DOR: https://csj.isi.org.ir/article_233351.html

DOI: https://doi.org/10.22034/csj.2025.538051.1114

بهار ۱۴۰۵، دوره ۱۱، شماره ۱، پیاپی ۴۰

دریافت مقاله: ۱۴۰۴/۰۵/۰۸

پذیرش مقاله: ۱۴۰۴/۰۸/۱۱

نوع مقاله: پژوهشی

ارائه یک روش ترکیبی برای داده‌افزایی و متوازن‌سازی مجموعه داده‌های نامتوازن

پرستو محقق

کارشناس ارشد مهندسی فناوری اطلاعات، دانشکده مهندسی برق و کامپیوتر، دانشگاه سیستان و بلوچستان، زاهدان، ایران

پست الکترونیکی: P.mohaghegh@pgs.usb.ac.ir

مهری رجائی*

استادیار، دانشکده مهندسی برق و کامپیوتر، دانشگاه سیستان و بلوچستان، زاهدان، ایران

پست الکترونیکی: rajayi@ece.usb.ac.ir

سمیرا نوفرستی

دانشیار، دانشکده مهندسی برق و کامپیوتر، دانشگاه سیستان و بلوچستان، زاهدان، ایران

پست الکترونیکی: snoferesti@ece.usb.ac.ir

چکیده

یکی از چالش‌های مهم در یادگیری ماشین، داده‌های نامتوازن است. در مجموعه داده‌های نامتوازن، توزیع ناهمگون نمونه‌ها بین کلاس‌های مختلف، باعث می‌شود مدل‌های طبقه‌بندی عملکرد ضعیفی در شناسایی کلاس اقلیت (کلاس با تعداد نمونه کمتر) داشته باشند. این مشکل زمانی برجسته‌تر می‌شود که در یک کاربرد مشخص کلاس اقلیت دارای اهمیت بیشتری است. در این مقاله، یک روش ترکیبی برای حل مسئله داده‌های نامتوازن پیشنهاد می‌شود. این روش که HDAM نامیده می‌شود با هدف تولید نمونه‌های موثرتر برای کلاس اقلیت، دو تکنیک شناخته شده داده‌افزایی به نام‌های H-SMOTE و LoRAS را ترکیب می‌کند. ارزیابی‌های انجام شده نشان می‌دهند به‌صورت میانگین روش پیشنهادی برای هر سه طبقه‌بند جنگل تصادفی، K نزدیک‌ترین همسایه و شبکه عصبی پرسپترون چندلایه و براساس معیارهای مختلف شامل فراخوانی، دقت، AUC و F1 عملکرد بهتری در مقایسه با دو روش پایه دارد. برای نمونه، روش HDAM بر اساس معیار F1 در مقایسه با H-SMOTE که عملکرد بهتری نسبت به LoRAS داشته است، کارایی طبقه‌بندهای جنگل تصادفی، K نزدیک‌ترین همسایه و شبکه عصبی را به‌ترتیب ۱/٦٤، ۰/۰۵ و ٦/۱ درصد بهبود داده است.

واژه‌های کلیدی: یادگیری ماشین، طبقه‌بندی، مجموعه‌داده‌های نامتوازن، داده‌افزایی، روش ترکیبی

۱- مقدمه

طبقه‌بندی یکی از تکنیک‌های یادگیری ماشین است که با آموزش بر روی یک مجموعه داده آموزشی از نمونه‌های برچسب‌خورده به ساخت یک مدل پیش‌بینی‌کننده می‌پردازد

* نویسندهٔ مسئول

که قادر است برچسب کلاس نمونه‌های جدید را تعیین کند. یکی از مشکلاتی که الگوریتم‌های طبقه‌بندی با آن مواجه هستند، توزیع نامتوازن کلاس‌های مختلف در مجموعه داده آموزشی است. مشکل عدم توازن زمانی رخ می‌دهد که تعداد نمونه‌ها در یک کلاس (که کلاس اقلیت نامیده می‌شود) در مقایسه با سایر کلاس‌ها بسیار کمتر باشد، در حالی که در بسیاری از مسائل واقعی، پیش‌بینی صحیح برچسب نمونه‌های کلاس اقلیت دارای اهمیت بیشتری است. برای نمونه، در مسائلی نظیر شناسایی و طبقه‌بندی آفات کشاورزی، برخی از آفات بسیار نادرند یا در تشخیص تقلب در کارت‌های اعتباری، تراکنش‌های مشکوک اندک هستند. در حالی که هدف اصلی این کاربردها شناسایی آفات نادر یا تراکنش‌های مشکوک است. به‌طورکلی مسئله داده‌های نامتوازن یکی از چالش‌های اصلی در یادگیری ماشین محسوب می‌شود، به‌ویژه در حوزه‌هایی مانند تشخیص نفوذ و طبقه‌بندی ناهنجاری‌ها که در آن رخدادهای ناهنجار بسیار کمتر از رویدادهای عادی رخ می‌دهند.

الگوریتم‌های طبقه‌بندی سنتی دقت لازم را در پیش‌بینی نمونه‌های کلاس اقلیت ندارند، زیرا نمونه‌های آموزشی کافی برای کلاس اقلیت وجود ندارد و این کلاس به درستی آموزش داده نمی‌شود. برای حل مشکل داده‌های نامتوازن، رویکردهای متعددی در سه سطح داده، مدل و ترکیبی ارائه شده‌اند [۱]. در این مقاله بر رویکرد سطح داده متمرکز می‌شویم که در مقایسه با رویکرد سطح مدل رایج‌تر است. هدف رویکرد سطح داده، کاهش نسبت عدم تعادل بین کلاس‌های اکثریت و اقلیت و به تبع آن بهبود دقت طبقه‌بندی است [۱].

