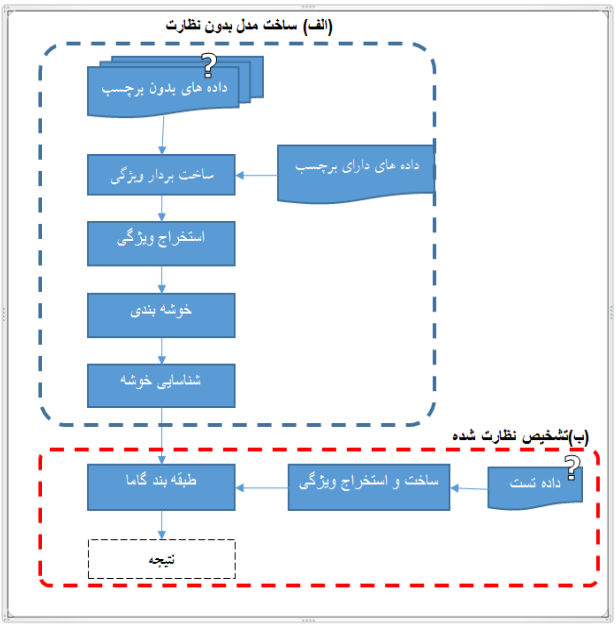
۳- خوشه‌بندی بدون نظارت بردارهای ویژگی با استفاده از DBScan

[1] Principal Component Analysis (PCA)

۴- شناسایی خوشه‌ها با شمارش داده‌های بر چسب‌دار در هر خوشه

۵- طبقه‌بندی جریان‌های ترافیک جدید توسط طبقه بند گاما

شکل (۱) بیانگر مراحل روش پیشنهادی است. مطابق این شکل چارچوب پیشنهادی شامل ترکیبی از داده‌های جریان ترافیک بر چسب‌دار و بدون برچسب به همراه ماژول‌هایی برای ساخت ویژگی، استخراج ویژگی، خوشه‌بندی و طبقه‌بندی است. مجموعه داده برچسب‌گذاری شده شامل مجموعه محدودی از جریان‌های ترافیک و انواع آن‌ها است.

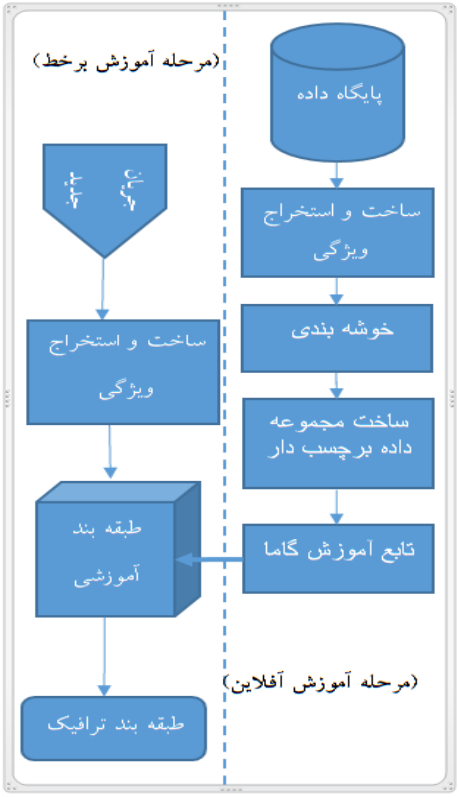


شکل (۱). مراحل رویکرد پیشنهادی

در بخش الف، هر نمونه پایگاه داده توسط یک بردار ویژگی متشکل از ویژگی‌های آماری جریان ترافیک آن توصیف می‌شود. در مرحله بعد ابعاد بردارهای ویژگی ایجادشده برای هر نمونه با استفاده از الگوریتم تجزیه‌وتحلیل مؤلفه‌های اصلی کاهش می‌یابد. هدف از این مرحله، بهبود سرعت پردازش نمونه‌ها در مراحل بعدی و همچنین کاهش اثر مخرب ویژگی‌های نامرتبط بر نتیجه طبقه‌بندی است. در مرحله بعد، نمونه‌های پایگاه داده توسط الگوریتم DBScan خوشه‌بندی می‌شوند. بنابراین، جریان‌های ترافیکی با ویژگی‌های آماری مشابه در خوشه‌های مشابه گروه‌بندی می‌شوند. مرحله بعدی شناسایی هر خوشه خواهد بود، به این معنی که نوع ترافیک صحیح برای هر خوشه تعیین می‌شود. برای انجام این کار، نمونه‌های برچسب گذاری‌شده در هر خوشه شمارش می‌شود و به هر خوشه برچسبی بر اساس اکثریت اختصاص می‌یابد. پس‌ازاین مرحله، داده‌های بدون برچسب، برچسب خوشه خود را خواهند گرفت.

مجموعه داده‌های برچسب‌گذاری شده برای ساخت طبقه‌بندی گاما برای طبقه‌بندی نظارت‌شده (بخش ب) استفاده خواهد شد.

این طبقه بند گاما برای طبقه‌بندی جریان‌های ترافیکی جدید (نمونه‌های آزمایشی) استفاده خواهد شد. بدین ترتیب، در مرحله نظارت‌شده روش پیشنهادی، بردار ویژگی نمونه‌های آزمایشی ساخته‌شده و ویژگی‌ها استخراج می‌شوند (مطابق بخش الف). سپس، ویژگی‌های استخراج‌شده به مدل طبقه بند گامای آموزش‌دیده ارسال می‌شود. مراحل در شکل (۲) نشان داده‌شده است.



شکل (۲). مراحل آموزش آفلاین و طبقه‌بندی برخط جریان ترافیک درروش پیشنهادی

الگوریتم (۱) شبه کد رویکرد پیشنهادی را نشان می‌دهد. در این الگوریتم، هر مرحله از رویکرد پیشنهادی به بخش مربوطه خود در مقاله مرتبط شده است.

الگوریتم (۱): شبه کد روش پیشنهادی

گام ۱: ایجاد بردار ویژگی برای هر نمونه ترافیک (بخش ۴-۲-۱)

گام ۲: استخراج ویژگی‌های مناسب توسط تجزیه‌وتحلیل مؤلفه‌های اصلی (بخش ۴-۲-۲)

گام ۳: اعمال خوشه‌بندی DBScan بر روی ویژگی‌های استخراج‌شده (بخش ۴-۲-۳)

گام ۴: اعمال مراحل زیر برای هر کلاستر (بخش ۴-۲-۴)

۴-۱: شمارش تعداد ترافیک‌ها بر اساس نمونه‌های برچسب دار خوشه

۴-۲: برچسب‌گذاری خوشه بر اساس بیشترین فراوانی برچسب‌های خوشه

۳-۴: اختصاص برچسب خوشه به هر یک از اعضای بدون برچسب خوشه

گام ۵: آموزش طبقه‌بندی کننده گاما را با استفاده از نمونه‌های آموزشی پایگاه داده (بخش ۴-۲-۵)

گام ۶: طبقه‌بندی نمونه‌های آزمایشی با استفاده از مدل طبقه‌بندی کننده گامای آموزش‌دیده

۳-۱. ساخت بردار ویژگی

مرحله اول در رویکرد پیشنهادی، توصیف هر جریان ترافیکی توسط یک بردار ویژگی با طول ثابت است. درروش پیشنهادی، یک مجموعه داده بر چسب‌دار کوچکی از جریان‌های ترافیک را در کنار یک مجموعه بدون برچسب بزرگ در نظر گرفته‌شده است. تفاوت بین این دو مجموعه داده در این است که مجموعه داده برچسب‌گذاری شده دارای یک فیلد اضافی در مورد کلاس‌های ترافیکی است. برچسب‌های این نمونه‌ها به‌صورت دستی تنظیم‌شده است و هر نمونه دارای یکی از برچسب‌های ۱- مرور، ۲- ایمیل، ۳- چت، ۴- استریم، ۵- انتقال فایل، ۶- VoIP، و ۷- P2P است [۲۲].

برای تعریف بردار ویژگی هر نمونه، نمونه‌های بدون برچسب و بر چسب‌دار باهم ترکیب می‌شوند تا یک مجموعه داده بدون برچسب جدید را تشکیل دهند. سپس ویژگی‌های آماری این نمونه‌ها در هر دو جهت جریان، یعنی روبه‌جلو (کلاینت به سرور) و عقب (سرور به کلاینت) استخراج می‌شود. درروش پیشنهادی، دو گروه آماری از ویژگی‌ها برای توصیف یک جریان ترافیک استفاده می‌شود:

۱- ویژگی‌های استخراج‌شده از نقاط زمانی خاص در جریان‌ها. این ویژگی‌های آماری می‌توانند ویژگی‌های مبتنی بر نمونه یک جریان ترافیک را تعریف کنند و از نقاط زمانی مشخص‌شده به‌طور تصادفی در طول اتصال استخراج می‌شوند. در این ویژگی‌ها، نقطه زمانی برای استخراج ویژگی k به‌صورت یک عدد تصادفی با توزیع نمایی و میانگین 2^k (ثانیه) تعیین می‌شود. به‌عنوان‌مثال، اگر k روی ۱۰ تنظیم شود، ۱۰ نقطه تصادفی با توزیع نمایی و مقادیر میانگین 2^1 تا 2^{10} ثانیه انتخاب می‌شود. سپس ویژگی‌های آماری این نقاط تصادفی از جریان‌ها استخراج می‌شود. در ادامه این گروه ویژگی‌های زمانی نامیده می‌شوند.