به‌طورکلی، تکنیک‌های سطح داده برای حل مسئله مجموعه داده‌های نامتوازن به دو دسته انتخاب نمونه و داده‌افزایی تقسیم می‌شوند. روش‌های انتخاب نمونه سعی در کاهش تعداد نمونه‌های کلاس اکثریت به‌منظور رسیدن به تعادل نسبی در اندازه کلاس‌ها دارد [۲-٤]. در مقابل، روش‌های داده‌افزایی، به‌صورت تصادفی یا در مناطقی که اهمیت بیشتری برای کاربرد دارد نمونه‌های کلاس اقلیت را کپی‌برداری یا تولید می‌کنند تا زمانی که تعادل بین نمونه‌های کلاس اقلیت و اکثریت حاصل شود [۵-۱۰]. روش‌های انتخاب نمونه موجب حذف نمونه‌هایی می‌شوند که برای جمع‌آوری و برچسب‌گذاری آن‌ها وقت و هزینه بسیاری صرف شده است و گاها نمونه‌های مهمی هستند که در آموزش طبقه‌بند نقش به‌سزایی دارند. از طرف دیگر، ممکن است مشکل انتخاب نمونه بیش از حد رخ می‌دهد که منجر به کاهش عملکرد طبقه‌بندی می‌شود. به همین دلیل، در پژوهش‌های اخیر، روش‌های داده‌افزایی بیشتر مورد توجه قرار گرفته‌اند.

در این مقاله، یک روش داده‌افزایی ترکیبی به نام HDAM[1] پیشنهاد می‌شود که با ترکیب دو روش برجسته H-SMOTE [۹] و LoRAS [۱۰] به گونه‌ای که هر داده جدید به‌صورت تصادفی از یکی از این دو روش تولید می‌شود، بر گسترش نمونه‌های کلاس اقلیت می‌پردازد. هدف از ترکیب دو روش نامبرده، استفاده از مزایا و کاهش مشکلات هر کدام از این روش‌ها است. به‌طور مشخص، H-SMOTE [۹] برای تولید نمونه‌های جدید با توزیع پایدار به‌کار می‌رود، در حالی که LoRAS تنوع بیشتری را با ترکیب خطی نمونه‌های محلی ایجاد می‌کند. ادغام این دو رویکرد منجر به ایجاد تعادلی میان کیفیت و تنوع نمونه‌های تولیدی می‌شود.

برای ارزیابی روش پیشنهادی، آزمایش‌ها روی چند مجموعه‌داده مرجع از جمله مجموعه داده تشخیص تقلب در کارت‌های اعتباری با ۲۸٤،۸۰۷ رکورد و عدم‌توازن شدید، مجموعه‌داده Pima Diabetes با ۷٦۸ نمونه و مجموعه‌داده نفوذ UNSW-NB15 با بیش از ۲/۵ میلیون رکورد، انجام شده است. نتایج نشان می‌دهد روش پیشنهادی نسبت به H-SMOTE که عملکرد بهتری در مقایسه با LoRAS داشته است، در معیار AUC به‌طور میانگین در طبقه‌بندهای جنگل تصادفی، K نزدیک‌ترین همسایه و شبکه عصبی به‌ترتیب ٤/۹۵ ، ۰/۹٦ و ٦/٤ درصد بهبود داشته است. همچنین روش پیشنهادی، معیار F1 طبقه‌بندهای مذکور را در

1-Hybrid Data Augmentation Method

مقایسه با H-SOMTE به میزان ۱/٦٤، ۰/۰۵ و ٦/۱ درصد افزایش داده است. این نتایج بیانگر کارایی روش پیشنهادی در حوزه‌های مختلف است.

به‌طورکلی، نوآوری‌های اصلی مقاله به‌صورت زیر ارائه می‌شود:

- برای نخستین بار ترکیب H-SMOTE [۹] و LoRAS [۱۰] به‌منظور مقابله با عدم توازن داده‌ها با هدف ایجاد تعادلی میان کیفیت و تنوع نمونه‌های تولید شده پیشنهاد شده است.
- ارزیابی جامعی روی چندین مجموعه‌داده از حوزه‌های مختلف (سلامت، مالی و امنیت سایبری) انجام گرفته است.
- نشان داده‌ایم که روش پیشنهادی بر اساس معیارهای AUC، دقت، فراخوانی و F1 در اغلب موارد عملکرد بهتری نسبت به روش‌های متداول دارد، بجز در برخی طبقه‌بندهای مبتنی بر فاصله که محدودیت‌های آن نیز به‌صراحت بحث شده است.

ادامه مقاله به‌صورت زیرسازمان‌دهی شده است: در بخش ۲ تحقیقات پیشین در زمینه طبقه‌بندی نامتوازن مرور می‌شوند. در بخش ۳ جزئیات روش پیشنهادی شرح داده می‌شود. در بخش ٤ ارزیابی کارایی روش پیشنهادی ارائه می‌گردد. در پایان، بخش ۵ به نتیجه‌گیری می‌پردازد.

۲-کارهای مرتبط

همان‌طور که پیشتر گفته شد، روش‌های موجود برای متعادل‌سازی مجموعه داده‌های نامتوازن به‌طورکلی به دو دسته انتخاب نمونه و داده‌افزایی تقسیم می‌شوند. در ادامه به معرفی کارهای مرتبط در هر دسته پرداخته می‌شود.

۲-۱- تکنیک‌های انتخاب نمونه

Tsai و همکاران [۲]، روشی با نام CBIS پیشنهاد داده‌اند که در گام اول نمونه‌های کلاس اکثریت را خوشه‌بندی می‌کند و هر خوشه به‌عنوان زیرکلاسی از کلاس اکثریت برچسب‌گذاری می‌شود. در گام بعد، الگوریتم انتخاب نمونه پایه مانند الگوریتم ژنتیک بر روی کلیه خوشه‌ها عمل پالایش را انجام می‌دهد و یک مجموعه داده کاهش‌یافته ایجاد می‌شود. نمونه‌های نوفه مجموعه کاهش‌یافته کنار گذاشته می‌شود و مجموعه داده کاهش‌یافته نهایی از کلاس اکثریت با نمونه‌های کلاس اقلیت ترکیب شده و به‌عنوان مجموعه آموزشی جدید در نظر گرفته می‌شود.

در [۳] روشی به نام DCIS معرفی شده است که از تکنیک تقسیم و غلبه برای تقسیم مجموعه داده آموزشی اصلی به چندین زیرمجموعه، اعمال الگوریتم انتخاب نمونه بر روی هر یک از زیرمجموعه‌ها و سپس ترکیب زیرمجموعه‌های کاهش‌یافته جهت ساخت مجموعه داده متوازن استفاده می‌کند. برای ترکیب زیرمجموعه‌های کاهش‌یافته دو تکنیک اجتماع و اشتراک پیشنهاد شده است. نتایج آزمایش‌های انجام گرفته نشان می‌دهد اجتماع زیرمجموعه‌ها به نتایج بهتری دست یافته است.

در [٤] روشی مبتنی بر یادگیری تقویتی برای انتخاب نمونه پیشنهاد شده است که ابتدا خوشه‌هایی از نمونه‌ها را بر اساس یک معیار شباهت می‌سازد. سپس عاملی در نظر می‌گیرد که می‌تواند به‌صورت تدریجی داده‌ها را به مجموعه کاهش یافته نهایی اضافه کند. در یک حلقه، این عامل خوشه‌ای از نمونه‌ها (عمل) را برای اضافه شدن به نمونه‌های برگزیده (حالت) انتخاب می‌کند و پاداشی متناسب با کاهش خطای مدل پیش‌بینی‌کننده دریافت می‌کند. خروجی الگوریتم (سیاست عامل) ماتریسی است که توازن بین بهبود مدل و اندازه داده‌ها را نشان می‌دهد. هر درایه از ماتریس ارزش افزودن یک خوشه به نمونه‌های موجود را نشان می‌دهد. این ماتریس توانایی متعادل کردن اهداف کاهش نوفه و حجم داده را دارد.

۲-۲- تکنیک‌های داده‌افزایی

در روش‌های داده‌افزایی ایده اصلی، گسترش تعداد نمونه‌های کلاس اقلیت برای متوازن‌سازی توزیع نمونه‌ها بین کلاس‌ها می‌باشد. روش SMOTE در بین روش‌های

داده‌افزایی از محبوبیت بیشتری برخوردار است [۵] زیرا نسخه‌های توسعه‌یافته زیادی برای آن ارائه شده است. ایده اصلی SMOTE پایه این است که برای هر نمونه کلاس اقلیت با توجه به میزان شباهتی که با دیگر نمونه‌های کلاس اقلیت دارد، K نزدیک‌ترین همسایگانش تعیین می‌شوند. سپس بین هر دو نمونه از همسایه‌ها، نمونه‌های جدید به‌صورت تصادفی تولید می‌شوند. مشکلی اصلی SMOTE این است که به‌دلیل محدود بودن ناحیه تولید نمونه‌ها نمی‌تواند به‌خوبی بر چالش توزیع نامتوازن داده‌ها غلبه کند. برای حل این مشکل، انواع مختلفی از SMOTE پیشنهاد شده است که به سه دسته کلی روش‌های مبتنی بر شناسایی مناطق امن در طبقه اقلیت، روش‌های متکی بر اهمیت ناحیه مرزی بین کلاس اقلیت و اکثریت و روش‌های مبتنی بر رابطه بین کلاس اقلیت و اکثریت تقسیم می‌شوند. در ادامه به معرفی برخی از روش‌های پیشنهاد شده برای هر دسته می‌پردازیم.

روش پایه SMOTE به این اهم رسیده است که همه مناطق در کلاس اقلیت برابر نیستند و برخی مناطق ممکن است مهم‌تر یا ایمن‌تر از سایر مناطق باشد. Douzas و Bacao [۶]، روشی به نام G-SMOTE پیشنهاد داده‌اند که در آن با تعیین یک شعاع امن در اطراف هر نمونه از کلاس اقلیت یک منطقه هندسی به نام ابر کروی برای تولید نمونه‌های مصنوعی جدید ایجاد می‌شود.

روش‌هایی که بر اهمیت ناحیه مرزی بین کلاس اقلیت و اکثریت تمرکز دارند، مبتنی بر این دیدگاه هستند که نمونه‌های کلاس اقلیت واقع در نزدیکی مرز تصمیم‌گیری طبقه‌بند از اهمیت بیشتری برخوردار هستند، بنابراین تولید نمونه‌های مصنوعی در این ناحیه مرزی می‌تواند به بهبود عملکرد طبقه‌بند کمک کند. Saez و همکاران [۷]، روش SMOTE-IPF را ارائه دادند که از یک پالایهٔ بخش‌بندی، برای اصلاح نمونه‌های مرزی و دارای نوفه تولید شده توسط SMOTE استفاده می‌کند.

سومین بینش در راستای بهبود عملکرد SMOTE این است که رابطه بین کلاس اکثریت و اقلیت اهمیت دارد. Bunkhumpornpat و همکاران [۸]، در روشی به نام SL-SMOTE، با محاسبه سطح ایمنی برای هر نمونه اقلیت که براساس فراوانی نمونه‌های اکثریت در K نزدیک‌ترین همسایگی نمونه اقلیت انجام می‌شود، داده‌افزایی را انجام داده‌اند.