۲- ویژگی‌هایی که با توجه به ویژگی‌های کلی جریان محاسبه می‌شوند. این ویژگی‌های آماری که ویژگی‌های کلی یک جریان را توصیف می‌کنند، می‌توانند در تشخیص نوع ترافیک شبکه نیز مفید باشند. شایان‌ذکر است که این نوع ویژگی‌ها نیازی به اتمام جریان ندارند. این ویژگی‌ها درواقع ویژگی‌های آماری تا زمان فعلی هستند (و برخلاف ویژگی‌های زمانی، به نقاط زمانی خاصی بستگی ندارند). در ادامه این گروه ویژگی‌های زمانی نامیده می‌شوند. مجموعه ویژگی‌های مورداستفاده برای ساخت بردار ویژگی‌ها درروش پیشنهادی در جدول (۱) توضیح داده‌شده است.

جدول (۱). ویژگی‌های آماری بکار رفته درروش پیشنهادی

ویژگی	تعداد ویژگی	نوع ویژگی	توصیف
بسته‌های نمونه‌برداری شده	$2 \times k$	زمانی	اندازه بسته‌های منتقل‌شده تا نقطه زمانی فعلی در هر جهت جریان
بایت‌های نمونه‌برداری شده	$2 \times k$	زمانی	بایت‌ها به نقطه زمانی فعلی در هر جهت جریان منتقل می‌شوند
اندازه نمونه‌برداری	$4 \times k$	زمانی	میانگین و انحراف معیار اندازه بسته تا نقطه زمانی فعلی در هر جهت جریان
زمان نمونه‌برداری	$4 \times k$	زمانی	میانگین و انحراف معیار زمان بین بسته تا نقطه زمانی فعلی در هر جهت جریان
بسته‌های جریان	2	جریان	تعداد بسته‌های منتقل‌شده در جهت جریان تا زمان فعلی
بایت‌های جریان	2	جریان	کل بایت‌های منتقل‌شده در هر جهت جریان تا زمان فعلی
اندازه جریان	6	جریان	میانگین، حداقل و حداکثر اندازه بسته در هر جهت جریان تا زمان فعلی
جریان زمانی	6	جریان	میانگین، حداقل و حداکثر زمان بین بسته‌ها در هر جهت جریان تا زمان جاری
طول بردار ویژگی			$16 + 10k$

۳-۲. استخراج ویژگی با استفاده از تجزیه‌وتحلیل مؤلفه‌های اصلی

برخی از ویژگی‌های ساخته‌شده ممکن است به نوع ترافیک شبکه مرتبط نباشد. این ویژگی‌ها علاوه بر کاهش سرعت پردازش نمونه، باعث کاهش دقت طبقه‌بندی نیز خواهد شد. بنابراین، در مرحله دوم رویکرد پیشنهادی از تجزیه‌وتحلیل مؤلفه‌های اصلی برای استخراج ویژگی‌ها و کاهش اثر مخرب ویژگی‌های نامربوط به انواع ترافیک استفاده می‌شود.

با در نظر داشتن ماتریس داده X^T با مقدار میانگین تجربی صفر، که هر سطر آن نشان‌دهنده یک نمونه و هر ستون آن مشخص‌کننده یک ویژگی است، تحلیل مؤلفه‌های اصلی را می‌توان به‌صورت زیر تعریف نمود [۲۳]:

$$Y^T = X^T W = VS \qquad (۱)$$

در رابطه (۱)، VSW^T نشان‌دهنده تجزیه مقدارهای منفرد ماتریس X^T است. بر اساس تعریف اولیه تحلیل مؤلفه‌های اصلی، هدف از این الگوریتم تبدیل ماتریس داده X با ابعاد M به ماتریس داده Y با ابعاد L است. بنابراین فرض می‌شود که ماتریس X از بردارهای ستونی $X_1, \dots, X_N$ تشکیل‌شده است که هریک از این بردارهای ستونی معرف یک ویژگی در X است. بدین ترتیب، ماتریس داده‌های X دارای ابعاد $M \times N$ خواهد بود. با در نظر

داشتن این ساختار، مراحل استخراج مؤلفه‌های اصلی X بر اساس ماتریس کوواریانس به‌صورت گام‌های زیر خواهد بود:

الف) محاسبه میانگین تجربی داده و نرمال‌سازی آن: میانگین تجربی داده، برداری است که به‌صورت رابطه زیر محاسبه می‌شود [۲۳]:

$$u[m]=\frac{1}{N}\sum_{i=1}^{N}X[m,i] \quad (۲)$$

میانگین تجربی داده در رابطه بالا، به‌صورت مشخص بر روی سطرهای ماتریس اعمال می‌شود. پس‌ازآن ماتریس فاصله داده با میانگین تجربی به‌صورت زیر محاسبه می‌شود [۲۳]:

$$B=X-uh \quad (۳)$$

در رابطه بالا، h یک بردار تماماً یک با اندازه $N\times 1$ است.

ب) محاسبه ماتریس کوواریانس: در گام دوم، ماتریس کوواریانس C دارای $M\times M$ از طریق رابطه زیر محاسبه می‌شود [۲۳]:

$$C=E[B\otimes B]=E[B.B^{*}]=\frac{1}{N}B.B^{*} \quad (۴)$$

در رابطه فوق، E مشخص‌کننده میانگین حسابی است. همچنین عملگر $\otimes$ نشان‌دهنده ضرب خارجی ماتریس بوده و B^{*} ترانهاده مزدوج ماتریس B است.

ج) تخمین مقادیر ویژه‌ماتریس کوواریانس و جایگشت بردارهای ویژه: در این گام، مقادیر ویژه و بردارهای ویژه در ماتریس کوواریانس C، با استفاده از رابطه زیر محاسبه می‌شوند [۲۳]:

$$V^{-1}CV=D \quad (۵)$$

در رابطه فوق؛ V ماتریس بردارهای ویژه بوده و D یک ماتریس قطری است که درایه‌های واقع در قطر اصلی آن مقادیر ویژه‌ماتریس هستند. هر مقدار ویژه در این ماتریس با یک بردار ویژه متناظر است. به عبارتی ماتریس V دارای ابعاد $M\times M$ بوده و ستون‌های آن بردارهای ویژه هستند؛ به صورتی که بردار ویژه V_qدر ستون q ام این ماتریس قرار دارد و مقدار ویژهq ام ماتریس D متناظر با آن است $(\square_q=D_{q,q})$. جاگشت بردارهای ویژه‌بر اساس بزرگی مقادیر ویژه متناظر با آن‌ها صورت می‌پذیرد. در این حالت بر اساس ترتیب نزولی مقادیر ویژه، بردارهای ویژه مرتب‌سازی می‌شوند.

د) انتخاب یک زیرمجموعه از بردارهای ویژه به‌عنوان مجموعه پایه: انتخاب یک زیرمجموعه از بردارهای ویژه از طریق تحلیل مقادیر ویژه صورت می‌پذیرد. زیرمجموعه نهایی بر اساس جایگشت حاصل از مرحله قبل به‌صورت $V_1,\dots,V_l$ تعیین می‌شود. در این گام می‌توان از انرژی تجمعی استفاده نمود که بر اساس آن [۲۳]:

$$g[m]=\sum_{q=l}^{m}\lambda_q \quad (۶)$$

انتخاب مقدار l باید به صورتی باشد که دارای مقدار کمینه بوده و درعین‌حال g از مقدار قابل قبولی برخوردار باشد. مثلاً می‌توان حداقل l را به صورتی انتخاب نمود که $g[m=l]\le 90\%$.

هـ) انتقال داده به فضای مختصات جدید: بدین منظور باید ابتدا تبدیلات زیر را اعمال نمود:

- ماتریس $s_{M,l}$ که نشان‌دهنده انحراف معیار مجموعه داده است، به‌صورت $s[i]=\sqrt{C[i,i]}$ محاسبه می‌گردد.
- داده به‌صورت کسر $Z=\frac{B}{s}$ تبدیل می‌شود.
- داده‌ها بر اساس رابطه زیر به فضای جدید نگاشت می‌یابند [۲۳]:

$$Y=W^{*}.ZY=W^{*}.Z \quad (۷)$$

پس از استخراج ویژگی‌های بارز بر اساس روند فوق، در گام بعدی از راهبرد خوشه‌بندی به‌منظور مقابله با مسئله محدودیت داده-های دارای برچسب استفاده می‌شود.

۳-۳- خوشه‌بندی بدون نظارت توسط DBScan

در این مرحله از رویکرد پیشنهادی، از روش *DBScan* برای خوشه‌بندی بدون نظارت ویژگی‌های استخراج‌شده استفاده می‌شود که نمونه‌های پرت از داده‌های معتبر تفکیک‌شده و بدین ترتیب می‌توان نمونه‌های پرت را در فرآیند برچسب‌گذاری نمونه‌ها فیلتر نمود. الگوریتم *DBScan* برای خوشه‌بندی نمونه‌ها بر اساس هر شاخص از دو پارامتر آستانه همسایگی ϵ و آستانه تعداد نقاط P بهره می‌گیرد و شامل گام‌های زیر است [۲۴]: **گام ۱**: مجموعه‌ی خوشه‌های C را به‌صورت تهی مقداردهی کن.

گام ۲: برای هر نمونه مشاهده نشده x از مجموعه داده ورودی گام‌های زیر را تکرار کن:
گام ۳: داده x را به مجموعه داده‌های مشاهده‌شده اضافه کن و نمونه‌های همسایه x را بر اساس آستانه همسایگی ϵ در N ذخیره کن.
گام ۴: اگر $|N|<P$، آنگاه x را به‌عنوان نویز نشانه‌گذاری کن و به گام ۲ برو؛ در غیر این صورت گام بعدی را اجرا کن.
گام ۵: درصورتی‌که $|N|\ge P$، آنگاه x را به مجموعه C اضافه کن. سپس، برای هر نمونه مانند $x'\in N$ گام‌های زیر را اجرا کن:
گام ۶: نمونه x' را از N حذف کن. سپس درصورتی‌که x' مشاهده‌شده است، به گام ۹ برو و در غیر این صورت گام‌های زیر را تکرار کن:
گام ۷: داده x' را به مجموعه مشاهده‌شده اضافه کن و نمونه‌های همسایه x' را بر اساس آستانه همسایگی ϵ در N'ذخیره کن.
گام ۸: درصورتی‌که $|N|\ge P$، آنگاه دو مجموعه N و N' را به‌صورت مجموعه N ادغام کن $(N=N\cup N')$.
گام ۹: اگر برای x' خوشه‌ای تعیین نشده است، آن را در خوشه‌ی C قرار بده و به گام ۵ برو

پس از دسته‌بندی نمونه‌ها، تمامی نمونه‌های نویز نادیده گرفته‌شده و برای سایر نمونه‌های خوشه‌بندی‌شده، عمل شناسایی دسته و انتشار برچسب صورت می‌پذیرد که در ادامه تشریح شده است.

۳-۴. شناسایی خوشه‌ها

پس از خوشه‌بندی نمونه‌های مجموعه داده ورودی، برچسب هر خوشه حاصل باید بر اساس نوع ترافیک تعیین شود. برای انجام این کار، از نتایج خوشه‌بندی نمونه‌های برچسب‌گذاری شده استفاده خواهد شد. به‌این‌ترتیب توزیع نمونه‌های بر چسب‌دار درنتیجه خوشه‌بندی بررسی‌شده و بیشترین تعداد برچسب در هر خوشه شمارش می‌شود. بنابراین، برچسب خوشه، برچسبی است که بیشترین فراوانی نمونه‌های بر چسب‌دار را در آن دارد. اگر بیش از یک برچسب بیشترین فراوانی را در خوشه داشته باشد، یک برچسب به‌طور تصادفی از بین آن‌ها انتخاب می‌شود. پس از انتخاب برچسب Li برای یک خوشه Ci، تمام نمونه‌های بدون برچسب در Ci به‌عنوان Li برچسب‌گذاری می‌شوند. نتیجه این مرحله یک مجموعه داده کامل دارای برچسب خواهد بود. این مجموعه داده برای طبقه‌بندی نیمه نظارتی در مرحله بعدی روش پیشنهادی استفاده خواهد شد.

۳-۵. طبقه‌بندی جریان‌های ترافیکی توسط طبقه بند گاما

در آخرین گام از روش پیشنهادی، از طبقه بند گاما به‌منظور ساخت مدل طبقه‌بندی جریان‌های ترافیکی استفاده خواهد شد. این مدل طبقه‌بندی، یک روش نظارت‌شده است که نام آن از عملگر شباهتی که استفاده می‌کند، عملگر گاما گرفته‌شده است. تعمیم‌یافته این عملگر دو بردار دودویی x و y و یک عدد صحیح مثبت مانند θ را به‌عنوان ورودی دریافت کرده و اگر هر دو بردار مشابه باشد مقدار ۱ و در غیر این صورت مقدار ۰ را برمی‌گرداند. عملگر گاما از عملگرهای دیگری (مانند α، β و ...) استفاده می‌کند که در ادامه به تعریف آن‌ها پرداخته می‌شود.

تعریف ۱) عملگرهای آلفا و بتا: با توجه به مجموعه $A = \{0,1\}$ و $B = \{0,1,3\}$، عملگرهای آلفا (α) و بتا (β) به‌صورت جدول (۲) خواهند بود.

جدول (۲). عملگرهای آلفا و بتا در طبقه بند گاما

α:AxA->B		
x	y	α(x,y)
0	0	1
0	1	0
1	0	2
1	1	1
β:BxA->A		
x	y	β (x,y)
0	0	0
0	1	0
1	0	0
1	1	1
2	0	1
2	1	1

تعریف ۲) اعمال عملگر آلفا بر بردارها: با فرض اینکه $x,y \in A^n$ بردارهای ستونی ورودی باشند. خروجی $\alpha(x,y)$ بردار n بعدی است که مؤلفه‌های آن به شرح زیر محاسبه می‌شود:

$$\alpha(x,y)_i = \alpha(x_i,y_i)\ \alpha(x,y)_i = \alpha(x_i,y_i) \quad (۸)$$

تعریف ۳) عملگر u_β: با فرض اینکه الگوی باینری $x \in A^n$ به‌عنوان ورودی، این عملگر عدد صحیح غیر منفی زیر را به‌عنوان خروجی تولید کرده و به‌صورت زیر محاسبه می‌شود:

$$u_\beta(x) = \sum_{i=1}^{n} \beta(x_i,x_i) \quad (۹)$$

تعریف ۴) عملگر هرس: با فرض اینکه $x \in A^n$ و $y \in A^m$ و $n < m$، دو بردار باینری باشند. سپس y هرس شده توسط x، به‌صورت $y\|x$ نمایش داده‌شده و برداری دودویی و n بعدی است که مؤلفه‌های آن به‌صورت زیر محاسبه‌شده است.

$$(y\|_x)_i = y_{i+m-n}, \quad (i = 1,2,\dots,n) \quad (۱۰)$$

عملگر گاما به یک بردار باینری به‌عنوان ورودی نیاز دارد. برای مقابله با بردارهای حقیقی یا عدد صحیح، روشی برای نمایش این بردارها به‌صورت باینری موردنیاز است. در این کار، معمولاً از کد جانسون-موبیوس اصلاح‌شده در [۲۵] استفاده می‌شود. با توجه به اینکه جزئیات کامل این الگوریتم در[۲۵] موردبحث قرارگرفته است، لذا از پرداختن به روند آن در این صرف‌نظر می‌شود.

تعریف ۵) عملگر گاما: عملگر شباهت گاما دو الگوی باینری مانند $x \in A^n$ و $y \in A^m$ و $n \le m$ و یک عدد صحیح غیرمنفی مانند θ را به‌عنوان ورودی گرفته و به ازای هریک از دو حالت زیر، یک خروجی دودویی تولید می‌کند:

حالت ۱: اگر $n = m$، خروجی مطابق رابطه زیر محاسبه می‌شود [۲۶]:

$$\gamma(x,y,\theta) = \begin{cases} 1, & if\ m - u_\beta[\alpha(x,y) mod\ 2] \le \theta \\ 0, & otherwise \end{cases} \quad (۱۱)$$

در رابطه فوق، mod نشانگر عملکرد باقیمانده تقسیم است.

حالت ۲: اگر $n < m$، خروجی، بجای y با استفاده از $y\|_x$ به‌صورت رابطه زیر محاسبه می‌شود [۲۶]:

$$\gamma(x,y,\theta) = \begin{cases} 1, & if\ m - u_\beta[\alpha(x,y\|_x) mod\ 2] \le \theta \\ 0, & otherwise \end{cases} \quad (۱۲)$$

با در نظر داشتن عملگرهای فوق، در ادامه به تشریح عملکرد طبقه بند گاما پرداخته می‌شود. با فرض اینکه الگوهای پایه به‌صورت $\{(x^u,y^u)|u = 1,2,\dots,p\}$ دارای کاردینالیته p بوده و الگوی آزمون به‌صورت $\tilde{x} \in R^n$ توصیف‌شده باشد. در این حالت، طبقه بند گاما از طریق گام‌های محاسباتی زیر نمونه آزمایشی $\tilde{x}$ را طبقه‌بندی می‌کند:

گام ۱) مجموعه پایه آموزشی را با استفاده از کد جانسون-موبیوس اصلاح‌شده به فرم دودویی تبدیل نموده تا برای هر مؤلفه در بردارهای n-بعدی مجموعه پایه‌یک مقدار e_m به‌صورت زیر محاسبه شود [۲۶]:

$$e_m(j) = \max_{i=1}^{p}(x_j^i), \qquad \forall j \in \{1,2,\dots,n\} \quad (۱۳)$$

گام ۲) پارامتر توقف به‌صورت $\rho = \max_{j=1}^{n}[e_m(j)]$ محاسبه می‌شود.

گام ۳) الگوی آزمون نیز با استفاده از کد جانسون-موبیوس اصلاح‌شده و با همان پارامترهایی که برای کدگذاری مجموعه اصلی استفاده‌شده است، کدگذاری خواهد شد. اگر هر e_j به‌دست‌آمده از $e_m(j)$ متناظرش بیشتر باشد، آن را با استفاده از تعداد بیشتری بیت، کد می‌شود.

گام ۴) شاخص همه الگوهای پایه را به دو شاخص تبدیل می‌شود: یکی برای کلاس آن‌ها و دیگری برای موقعیت آن‌ها در کلاس.

گام ۵) مقدار اولیه پارامتر θ برابر صفر تنظیم می‌شود.