Chao و Zhang [۹]، روش H-SMOTE را با هدف تولید نمونه‌های مصنوعی یکنواخت‌تر و نزدیک‌تر به مرکز کلاس اقلیت، معرفی کردند. این روش از دو عمل نمونه‌برداری و پالایش نمونه‌های دارای نوفه تشکیل شده است. ابتدا نمونه مرکزی کلاس اقلیت با میانگین‌گیری از ویژگی‌ها مشخص می‌شود و سپس فاصله منهتن هر نمونه از کلاس اقلیت با مرکز محاسبه می‌شود و نمونه جدید در سطح اول تولید می‌گردد. این عمل برای نمونه‌های جدید و نمونه‌های کلاس اقلیت در سطح دوم نیز تکرار می‌شود.

Bej و همکاران [۱۰]، یک روش داده‌افزایی با نام LoRAS را معرفی کردند. در این روش برای هر نمونه P از کلاس اقلیت، ناحیه‌ای چندضلعی در نظر گرفته شده و سپس K نزدیک‌ترین همسایه‌های نمونه P در آن ناحیه انتخاب می‌شوند. برای هر نمونه که در همسایگی نمونه P قرار دارد، به اندازه معینی، نمونه سایه ایجاد می‌شود. نمونه‌های سایه با ایجاد نوفه در محدوده تابع توزیع نرمال با انحراف معیار استاندارد ایجاد می‌شوند. سپس به‌صورت تصادفی به تعداد ویژگی‌ها از لیست نمونه‌های سایه، نمونه انتخاب می‌شود و در وزن‌های بردار آفین ضرب می‌شوند و خروجی بردار به‌عنوان نمونه جدید در نظر گرفته می‌شود

Temraz و Keane [۱۱]، روشی مبتنی بر جفت‌های متضاد برای داده‌افزایی پیشنهاد دادند. این روش که CFA نام دارد، ابتدا هر دو نمونه از کلاس اکثریت (نمونه x و اقلیت (نمونه p) را که حداکثر در دو ویژگی تفاوت دارند، به‌صورت جفت در مجموعه داده آموزشی در نظر می‌گیرد که این جفت‌ها را جفت‌های متضاد می‌نامند. سپس برای هر نمونه از کلاس اکثریت که با هیچ نمونه‌ای از کلاس اقلیت

جفت نشده است (نمونه x')، با استفاده از فاصله اقلیدسی نزدیک‌ترین همسایه آن از جفت‌های متضاد، شناسایی شده و یک نمونه مصنوعی (نمونه p') با استفاده از ویژگی‌های این جفت متضاد تولید می‌شود. نمونه p' ویژگی‌های مشابه را از p و ویژگی‌های متفاوت خود را از x' می‌گیرد. اگرچه CFA عملکرد خوبی داشته است اما همچنان با محدودیت‌هایی همراه است. از جمله این محدودیت‌ها نیاز به شرکت حداقل ۵ درصد از نمونه‌های کلاس اکثریت در جفت‌های متضاد است. بنابراین این روش بدون جفت‌های متضاد کافی در تولید نمونه‌های اقلیت مصنوعی به شدت با مشکل روبه‌رو خواهد شد.

در سال‌های اخیر، به‌کارگیری مدل‌های زبانی بزرگ برای داده‌افزایی مورد توجه محققان قرار گرفته است. در این رویکرد، با طراحی پرسش‌های مناسب، از مدل زبانی بزرگ خواسته می‌شود نمونه‌های مصنوعی مرتبط با داده‌های اولیه تولید کند. این نمونه‌ها به همراه داده‌های اصلی برای افزایش حجم و تنوع داده‌ها در فرآیند آموزش مدل‌های یادگیری ماشین به‌کار می‌روند [۱۲، ۱۳].

در جدول (۱)، خلاصه روش‌های داده‌افزایی معرفی شده به همراه نقاط ضعف و قوت هر کدام و کارایی آنها گزارش داده شده است. علی‌رغم تحقیقات گسترده، مجموعه‌داده‌های نامتوازن همچنان یک مسئله چالش‌برانگیز در حوزه یادگیری ماشین به‌شمار می‌آید. دلیل این امر آن است که بسیاری از روش‌های موجود یا صرفاً برای یک کاربرد خاص طراحی شده‌اند، یا از مقیاس‌پذیری کافی برخوردار نیستند، یا توانایی تولید نمونه‌های کافی برای رفع عدم توازن شدید را ندارند و یا در شرایط خاص، پیچیدگی محاسباتی بالایی ایجاد می‌کنند. روش H-SMOTE با وجود تولید نمونه‌های جدید با توزیع پایدار و معقول و عملکرد بهتر نسبت به الگوریتم‌های پایه مانند SMOTE، همچنان در سطوح بالای عدم‌توازن با مشکل کمبود نمونه‌های مصنوعی مطابق با الگوی واقعی داده‌ها مواجه است. از سوی دیگر، روش LoRAS می‌تواند مجموعه‌داده‌های با نرخ عدم تعادل بالا و حجم کلان را به‌خوبی مدیریت کند، اما در مواردی که داده‌های اصلی پراکنده باشند، احتمال تولید نمونه‌های غیرقابل اعتماد وجود دارد و دارای پیچیدگی محاسباتی بالاتری نسبت به H-SMOTE است. در این مقاله تلاش کرده‌ایم با ترکیب این دو روش بر محدودیت‌های استفاده تکی غلبه کرده و رویکردی کاراتر برای طبقه‌بندی متوازن ارائه دهیم، به‌گونه‌ای که مزایای هر دو روش هم زمان مورد بهره‌برداری قرار گیرد.

۳- روش پیشنهادی

در روش پیشنهادی با نام HDAM، برای متوازن‌سازی مجموعه داده، از ترکیب دو الگوریتم H-SMOTE و LoRAS برای تولید نمونه‌های مصنوعی در کلاس اقلیت استفاده می‌شود. دلیل انتخاب دو الگوریتم H-SMOTE و LoRAS، در انتهای بخش ۲ عنوان شده است.