گام ۶) اگر $\theta = 0$، آنگاه با محاسبه $\gamma(x_j^u, \tilde{x}_j, 0)$ بررسی می‌شود که آیا $\tilde{x}$ یک الگوی پایه است و سپس وزن انباشته اولیه c_u^0 برای هر الگوی پایه به‌صورت زیر محاسبه می‌شود [۲۶]:

$$c_u^0 = \sum_{i=1}^{n} \gamma(x_j^u, \tilde{x}_j, 0), \quad for\ u = 1,2,\dots,p \qquad (۱۴)$$

اگر یک مقدار بیشینه یکتا برابر با n وجود دارد؛ آنگاه کلاس متناظر با این مقدار بیشینه به الگوی آزمون اختصاص داده می‌شود [۲۶]:

$$\tilde{y} = y^{\sigma} \quad such\ that \quad \max_{i=1}^{p} c_i^0 = c_\sigma^0 = n \qquad (۱۵)$$

گام ۷) مقدار $\gamma(x_j^{i\omega}, \tilde{x}_j, \theta)$ را برای هر مؤلفه الگوهای پایه محاسبه می‌شود.

گام ۸) مقدار مجموع وزن‌دار c_i را برای هر کلاس با استفاده از رابطه زیر محاسبه خواهد شد [۲۶]:

$$c_i = \frac{\sum_{\omega=1}^{k_i}\sum_{j=1}^{n}\gamma(x_j^{i\omega}, \tilde{x}_j, \theta)}{k_i} \qquad (۱۶)$$

در رابطه فوق، k_i نشان‌دهنده کاردینالیته مجموعه پایه کلاس i است.

گام ۹) اگر بیش از یک مقدار بیشینه در بین c_i های مختلف وجود داشته باشد، مقدار θ را یک واحد افزایش داده و گام‌های ۷ و ۸ را تا زمانی که تنها یک مقدار بیشینه وجود داشته باشد؛ یا شرط خاتمه $\theta > \rho$ برآورده گردد؛ تکرار خواهد شد.

گام ۱۰) اگر یک مقدار بیشینه یکتا در بین c_i های مختلف وجود داشته باشد؛ آنگاه کلاس مربوط به الگوی آزمایشی $\tilde{x}$ با استفاده از رابطه زیر محاسبه می‌شود [۲۶]:

$$\tilde{y} = y^{j} \quad such\ that \quad max\, c_i = c_j \qquad (۱۷)$$

گام ۱۱) در غیر این صورت الگوی $\tilde{x}$ به کلاس دارای اولین مقدار بیشینه اختصاص خواهد یافت.

۴. مقایسه و ارزیابی

در این مطالعه، از نمونه‌های مجموعه داده ISCXVPN2016 [۲۷] برای ارزیابی عملکرد روش پیشنهادی استفاده شد و شامل مجموعه‌ای از داده‌های ترافیکی واقعی است که ازنظر کمیت و تنوع غنی است. نمونه‌های این مجموعه داده بر اساس سناریوی آلیس و باب ثبت‌شده است. برای انجام این کار، دو حساب کاربری (آلیس و باب) ایجادشده و جریان ترافیک این کاربران در استفاده از سرویس‌هایی مانند اسکایپ، یوتیوب، Hangout و غیره ضبط‌شده است. برنامه Wireshark برای ضبط جریان‌های ترافیک از دستگاه‌های شبکه استفاده‌شده است و هر داده گرفته‌شده به‌عنوان یک فایل pcap. یا pcapng. در مجموعه داده ذخیره می‌شود. در ادامه انواع ترافیک و برنامه‌های کاربردی مورداستفاده در مجموعه داده ISCXVPN2016 برای هر نوع ترافیک تشریح خواهد شد.

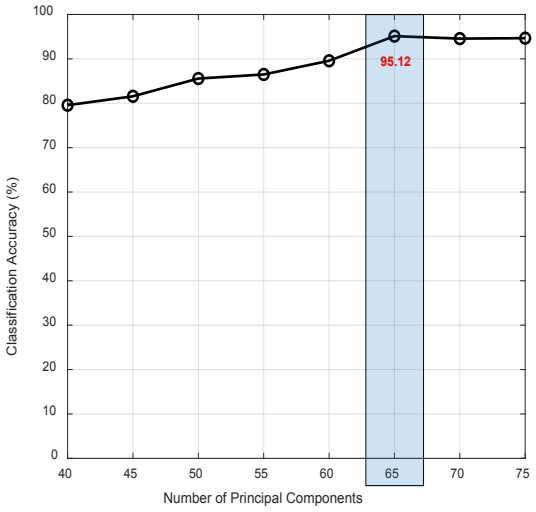
ISCXVPN2016 حاوی داده‌های ترافیکی ضبط‌شده برای ۷ کلاس ترافیک در شبکه‌های VPN و اتصالات غیر VPN است. بنابراین، تعداد کل کلاس‌های منحصربه‌فرد در این مجموعه داده برابر با ۱۴ خواهد بود: Email، VPN-Email، Chat، VPN-Chat و غیره. در این مطالعه، از ۴۷۳۹ جریان از مجموعه داده‌های ISCXVPN2016 استفاده‌شده است که ۶۰ جریان از هر کلاس برچسب‌گذاری شده است. بنابراین تعداد کل جریان‌های برچسب‌گذاری شده ۴۲۰ است. لیست انواع ترافیک این مجموعه داده عبارت‌اند از:

1. Browsing
2. Email
3. Chat
4. Streaming
5. File Transfer
6. VoIP
7. P2P

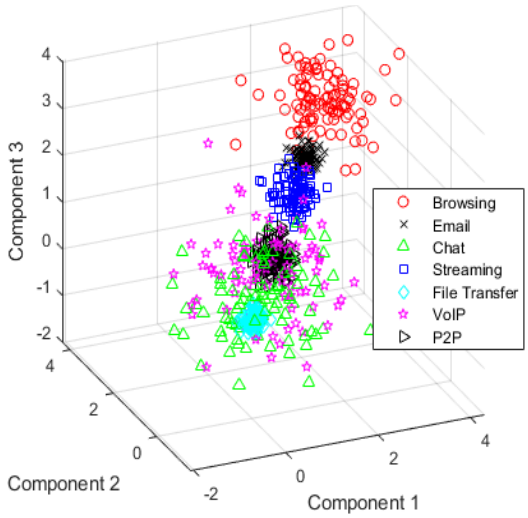
برای ثبت این نوع ترافیک‌ها از Wireshark و tcpdump استفاده‌شده است و ۱۵۰ فایل. pcap/.pcapng با حجم کل ۲۸ گیگابایت ضبط‌شده است.

روش پیشنهادی در محیط MALTAB 2020a پیاده‌سازی شده است. از tshark (بخشی از برنامه Wireshark) برای ساخت بردارهای ویژگی فایل‌های مجموعه استفاده‌شده است. با توجه به تعداد محدود پارامترهای مدل پیشنهادی، به‌منظور پیکربندی این پارامترها از راهبرد جستجوی جامع و بررسی تجربی نتایج استفاده‌شده است. در این روند، ترکیبات مختلف مقادیر پارامترها موردبررسی قرارگرفته و پیکربندی که منجر به بیشینه نمودن عملکرد مدل گردد، انتخاب‌شده است. این مجموعه پارامترها عبارت‌اند از: مقدار پارامتر k در گام توصیف ویژگی (که مشخص‌کننده طول بردار ویژگی‌های مبتنی بر زمان استخراج‌شده از جریان ترافیکی است)، ابعاد ویژگی استخراج‌شده توسط الگوریتم تجزیه‌وتحلیل مؤلفه‌های اصلی و تعداد تکرار درروند اعتبار سنجی متقاطع. در طی فرآیند ساخت بردار ویژگی، مقدار k برابر ۱۰ در نظر گرفته‌شده است که ۴۷۳۹ نمونه و بردار ویژگی به طول ۱۶+۱۰×۱۰=۱۱۶ برای توصیف هر یک به‌دست‌آمده است. این مقدار منجر به شکل‌گیری بردارهای ویژگی با غنای اطلاعاتی

کافی برای توصیف خصوصیات مرتبط با نوع جریان‌های ترافیکی خواهد شد. از طرفی، بررسی مقادیر مختلف پارامتر تعداد تکرار (به ازای مقادیر ۵، ۱۰، ۲۰ و ۳۰ تکرار) در اعتبارسنجی متقاطع اختلاف معناداری در نتایج نهایی را نشان نداد. لذا آزمایش‌های این بخش‌بر اساس ۱۰ بار تکرار در رویکرد اعتبارسنجی متقاطع انجام‌شده است. بعلاوه، راهبرد جستجوی جامع نشان داد که درروش پیشنهادی، کاهش ابعاد پایگاه داده به ۶۵ ویژگی با استفاده از الگوریتم تجزیه‌وتحلیل مؤلفه‌های اصلی بهترین عملکرد را درپی دارد. برای این کار، دقت روش پیشنهادی برای مقادیر مختلف این پارامتر بررسی‌شده است.



(الف)

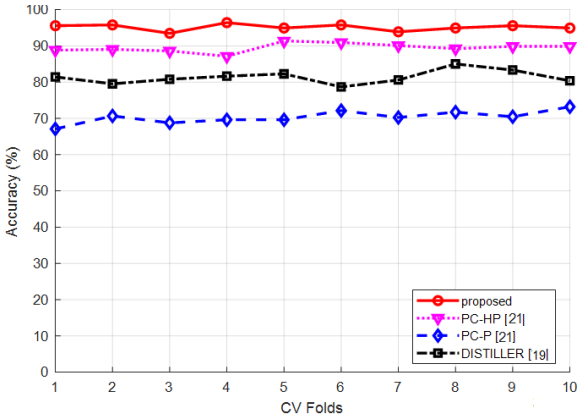


(ب)

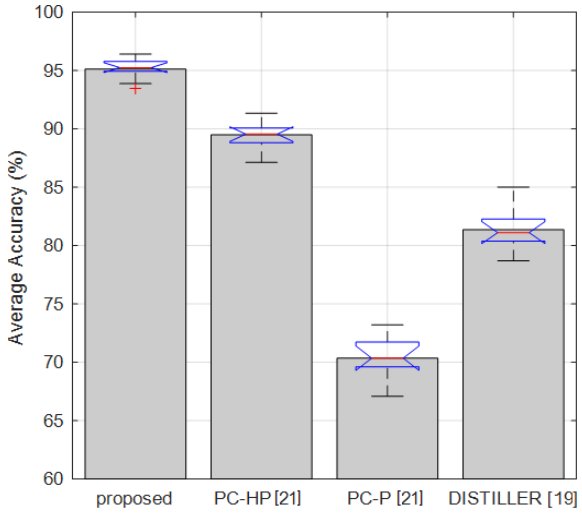
شکل ۳. (الف) دقت طبقه‌بندی روش پیشنهادی به ازای تغییرات تعداد ویژگی و (ب) توزیع نمونه‌ها در دسته‌های هدف بر اساس سه مؤلفه برتر استخراج‌شده توسط تجزیه‌وتحلیل مؤلفه‌های اصلی

شکل(۳). الف، نتایج به‌دست‌آمده را نشان می‌دهد. همچنین در شکل (۳).ب، توزیع نمونه‌ها در دسته‌های هدف بر اساس سه مؤلفه برتر استخراج‌شده توسط الگوریتم تجزیه‌وتحلیل مؤلفه‌های اصلی نمایش داده‌شده است.

شکل (۴).الف نتایج طبقه‌بندی روش پیشنهادی را ازنظر دقت، در طول ۱۰ با تکرار آزمایش‌ها نشان می‌دهد. در این شکل، عملکرد مدل پیشنهادی در طی تکرارهای مختلف با روش‌های PC-HP [۲۱]، PC-P [۲۱] و DISTILLER [۱۹] مقایسه شده است. با توجه به نتایج این شکل، دقت روش پیشنهادی در طول این تکرارها، برتر از روش‌های مقایسه شده و میانگین دقت آن در طبقه‌بندی تمام نمونه‌های داده ۹۵,۱۲ درصد است که توانایی بالای کلی آن را در شناسایی انواع ترافیک نشان می‌دهد.



(الف)



(ب)

شکل (۴). مقایسه دقت طبقه‌بندی (الف) تغییرات دقت در طول ۱۰ بار تکرار آزمایش‌ها و (ب) نمودار جعبه‌ای تغییرات دقت

مطابق شکل (۳). الف، حداکثر مقدار دقت زمانی به دست می‌آید که تعداد ویژگی‌های استخراج‌شده برابر با ۶۵ باشد. طبق نتایج تجربی، استفاده از مقادیر بالاتر برای تعداد ویژگی‌های استخراج‌شده، طبقه‌بندی روش پیشنهادی را بهبود نمی‌بخشد. بنابراین، در ادامه این آزمایش‌ها، روش پیشنهادی با استفاده از این مقادیر بهینه ارزیابی خواهد شد. از طرفی، شکل (۳).ب به‌خوبی نشان می‌دهد که با بهره‌گیری از مؤلفه‌های اصلی استخراج‌شده می‌توان نمونه‌های متعلق به دسته‌های مختلف را از هم تفکیک نمود. شکل (۳).الف نشان می‌دهد که با در نظر گرفتن مؤلفه‌های اصلی بیشتر این قابلیت تفکیک نیز تقویت

می‌شود و براین اساس می‌توان کارایی راهبرد تجزیه‌وتحلیل مؤلفه‌های اصلی در توصیف کارآمد ویژگی‌ها را نتیجه گرفت.

از طرفی، در شکل(۴).ب، مقادیر میانگین دقت به‌دست‌آمده توسط روش‌های مقایسه شده در کنار بازه تغییرات دقت آن‌ها با استفاده از یک نمودار جعبه‌ای نمایش داده‌شده است. در این نمودار، تغییرات دقت در طی تکرارهای مختلف به چارک‌های مختلف تقسیم‌بندی شده و مقدار میانه دقت به‌صورت نقطه میانی هر جعبه نمایش داده‌شده است. بررسی این نمودار نشان می‌دهد که بازه تغییرات دقت روش پیشنهادی در طی ۱۰ تکرار اعتبارسنجی متقاطع به بازه ۹۳,۴۶ تا ۹۶,۴۱ محدوده شده است که محدوده فشرده‌تری را در مقایسه با روش‌های مقایسه شده از خود نشان می‌دهد. از طرفی، مقادیر میانه و میانگین دقت به‌دست‌آمده توسط روش پیشنهادی در سطوح بالاتری نسبت به روش‌های مقایسه شده قرار دارند که تائید کننده قابلیت اعتماد بالاتر مدل پیشنهادی در پیش‌بینی صحیح نوع جریان‌های ترافیکی شبکه است.

در شکل‌های (۵).الف تا (۵).ج، مقادیر میانگین F-measure، پوشش و صحت در طول تکرارهای اعتبار سنجی متقاطع نشان داده‌شده است. این نتایج برای روش پیشنهادی در مقایسه با روش‌های[21] PC-HP، [21] PC-P و[19] DISTILLER است.

ازآنجایی‌که طبقه‌بندی ترافیک یک مسئله چند کلاسه است، ابتدا مقادیر دقت، پوشش و F-Measure برای هر کلاس به‌طور جداگانه محاسبه‌شده و سپس میانگین ارزش این معیارها محاسبه‌شده است. در هنگام محاسبه این معیارها برای هر کلاس، آن طبقه مثبت و سایر طبقات منفی در نظر گرفته‌شده است. در این مورد، معیار صحت می‌تواند دقت الگوریتم را برای هر کلاس به‌طور جداگانه توصیف کند:

$$Precision = \frac{TP}{TP + FP} \times 100 \qquad (۱۸)$$

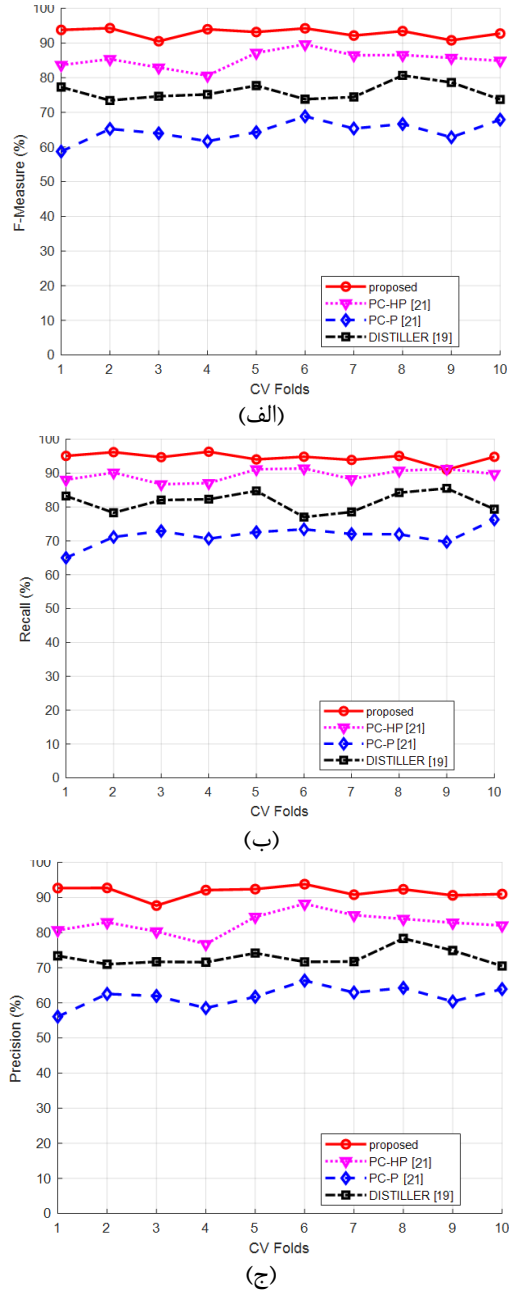
در این معادله، TP تعداد نمونه‌های طبقه موردمطالعه را که به‌درستی شناسایی‌شده‌اند (مثبت صحیح)، نشان می‌دهد و FP به نمونه‌های متعلق به سایر طبقات که در کلاس مثبت طبقه‌بندی‌شده‌اند، اشاره می‌کند (مثبت کاذب). معیار پوشش نسبت نمونه‌های مثبتی را نشان می‌دهد که الگوریتم می‌تواند به‌درستی شناسایی کند:

$$recall = \frac{TP}{TP + FN} \times 100 \qquad (۱۹)$$

در این معادله، FN (منفی کاذب) به نمونه‌های متعلق به کلاس هدف اشاره دارد که در کلاس‌های دیگر طبقه‌بندی‌شده‌اند. با استفاده از مقادیر صحت و پوشش، میانگین هماهنگ این معیارها را می‌توان به‌صورت F-Measure محاسبه کرد:

$$F - Measure = \frac{2 \times Precision \times Recal}{Precision + Recal} \qquad (۲۰)$$

مطابق نمودارهای ترسیم‌شده در شکل (۵)، مقادیر صحت، پوشش و F-Measure درروش پیشنهادی پس از ۱۰ بار تکرار آزمایش‌ها به ترتیب برابر با ۹۱,۰۷، ۹۴,۵۹ و ۹۳,۰۵ است. این مقادیر ارتقای قابل‌توجهی را در مقایسه با روش‌های مقایسه شده نشان می‌دهد. مقادیر بالاتر صحت روش پیشنهادی در مقایسه با سایر روش‌ها نشان می‌دهد که به‌صورت میانگین و برای هر دسته هدف، نسبت بالاتری از برچسب‌های پیش‌بینی‌شده برای آن دسته صحیح بوده است.