در روش پیشنهادی تا زمانی که تعداد نمونه‌های کلاس اقلیت (C_{min}) برابر با تعداد نمونه‌های کلاس اکثریت (C_{maj}) نشده است، به صورت تصادفی یکی از دو الگوریتم H-SMOTE و LoRAS انتخاب می‌شود و با الهام گرفتن از ایده آن الگوریتم یک نمونه اقلیت جدید ایجاد می‌گردد. پس از بررسی تکراری نبودن، این نمونه به کلاس اقلیت گسترش یافته که با C'_{min} نشان داده می‌شود و در ابتدا شامل نمونه‌های C_{min} است، اضافه می‌شود. در پایان، کلاس اقلیت گسترش یافته تعداد نمونه‌هایی برابر با کلاس اکثریت خواهد داشت.

شکل (۱) شبه کد روش پیشنهادی را نشان می‌دهد. مطابق شکل (۱)، نمونه‌های دو کلاس اقلیت و اکثریت به‌عنوان ورودی دریافت و پارامترهای الگوریتم LoRAS مقداردهی می‌شوند. سپس تا زمانی که تعداد نمونه‌های کلاس اقلیت برابر با تعداد نمونه‌های کلاس اکثریت نشده است (خط ۱)، به تصادف یکی از دو الگوریتم H-SMOTE و LoRAS انتخاب می‌شود (خط ۲) و یک نمونهٔ اقلیت بر اساس نمونه‌های موجود در C_{min} تولید شده و پس از

جدول ۱. مقایسه روش‌های پیشین داده‌افزایی

مرجع	ایده اصلی	مزایا	معایب	نتایج		
				روش ارزیابی	طبقه‌بند	کارایی
SMOTE [۵]	درون‌یابی خطی بین نمونه‌های کلاس اقلیت	پیاده‌سازی ساده، تنوع داده‌های تولید شده	تولید داده‌های نامعتبر در صورت وجود نمونه‌های نویز یا مرزی در همسایگی، تولید نمونه‌های تکراری	میانگین نتایج حاصل از روش اعتبارسنجی متقابل با ۵ حلقه بر روی چند مجموعه داده مختلف از مخزن UCI	DT	AUC=۰/۸۶۵ F^1=۰/۶۴۱
					KNN	AUC=۰/۸۴۲ F^1=۰/۵۹۰
G-SMOTE [۶]	درون‌یابی خطی بین نمونه‌های موجود در مناطق امن تعیین شده برای کلاس اقلیت	کاهش نمونه‌های همپوشان با کلاس اکثریت، تنوع بیشتر نمونه‌های تولید شده در مقایسه با SMOTE	محاسبات پیچیده، نیاز به تنظیم دقیق پارامترهای متعدد، تولید نمونه‌های تکراری	میانگین نتایج حاصل از روش اعتبارسنجی متقابل با ۵ حلقه بر روی چند مجموعه داده دودویی از مخازن UCI و KEEL	DT	AUC=۰/۸۹۴ F^1=۰/۶۹۰
					KNN	AUC=۰/۸۶۳ F^1=۰/۶۲۰
SMOTE-IPF [۷]	به‌کارگیری فیلتر پارتیشین‌بندی برای اصلاح نمونه‌های مرزی و دارای نویز	کاهش نمونه‌های نوفه و ایجاد مرز تصمیم یکنواخت‌تر نسبت به SMOTE، عملکرد بهتر روی مجموعه داده‌های نویزدار	هزینه محاسباتی و زمان اجرای بالا برای مجموعه‌داده‌های بزرگ، پیاده‌سازی پیچیده، حذف نمونه‌های مرزی مفید توسط پالایه‌ها در برخی مجموعه‌ها	میانگین نتایج حاصل از روش اعتبارسنجی متقابل با ۵ حلقه بر روی چند مجموعه داده واقعی	C۴,۵	AUC=۰/۸۴۰
SL-SMOTE [۸]	تولید نمونه‌های اقلیت بر اساس فراوانی نمونه‌های اکثریت در همسایگی آن	کاهش تولید نمونه‌های نامعتبر در مرز تصمیم	نیاز به برآورد دقیق سطح امن، پیچیدگی بالاتر نسبت به SMOTE	میانگین نتایج حاصل از روش اعتبارسنجی متقابل با ۱۰ حلقه بر روی مجموعه داده Satimage	C۴,۵	F^1=۰/۹۲۵
H-SMOTE [۹]	نمونه‌برداری و فیلتر نمونه‌های دارای نوفه	توزیع یکنواخت‌تر و پایدارتر نمونه‌ها نسبت به SMOTE	پیچیدگی محاسباتی بیشتر در مقایسه با SMOTE پایه، عدم تطابق نمونه‌های تولید شده با الگوی واقعی داده‌ها در مجموعه داده‌های بسیار نامتوازن	میانگین نتایج بر روی ۱۲ مجموعه داده مخزن UCI که با روش تقسیم تصادفی به نسبت ۳۰ به ۷۰ به دو مجموعه آزمون و آموزش تقسیم شده‌اند	SVM	AUC=۰/۹۵۳ F^1=۰/۹۵۲
LoRAS [۱۰]	تولید نمونه‌های اقلیت با میانگین‌گیری از زیرمجموعه‌های همسایه به‌جای خط مستقیم	عملکرد بالا در مجموعه‌داده‌های بسیار نامتوازن	احتمال تولید نمونه‌های غیرقابل اعتماد، پیچیدگی محاسباتی بالاتر نسبت به H-SMOTE، نیاز به تنظیم پارامترهای متعدد	میانگین نتایج حاصل از روش اعتبارسنجی متقابل با ۳ و ۱۰ حلقه بر روی ۱۴ مجموعه داده از مخازن UCI و KEEL	KNN, LR, RF	F^1=۰/۴۳۳ Balanced-accuracy=۰/۷۸۳
CFA [۱۱]	تولید نمونه‌های اقلیت با نمونه‌های کلاس مقابل به‌جای درون‌یابی	قابل‌توضیح، سازگار با قیود دامنه	نیاز به تعریف سازوکار معتبر برای جفت‌های متضاد در هر کاربرد	میانگین نتایج حاصل از روش اعتبارسنجی متقابل با ۵ حلقه بر روی ۲۵ مجموعه داده از مخازن UCI و KEEL	KNN	AUC=۰/۹۸۸
					MLP	AUC=۰/۹۴۱
IIoT-IDS [۱۲]	ترکیب داده‌سازی مبتنی بر مدل زبانی (توصیف/تولید ردیف‌های عددی) با روش‌های درون‌یابی برای طبقه‌بندی حملات	بهبود پوشش الگوهای نادر، مناسب برای طبقه‌بندی چندکلاسه	خاص کاربرد، ریسک عدم تطابق با توزیع واقعی داده‌ها	میانگین نتایج بر روی ۸ مجموعه داده نفوذ در IIoT که با روش تقسیم تصادفی به نسبت ۲۰ به ۸۰ به دو مجموعه آزمون و آموزش تقسیم شده‌اند	RF	F^1=۰/۸۰۷
					MLP	F^1=۰/۷۱۶