شکل (۵). مقادیر میانگین (الف) F-Measure، (ب) پوشش و (ج) صحت، در ۱۰ بار تکرار آزمایش‌ها در روش‌های ارزیابی‌شده

این ویژگی تائید می‌کند که به‌صورت کلی برچسب‌های پیش‌بینی‌شده توسط روش پیشنهادی از احتمال صحت بالاتری برخوردارند. در مقابل، بالاتر بودن مقدار پوشش روش پیشنهادی در طی تکرارهای مختلف نشان می‌دهد که این راهبرد قادر بوده

است تا نرخ بالاتری از نمونه‌های متعلق به هر دسته ترافیکی را به‌درستی شناسایی کند. به بیانی دیگر، انواع جریان‌های ترافیکی با احتمال کمتری در سایر دسته‌ها طبقه‌بندی خواهند شد.

درنهایت، سطوح بالاتر F-measure کسب‌شده توسط روش پیشنهادی تائید می‌کند که این راهکار از جنبه‌های عملکردی بررسی‌شده دارای برتری کلی نسبت به روش‌های مقایسه شده پیشین است. تجزیه‌وتحلیل دقیق‌تر عملکرد مدل پیشنهادی از طریق بررسی این معیارها به تفکیک هر دسته ممکن خواهد بود.

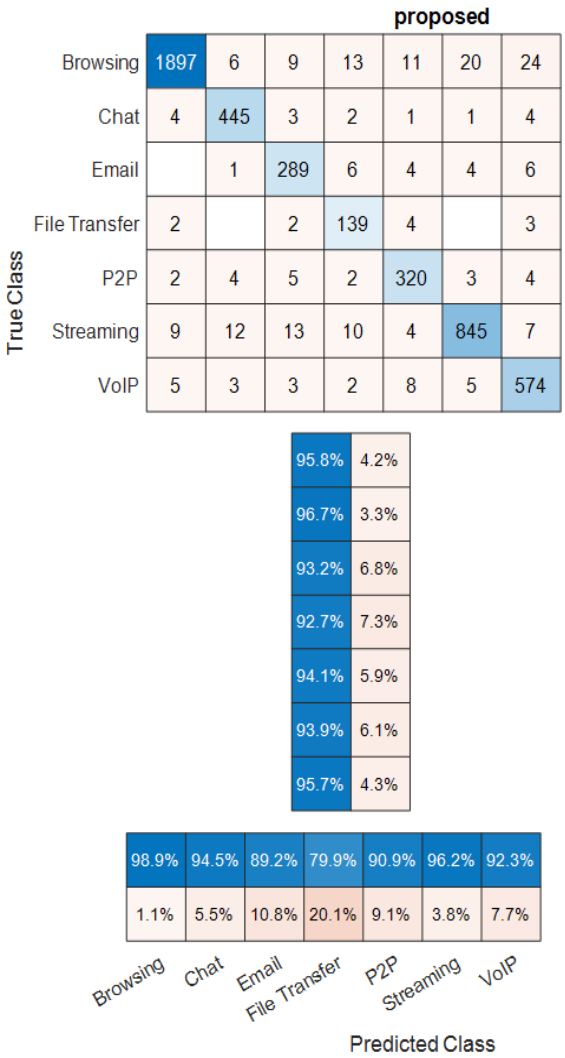
جدول(۳) نتایج ارزیابی الگوریتم پیشنهادی برای طبقه‌بندی جریان‌های ترافیکی مجموعه داده ISCXVPN2016 را خلاصه می‌کند. در این جدول معیارهای موردمطالعه برای هر کلاس ترافیکی به تفکیک محاسبه‌شده است. همچنین، کیفیت طبقه‌بندی رویکرد پیشنهادی باکارهای قبلی مقایسه شده است. روش‌های مقایسه شده نیز از همین مجموعه داده برای ارزیابی رویکردهای خود استفاده کرده‌اند.

مطابق این جدول، روش پیشنهادی بهتر از الگوریتم‌های مقایسه شده عمل می‌کند. بهترین مقدار دقت درروش پیشنهادی، متعلق به کلاس‌های ترافیک «Browsing»، «Chat» و «streaming» است. این بدان معنی است که خروجی‌های ساخته‌شده با روش پیشنهادی در این کلاس‌ها با احتمال بالایی درست هستند. همچنین، بهترین مقدار پوشش برای رویکرد پیشنهادی، متعلق به همین کلاس‌ها است. این بدان معنی است که روش پیشنهادی می‌تواند اغلب این نمونه‌های ترافیک را به‌درستی شناسایی کند.

جدول (۳). مقایسه کیفیت طبقه‌بندی روش پیشنهادی

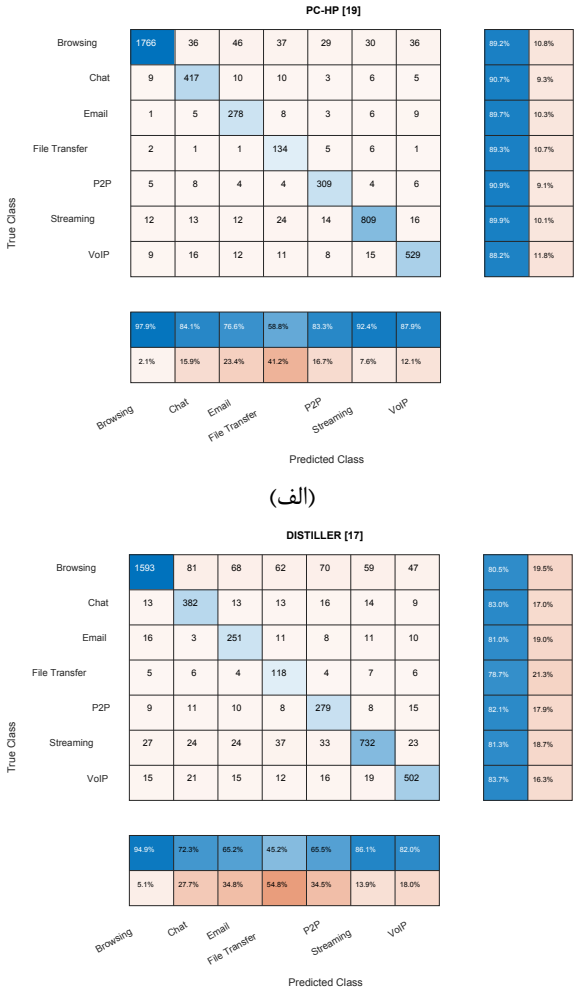
		Browsing	**Email**	**Chat**	**Streaming**
روش پیشنهادی	%Precision	98.9	89.2	94.5	96.2
	%Recall	95.8	93.2	96.7	93.9
	%F-Measure	97.3	91.16	95.59	95.04
PC-HP[۲۱]	%Precision	-	65.77	85.35	79.96
	%Recall	-	76.45	76.45	91.50
	%F-Measure	-	80.66	80.66	85.34
PC-P [۲۱]	%Precision	-	59.97	74.11	48.08
	%Recall	-	69.84	47.31	41.88
	%F-Measure	-	64.53	57.75	44.77
DISTILLER [۱۹]	%Precision	-	78.28	61.12	78.96
	%Recall	-	77.17	72.13	75.78
	%F-Measure	-	77.72	66.17	77.34
		Transfer	**VoIP**	**P2P**	**Average**
روش پیشنهادی	%Precision	79.9	92.3	90.9	91.70
	%Recall	92.7	95.7	94.1	94.5
	%F-Measure	85.8	93.9	92.4	93.0
PC-HP[۲۱]	%Precision	98.4	87.24	99.3	86.01
	%Recall	90.2	84.47	99.97	86.51
	%F-Measure	94.1	85.83	99.63	87.71
PC-P[۲۱]	%Precision	67.5	76.33	69.66	65.94
	%Recall	77.1	85.15	77.33	66.45
	%F-Measure	72.0	80.50	73.29	65.47
DISTILLER [۱۹]	%Precision	82.3	80.53	93.33	79.09
	%Recall	75.2	73.74	97.77	78.63
	%F-Measure	78.6	76.99	95.50	78.72

شکل (۶) ماتریس درهم‌ریختگی به‌دست‌آمده از طبقه‌بندی جریان‌های ترافیکی شبکه توسط مدل پیشنهادی را نشان می‌دهد. این ماتریس نتایج طبقه‌بندی تمامی نمونه‌های مجموعه داده در تمامی تکرارهای ارزیابی متقاطع را نشان می‌دهد. در این ماتریس، ستون‌ها با خروجی طبقه‌بندی روش پیشنهادی و سطرها باکلاس واقعی نمونه‌های آزمایشی مطابقت دارند. به‌عنوان‌مثال، مجموع تمام مقادیر در سطر اول ماتریس نشان می‌دهد که تعداد نمونه‌های دسته‌بندی «Browsing» در طول این آزمایش 1980 مورد بوده که روش پیشنهادی 1897 نمونه از آن‌ها را به‌درستی طبقه‌بندی می‌کند.



شکل (۶). ماتریس درهم‌ریختگی برای طبقه‌بندی ترافیک توسط الگوریتم پیشنهادی

در شکل (۷) ماتریس درهم‌ریختگی روش‌های مقایسه شده مشخص‌شده است. مقایسه ماتریس ارائه‌شده در شکل (۶) با ماتریس‌های به‌دست‌آمده از روش‌های مقایسه شده نشان می‌دهد که تعداد خطای طبقه‌بندی روش پیشنهادی برای هر دسته هدف به‌صورت کلی پایین‌تر از سایر روش‌ها است. بررسی مقادیر این ماتریس‌ها بر اساس سطرهای تشکیل‌دهنده آن نشان می‌دهد که نسبت بالاتری از نمونه‌های هر دسته‌ی هدف توسط روش پیشنهادی با برچسب صحیح مقداردهی شده‌اند (پوشش بالاتر) و درعین‌حال؛ با مقایسه این ماتریس‌ها بر اساس ستون‌های تشکیل‌دهنده می‌توان دریافت که درروش پیشنهادی نسبت بالاتری از برچسب‌های تولیدشده برای هر دسته صحیح است (صحت بالاتر).



(الف)

(ب)

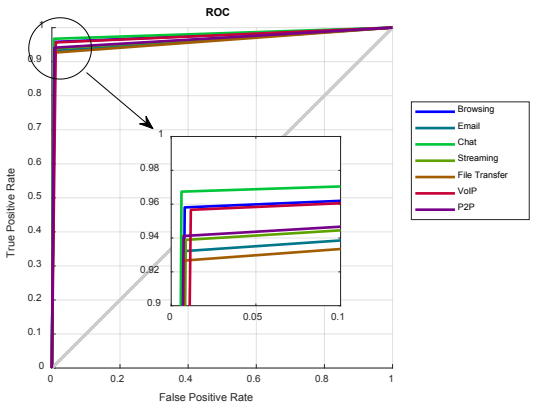
شکل(۷). ماتریس درهم‌ریختگی برای طبقه‌بندی ترافیک توسط (الف) روش PC-HP، و (ب) روش DISTILLER

منحنی ویژگی‌های عملیاتی دریافتی[1] یک نمودار گرافیکی است که برای نشان دادن حساسیت و ویژگی یک الگوریتم طبقه‌بندی برای هر کلاس به‌طور جداگانه استفاده می‌شود. در این منحنی، نرخ‌های مثبت صحیح به‌دست‌آمده از هر کلاس در محور عمودی در برابر نرخ‌های منفی کاذب در محور افقی رسم می‌شود. بدیهی است که طبقه‌بندی کارآمد منجر به نرخ مثبت صحیح بیشتر و نرخ منفی کاذب کمتر می‌شود. بنابراین، مساحت بیشتر زیر منحنی[2] نشانه کار آیی بیشتر در طبقه‌بندی است. در شکل (۸).الف، منحنی‌های ویژگی‌های عملیاتی دریافتی حاصل‌شده از طریق طبقه‌بندی ترافیک توسط رویکرد پیشنهادی نشان داده‌شده است. در این شکل منحنی هر کلاس هدف به‌طور جداگانه نشان داده‌شده است.

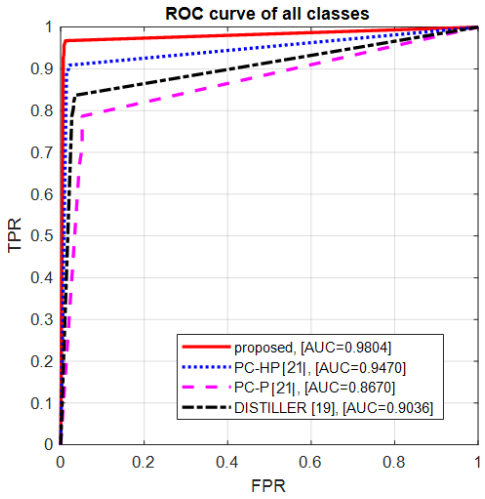
طبق این شکل، کلاس "Chat" بالاترین مقدار مساحت زیر منحنی را دارد و کلاس‌های "Browsing" و "Voip" در رتبه

1 Receiver operating characteristic (ROC)
2 Area Under Curve (AUC)

بعدی هستند. از طرف دیگر، کمترین مقدار این منحنی متعلق به دسته‌های «Transfer» و «Email» است. به‌طورکلی نتایج این شکل نشان می‌دهد که روش پیشنهادی می‌تواند نمونه‌های هر کلاس را با دقت بالایی طبقه‌بندی کند.



(الف)



(ب)

شکل (۸). منحنی‌های ROC به‌دست‌آمده از طریق طبقه‌بندی ترافیک (الف) منحنی به‌دست‌آمده برای رویکرد پیشنهادی به تفکیک هر دسته و (ب) منحنی‌های ROC روش‌های مختلف حاصل از تجمیع خروجی دسته‌های مختلف

شکل (۸).ب به مقایسه منحنی‌های ویژگی‌های عملیاتی دریافتی برای روش پیشنهادی و سایر روش‌ها می‌پردازد. در این شکل، نقاط آستانه به‌دست‌آمده برای کلاس‌های مختلف باهم تجمیع شده تا عملکرد کلی هر مدل در قالب یک منحنی توصیف گردد. مقایسه نمودارهای ترسیم‌شده در این شکل نشان می‌دهد که روش پیشنهادی قادر به افزایش مساحت زیر نمودار در مقایسه با سایر روش‌هاست. این افزایش از کاهش نرخ مثبت کاذب و افزایش نرخ مثبت صحیح به‌صورت هم‌زمان ناشی می‌شود. لذا می‌توان بر اساس این نمودار، برتری عملکردی مدل پیشنهادی در شناسایی انواع جریان‌های ترافیکی شبکه را نتیجه گرفت.

۵- نتیجه‌گیری

در این تحقیق، یک روش جدید مبتنی بر یادگیری ماشین برای طبقه‌بندی جریان‌های ترافیکی شبکه پیشنهاد گردید. روش پیشنهادی شامل مجموعه‌ای از الگوریتم‌های استخراج ویژگی، خوشه‌بندی و طبقه‌بندی برای شناسایی کلاس صحیح جریان‌های شبکه است. این روش از یک مجموعه کوچک از جریان‌های ترافیکی برچسب‌گذاری شده، در کنار یک مجموعه داده بزرگ که نمونه‌های ترافیک آن بدون برچسب هستند؛ استفاده می‌کند که باعث می‌شود رویکرد پیشنهادی برای کاربردهای واقعی قابل‌استفاده باشد، زیرا در موقعیت‌های واقعی اغلب امکان برچسب‌گذاری تمام نمونه‌های موجود وجود ندارد. رویکرد پیشنهادی، از مجموعه‌ای از ویژگی‌های آماری برای توصیف هر جریان استفاده می‌کند. ازآنجایی‌که ممکن است تعداد ویژگی‌ها زیاد باشد و برخی از آن‌ها به کلاس‌های ترافیکی ربطی نداشته باشند، از الگوریتم تجزیه‌وتحلیل مؤلفه‌های اصلی برای کاهش ابعاد ویژگی و بهبود سرعت پردازش نمونه‌ها در مراحل بعدی و همچنین کاهش اثر مخرب ویژگی‌های نامرتبط استفاده گردید. در ادامه از DBScan برای گروه‌بندی جریان‌های ترافیکی با ویژگی‌های آماری مشابه در خوشه‌های یکسان استفاده‌شده است. نتایج شبیه‌سازی نشان می‌دهد که روش پیشنهادی قادر به طبقه‌بندی دقیق جریان‌های ترافیکی است و می‌تواند به‌عنوان یک ابزار کارآمد برای سناریوهای دنیای واقعی استفاده شود.

تحقیق حاضر باوجود دستیابی به نتایج مطلوب در زمینه طبقه‌بندی جریان‌های ترافیکی شبکه، با محدودیت‌های زیر مواجه است:

محدودیت‌های داده: کار آیی روش پیشنهادی به کیفیت و کمیت داده‌های موجود بستگی دارد. در مواردی که مجموعه داده‌های برچسب‌گذاری شده محدود یا مغرضانه است، عملکرد مدل ممکن است به خطر بیفتد. اگرچه مدل پیشنهادی با استفاده از راهبرد خوشه‌بندی در جهت رفع این محدودیت تلاش کرده است، اما همچنان توانایی مدل در تشخیص دقیق ویژگی‌های متنوع ترافیک شبکه به نمایندگی مجموعه داده مورداستفاده برای آموزش بستگی دارد.

پیچیدگی محاسباتی: پیچیدگی محاسباتی روش پیشنهادی، به‌ویژه مراحل استخراج ویژگی و خوشه‌بندی، برای مجموعه داده‌های بزرگ می‌تواند قابل‌توجه باشد. اگرچه این پیچیدگی در مقایسه با راهبردهای یادگیری عمیق ناچیز است، اما ممکن است

همچنان مقیاس‌پذیری آن را برای کاربردهای بلادرنگ یا استقرار بر روی دستگاه‌های دارای منابع پردازشی ناچیز محدود کند.

با توجه به محدودیت‌های ذکرشده و پتانسیل‌های موجود در این حوزه تحقیقاتی، می‌توان پیشنهادهایی را برای ادامه این مسیر تحقیقاتی ارائه داد:

توسعه گام مهندسی ویژگی: توسعه روش‌های مهندسی ویژگی مؤثرتر برای ترافیک شبکه‌های کامپیوتری می‌تواند توانایی مدل را در تشخیص الگوهای مرتبط و کاهش ابعاد فضای ویژگی بهبود بخشد.