HDAM Algorithm

Inputs:

C_{min}: Minority class data points

C_{maj}: Majority class data points

k: Number of nearest neighbors to be considered per parent data point (default value : 30 if $|C_{min}| > 100$, 5 otherwise)

$|S_p|$: Number of generated shadow samples per parent data point (default value : $\max\left(\left\lceil \frac{2|F|}{k} \right\rceil, 40\right)$)

N_{aff}: Number of shadow points to be chosen for a random affine combination (default value : $|F|$)

Outputs:

C'_{min}: Minority data points and non_duplicate generated samples

1 **WHILE** $|C'_{min}| < |C_{maj}|$ **do**

2 Method ← Randomly select an algorithm (H-SMOTE or LoRAS)

3 **IF** Method == LoRAS **then**

4 Execute LoRAS Algorithm to generate a new minority sample (p)

5 **ELSE**

6 Execute H-SMOTE Algorithm to generate a new minority sample (p)

7 **IF** p is not a duplicate sample **then**

8 Add p to C'_{min}

9 **End WHILE**

10 **Return** C'_{min}

شکل ۱. شبه کد روش پیشنهادی HDAM

بررسی تکراری نبودن، این نمونه به مجموعه C'_{min} اضافه می‌گردد (خطوط ۳ تا ۸).

به‌صورت دقیق‌تر، اگر الگوریتم LoRAS انتخاب شود (خط ۳)، یک نمونه p_i از کلاس اقلیت C_{min} انتخاب و ناحیه‌ای مکعب مستطیل شکل اطراف آن در نظر گرفته می‌شود. سپس k نزدیک‌ترین همسایه‌های نمونه p_i تعیین می‌شوند. برای هر نمونه q والد که در همسایگی نمونه p_i قرار دارد، مطابق رابطه (۱) به اندازه $|S_p|$ نمونه سایه ایجاد می‌شود [۱۰]:

(۱)

$$|S_p| = \left(\max\left(\left\lceil \frac{۲|F|}{k} \right\rceil, ۴۰ \right) \right)$$

که F تعداد ویژگی‌ها است. نمونه‌های سایه به وسیله تولید نوفه با کمک تابع توزیع نرمال با انحراف معیار استاندارد L_σ ایجاد می‌شوند و مقدار پیش‌فرض آن [۰/۰۰۵, ... ,۰/۰۰۵] می‌باشد. سپس به‌صورت تصادفی به اندازه N_{aff} که به تعداد ویژگی‌ها است از لیست نمونه‌های سایه، نمونه انتخاب می‌شود و مطابق رابطه (۲) در وزن‌های بردار آفین ضرب می‌شوند. خروجی بردار به‌عنوان نمونه جدید به مجموعه نمونه‌های مصنوعی افزوده می‌شود. یک بردار آفین با وزن‌های تصادفی به‌صورت $\alpha_۱ + \alpha_۲ + \alpha_۳ = ۱$ تعریف می‌شود و نمونه جدید از طریق رابطه (۲) به‌دست می‌آید [۱۰]:

(۲)

$$L = \alpha_۱ s_۱ + \alpha_۲ s_۲ + \alpha_۳ s_۳$$

اگر الگوریتم انتخابی H-SMOTE باشد (خط ۵ شبه کد)، به تصادف، تولید نمونه سطح اول یا دوم انتخاب می‌شود. اگر تولید نمونه سطح اول انتخاب شود، برای نمونه p_i از کلاس اقلیت که به تصادف انتخاب شده است، مطابق رابطه (۳)، نمونه جدید x_i میان نمونه مرکزی Z (که با میانگین گرفتن از نمونه‌های کلاس اقلیت محاسبه می‌شود) و نمونه p_i تولید می‌شود و در کلاس اقلیت قرار می‌گیرد [۹].

$$x_i = p_i + ۰/۵ \times (Z - p_i) \qquad (۳)$$

اگر تولید نمونه سطح دوم انتخاب شود، دو نمونه سطح اول x_i و x_j مطابق رابطه (۳) تولید می‌شوند. سپس مطابق رابطه (٤) نمونه جدید x_{ii} در سطح دوم ایجاد می‌شود [۹].

$$x_{ii} = x_i + ۰/۵ \times (x_j - x_i) \qquad (٤)$$

این روال تا زمان برآورده شـــدن شرط توقف (برابری اندازه کلاس اقلیت و اکثریت) ادامه خواهد داشـــت. کلاس اقلیت نهایی تعداد نمونه‌های برابر با تعداد نمونه‌های کلاس اکثریت دارد و بنابراین مشکل عدم توازن مجموعه داده‌ای حل می‌شود. در پایان، مجموعه آموزش متوازن جدید که متشـــکل از نمونه‌های دو کلاس C'_{min} و C_{maj} است برای ساخت مدل پیش‌بینی کننده استفاده می‌شود.

۴ - نتایج

در ادامـــه ابتدا مشـــخصات مجموعـــه داده‌های مورد استفاده و نیز ابزار و تنظیمات شبیه‌سازی معرفی می‌شود. سپس نتایج ارزیابی کارایی روش پیشنهادی ارائه می‌گردد

۴-۱- مجموعه داده‌ها، ابزار و تنظیمات شبیه‌سازی

جـــدول (۲) ویژگی‌هـــای مجموعه داده‌هـــای نامتوازن انتخاب شـــده از دو مخزن UCI [۱۴] و KEEL [۱۵] را نشان می‌دهد. روش‌های موجود داده‌افزایی عمدتا روی مســـائل طبقه‌بندی دودویی تمرکز دارند. بنابراین مجموعه داده‌های چند کلاسه بایستی به مجموعه‌هایی با کلاس‌های دودویی تبدیل شـــوند. برای انجام این تبدیل از دو روش یک به یک (OVO) [۱۶] و یک به چند (OVR) [۱۷] اســـتفاده می‌شـــود. روش OVO یک مجموعه داده چندکلاسه را به یک مجموعه داده دودویی براســـاس هر جفت کلاس تقسیم می‌کند. در مقابل، رویکرد OVR یکی از کلاس‌های چندگانه را انتخاب می‌کند و آن را در برابر همه کلاس‌های دیگر قرار می‌دهد، به‌طوری که یکی از کلاس‌ها به‌عنوان کلاس مثبت (اقلیت) و سایر کلاس‌ها به‌عنوان کلاس منفی (اکثریت) تلقی می‌شود و بدین‌ترتیب نســـبت عدم توازن کلاس‌هـــا تغییر می‌کند. نسبت عدم توازن (IR) مطابق رابطه (۵) محاسبه می‌گردد:

$$IR = \frac{N_{maj}}{N_{\min}} \qquad (۵)$$

که N_{min} تعداد نمونه‌هـــای کلاس اقلیت و N_{maj} تعداد

جدول ۲. مشخصات مجموعه داده‌های مورد استفاده

مجموعه داده	شناسه	تعداد ویژگی‌ها	تعداد کلاس‌ها	تعداد نمونه‌ها	تعداد نمونه‌های اقلیت	تعداد نمونه‌های اکثریت	IR	OVO/OVR
Vehicle	D1	۱۹	۴	۸۴۶	۱۹۹	۶۴۷	۳/۲۵	OVR
Yeast3	D2	۹	۱۰	۱۴۸۴	۱۶۳	۱۳۲۱	۸/۱	OVR
Page-blocks0	D3	۱۱	۵	۵۴۷۲	۵۵۹	۴۹۱۳	۸/۷۸	OVR
Ecoli3	D4	۸	۸	۳۳۶	۳۵	۳۰۱	۸/۶	OVR
Yeast4	D5	۹	۱۰	۱۴۸۴	۵۱	۱۴۳۳	۲۸/۰۹	OVR
Yeast6	D6	۹	۱۰	۱۴۸۴	۳۵	۱۴۴۹	۴۱/۴	OVR
Glass3	D7	۱۰	۷	۲۱۴	۱۷	۱۹۷	۱۱/۵۸	OVR
Abalone9-vs-13	D8	۹	۲	۸۹۲	۲۰۳	۶۸۹	۳/۳۹	OVO
Abalone13	D9	۹	۲۸	۴۱۷۷	۲۰۳	۳۹۷۴	۱۹/۵۷	OVR
WineQuality-White3-vs-7	D10	۱۲	۲	۹۰۰	۲۰	۸۸۰	۴۴	OVO

نمونه‌های کلاس اکثریت را نشـــان می‌دهد. هر چه نســـبت عـــدم توازن (IR) بـــه ۱ نزدیک‌تر باشـــد آن مجموعه داده متعادل‌تر است.

جهـــت ارزیابی دقـــت الگوریتم‌های طبقه‌بنـــدی نیاز به یک مجموعه آزمون اســـت که نمونه‌های آن برای آموزش طبقه‌بنـــد مورد اســـتفاده قـــرار نگرفته باشـــند، برای این منظور هر یـــک از مجموعه داده‌های جـــدول (۲) با روش اعتبارسنجی متقابل با k حلقه با نمونه‌گیری متوازن[۲] به دو بخش آموزش و آزمون تقسیم شده‌اند. در این مقاله مقدار k برابر ۵ در نظر گرفته شده است.

جهت پیاده‌سازی روش پیشنهادی از زبان برنامه‌نویسی پایتون استفاده شده است. برای ارزیابی روش پیشنهادی، خروجی این روش به‌عنوان مجموعه آموزش الگوریتم‌های طبقه‌بندی K نزدیک‌ترین همســـایه (KNN)، جنگل تصادفی (RF) و شـــبکه عصبی پرســـپترون چند لایه (MLP) به‌کار گرفته شده است. مقادیر فراپارامترهای طبقه‌بندهای مذکور با روش جستجوی شـــبکه‌ای[۳] تنظیم شده است. در جدول (۳) مقادیر بررسی شده برای فراپارامترهای الگوریتم‌های طبقه‌بنـــدی مختلف و مقدار انتخاب شـــده (مقادیر پررنگ