بهره‌گیری از هوش مصنوعی قابل توضیح: بررسی روش‌هایی برای قابل توضیح کردن مدل می‌تواند به ایجاد اعتماد به پیش‌بینی‌های آن و تسهیل درک فرآیند تصمیم‌گیری آن کمک کند.

کاربردهای بلادرنگ: در تحقیقات آینده می‌توان به چالش‌های محاسباتی موجود به‌منظور استقرار بلادرنگ مدل برای طبقه‌بندی ترافیک آنلاین رسیدگی نمود. همچنین بررسی چارچوب‌های محاسباتی توزیع‌شده یا شتاب‌دهنده‌های سخت‌افزاری برای بهبود مقیاس‌پذیری مدل برای رسیدگی به استقرارهای بزرگ‌مقیاس شبکه‌های کامپیوتری می‌تواند موضوعی برای ادامه تحقیقات باشد.

انطباق با الگوهای در حال تکامل ترافیک: توسعه سازوکارهایی برای سازگاری مدل با تغییرات در الگوهای ترافیک در طول زمان، مانند دستگاه‌های نوظهور اینترنت اشیاء یا فن‌های جدید حمله می‌تواند تأثیری قابل‌توجه بر عمومیت کاربرد مدل پیشنهادی داشته باشد.

با رسیدگی به این محدودیت‌ها و دنبال کردن جهت‌گیری‌های پیشنهادشده برای تحقیقات آینده، می‌توان رویکرد پیشنهادی را بیشتر اصلاح و گسترش داد تا چالش‌ها و فرصت‌های در حال تکامل در طبقه‌بندی ترافیک شبکه را برطرف کند.

۶. مراجع

[1] M. Bazooband and H. Bahramghiri, "A New Hybrid Approach for Traffic Identification and Classification in Wireless Networks," Sci. J. Electron. Cyber Def., vol. 10, no. 2, 2022.(In Persian). dor: 20.1001.1.23224347.1401.10.2.4.0

[2] S. Z. Majidian, "Reducing the Destructive Effect of Misbehaving Users in Cooperative Spectrum Sensing using Reinforcement Learning," Sci. J. Electron. Cyber Def., vol. 10, no. 4, 2023. (In Persian).dor: 20.1001.1.23224347.1401.10.4.1.1

[3] M. Cotton, L. Eggert, J. Touch, M. Westerlund, and S. Cheshire, "RFC 6335: Internet Assigned Numbers Authority (IANA) Procedures for the Management of the Service Name and Transport Protocol Port Number Registry." RFC Editor, USA, 2011. doi: 10.17487/RFC6335

[4] C. Defense, "Fault Proness Estimation of Software Modules Using Machine learning," Sci. J. Electron. Cyber Def., vol. 11, no. 4, 2024.(In persion). dor: 20.1001.1.23224347.1402.11.4.4.1

[5] E. Kazemi and A. Taei, "Presenting a New Sentiment Analysis Method Based on Multi-objective Archimedes Optimization Algorithm and Machine Learning," Sci. J. Electron. Cyber Def., vol. 11, no. 4, 2024. (In persion). dor: 20.1001.1.23224347.1402.11.4.12.9

[6] A. J. R. K. Dadashtabar Ahmadi, M. Kheirkhah, "Detection of advanced Cyber Attacks, Using Behavior Modeling Based on Natural Language Processing," J. Electron. Cyber Def., vol. 3, pp. 141–151, 2018. (In persion). dor: 20.1001.1.23224347.1397.6.3.12.2

[7] G. Aceto, D. Ciuonzo, A. Montieri, and A. Pescapé, "Mobile Encrypted Traffic Classification Using Deep Learning," in 2018 Network Traffic Measurement and Analysis Conference (TMA), 2018, pp. 1–8. doi: 10.23919/TMA.2018.8506558.

[8] J. Yan, "A Survey of Traffic Classification Validation and Ground Truth Collection," in 2018 8th International Conference on Electronics Information and Emergency Communication (ICEIEC), 2018, pp. 255–259. doi: 10.1109/ICEIEC.2018.8473477.

[9] F. Pacheco, E. Exposito, M. Gineste, C. Baudoin, and J. Aguilar, "Towards the Deployment of Machine Learning Solutions in Network Traffic Classification: A Systematic Survey," IEEE Commun. Surv. Tutorials, vol. 21, no. 2, pp. 1988–2014, 2019, doi: 10.1109/COMST.2018.2883147.

[10] M. Shen, Y. Liu, L. Zhu, K. Xu, X. Du, and N. Guizani, "Optimizing Feature Selection for Efficient Encrypted Traffic Classification: A Systematic Approach," IEEE Netw., vol. 34, no. 4, pp. 20–27, 2020, doi: 10.1109/MNET.011.1900366.

[11] S. Dong, "Multi class SVM algorithm with active learning for network traffic classification," Expert Syst. Appl., vol. 176, p. 114885, 2021, doi: 10.1016/j.eswa.2021.114885.

[12] Z. Bu, B. Zhou, P. Cheng, K. Zhang, and Z.-H. Ling, "Encrypted Network Traffic Classification Using Deep and Parallel Network-in-Network Models," IEEE Access, vol. 8, pp. 132950–132959, 2020, doi: 10.1109/ACCESS.2020.3010637.

[13] A. A. Afuwape, Y. Xu, J. H. Anajemba, and G. Srivastava, "Performance evaluation of secured network traffic classification using a machine learning approach," Comput. Stand. Interfaces, vol. 78, p. 103545, 2021, doi: 10.1016/j.csi.2021.103545.

[14] G. D'Angelo and F. Palmieri, "Network traffic classification using deep convolutional recurrent autoencoder neural networks for spatial–temporal features extraction," J. Netw. Comput. Appl., vol. 173, p. 102890, 2021, doi: 10.1016/j.jnca.2020.102890.

[15] R. Kumar, M. Swarnkar, G. Singal, and N. Kumar, "IoT Network Traffic Classification Using Machine Learning Algorithms: An Experimental Analysis," IEEE Internet Things J., vol. 9, no. 2, pp. 989–1008, 2022, doi: 10.1109/JIOT.2021.3121517.

[16] J. Guan, J. Cai, H. Bai, and I. You, "Deep transfer learning-based network traffic classification for scarce dataset in 5G IoT systems," Int. J. Mach. Learn. Cybern., vol. 12, no. 11, pp. 3351–3365, 2021, doi: 10.1007/s13042-021-01415-4.

[17] M. M. Raikar, M. S M, M. M. Mulla, N. S. Shetti, and M. Karanandi, "Data Traffic Classification in Software Defined Networks (SDN) using supervised-learning," Procedia Comput. Sci., vol. 171, pp. 2750–2759, 2020, doi: 10.1016/j.procs.2020.04.299.

[18] G. Aceto, D. Ciuonzo, A. Montieri, and A. Pescapé, "Toward effective mobile encrypted traffic classification through deep learning," Neurocomputing, vol. 409, pp. 306–315, 2020, doi: 10.1016/j.neucom.2020.05.036.

[19] G. Aceto, D. Ciuonzo, A. Montieri, and A. Pescapé, "DISTILLER: Encrypted traffic classification via multimodal multitask deep learning," J. Netw. Comput. Appl., vol. 183–184, p. 102985, 2021, doi: 10.1016/j.jnca.2021.102985.

[20] S. Rezaei and X. Liu, "Deep Learning for Encrypted Traffic Classification: An Overview," IEEE Commun. Mag., vol. 57, no. 5, pp. 76–81, 2019, doi: 10.1109/MCOM.2019.1800819.

[21] A. M. Sadeghzadeh, S. Shiravi, and R. Jalili, "Adversarial Network Traffic: Towards Evaluating the Robustness of Deep-Learning-Based Network Traffic Classification," IEEE Trans. Netw. Serv. Manag., vol. 18, no. 2, pp. 1962–1976, 2021, doi: 10.1109/TNSM.2021.3052888.

[22] J. Höchst, L. Baumgärtner, M. Hollick, and B. Freisleben, "Unsupervised Traffic Flow Classification Using a Neural Autoencoder," in 2017 IEEE 42nd Conference on Local Computer Networks (LCN), 2017, pp. 523–526. doi: 10.1109/LCN.2017.57.

[23] F. L. Gewers et al., "Principal Component Analysis: A Natural Approach to Data Exploration," ACM Comput. Surv., vol. 54, no. 4, May 2021, doi: 10.1145/3447755.

[24] D. Deng, "DBSCAN Clustering Algorithm Based on Density," in 2020 7th International Forum on Electrical Engineering and Automation (IFEEA), 2020, pp. 949–953. doi: 10.1109/IFEEA51475.2020.00199.

[25] M. Raja, P. Hasan, M. Mahmudunnobe, M. Saifuddin, and S. N. Hasan. Membership determination in open clusters using the DBSCAN Clustering Algorithm. Astronomy and Computing, 47, 100826, 2024. doi: 10.1016/j.ascom.2024.100826.

[26] J. L. Velazquez-Rodriguez, Y. Villuendas-Rey, O. Camacho-Nieto and C. Yanez-Marquez. A novel and simple mathematical transform improves the perfomance of lernmatrix in pattern classification. Mathematics, 8(5), 732, 2020. doi: /10.3390/math8050732.

[27] Y. Zhao and X. Shu. Speech emotion analysis using convolutional neural network (CNN) and gamma classifier-based error correcting output codes (ECOC). Scientific Reports, 13(1), 20398, 2023. doi: 10.1038/s41598-023-47118-4.