2-Stratified k-fold cross validation
3-Grid Search

جدول ۳. مقادیر مورد بررسی برای فراپارامترهای طبقه‌بندهای RF،KNN و MLP

طبقه‌بند	مقادیر فراپارامترها
جنگل تصادفی (RF)	ntree = {50, **100**, 200, 400, 600} $\max_{depth}$ = {**None**, 4, 6, 10, 20, 30, 50, 80, 100}
K-نزدیک‌ترین همسایه (KNN)	$n_{neighbors}$ = {3, **5**, 7, 10, 20, 30, 50} metric = {'**Minkowski**', 'Manhattan', 'Euclidean'}
پرسپترون چند لایه (MLP)	hidden_layer_sizes = (**100,50**) activation = {'identity', 'logistic', 'tanh', '**relu**'} solver = {'sgd', 'lbfgs', '**adam**'} alpha = **0/0001**

در جدول) نشـــان داده شـــده اســـت. مقادیر فراپارامترها بر اســـاس مقادیر پیش‌فرض برای هر طبقه‌بند در ســـایت scikit-learn.org و مقادیر رایج در تحقیقات مرتبط، انتخاب شده‌اند [۱۱].

مقادیـــر پارامترهـــای روش LoRAS برای هر مجموعه داده مطابق جدول (٤) و بر اساس مرجع [۱۰] انتخاب شده است. در روش LoRAS، به k در صورتی که $|C_{\min}| >= ۱۰۰$ باشـــد، به‌صورت پیش‌فرض مقدار ۳۰ داده می‌شود و در غیر این صورت مقدار ۵ می‌گیرد. اندازه N_{aff} برابر تعداد ویژگی‌ها است [۱۱].

۴-۲- ارزیابی روش پیشنهادی

بـــرای ارزیابی کارایی روش پیشـــنهادی از معیارهای رایـــج دقت، فراخوانی، F1 و AUC اســـتفاده شـــده اســـت. همچنین کارایی روش پیشـــنهادی با دو روش H-SMOTE و LoRAS مقایســـه شـــده اســـت. تمامی روش‌های مذکور بـــر روی ۱۰ مجموعه داده که مشـــخصات آنها در جدول (۲) آورده شـــده است و با به‌کارگیری ۳ طبقه‌بند یادگیری ماشین جنگل تصادفی (RF)، K نزدیک‌ترین همسایه (KNN) و شـــبکه عصبی پرسپترون چند لایه (MLP) مورد ارزیابی قرار گرفته‌اند. نتایج ارائه شده در این بخش میانگین ۱۰ بار اجرای روش پیشنهادی است.

جدول ۴. مقادیر پارامترهای روش LoRAS [۱۰]

مجموعه داده	تعداد نمونه‌های اقلیت	K	N_{aff}
Vehicle	۱۹۹	۳۰	۱۸
Yeast3	۱۶۳	۳۰	۸
Page-blocks0	۵۵۹	۳۰	۱۰
Ecoli3	۳۵	۵	۷
Yeast4	۵۱	۵	۸
Yeast6	۳۵	۵	۸
Glass3	۱۷	۵	۹
Abalone9-vs-13	۲۰۳	۳۰	۸
Abalone13	۲۰۳	۳۰	۸
WineQuality-White3-vs-7	۲۰	۵	۱۱

جدول ۵. مقایسه کارایی روش پیشنهادی بر اساس معیار AUC برای طبقه‌بند RF

مجموعه داده	بدون داده‌افزایی	LoRAS	H-SMOTE	HDAM
D1	۰/۹۳۷۴	۰/۹۰۲۶	۰/۹۵۳۴	**۰/۹۷۷۳**
D2	۰/۷۸۳۳	۰/۷۹۷۶	۰/۸۲۲۹	**۰/۸۶۰۳**
D3	۰/۹۳۰۳	۰/۸۷۳۱	۰/۹۲۳۵	**۰/۹۴۳۱**
D4	۰/۶۸۹۰	۰/۵۹۹۲	۰/۷۰۶۹	**۰/۷۸۳۳**
D5	۰/۵۲۷۰	۰/۵۰۰۰	۰/۵۵۷۳	**۰/۶۱۶۱**
D6	۰/۵۸۳۳	۰/۶۱۲۱	۰/۷۲۵۷	**۰/۷۶۹۳**
D7	۰/۵۰۰۰	۰/۴۹۷۵	۰/۵۰۰۰	**۰/۶۱۶۶**
D8	۰/۶۷۶۴	۰/۶۱۶۳	۰/۶۵۴۶	**۰/۷۴۲۵**
D9	۰/۴۹۹۵	۰/۴۹۸۴	۰/۴۹۷۸	**۰/۵۰۲۲**
D10	۰/۵۰۰۰	۰/۴۹۷۱	۰/۴۹۸۲	**۰/۵۲۴۴**
میانگین	۰/۶۶۲۶	۰/۶۳۹۳	۰/۶۸۴۰	**۰/۷۳۳۵**

جداول (۵ تا ۷) مقادیر معیار AUC را برای سه طبقه‌بند KNN، RF و MLP نشـــان می‌دهند، زمانی که این طبقه‌بندها روی مجموعـــه داده پایـــه (بدون داده‌افـــزایی) و مجموعه داده‌های تولید شـــده توسط روش‌های H-SMOTE، HDAM و LoRAS آموزش دیده‌اند. در این جداول، بهترین عملکرد به‌صورت پررنگ و رتبه دوم به‌صورت زیرخط‌دار نشان داده شده است.

جدول (۵) نشـــان می‌دهد که بـــرای طبقه‌بند RF، روش HDAM در تمامی مجموعه داده‌ها بهتر از سایر روش‌های موجود عمل کرده است. به‌طور میانگین، روش پیشنهادی HDAM با امتیاز ۰/۷۳۳۵ از AUC بالاتری برای طبقه‌بند RF برخوردار است.

مطابق جدول (٦)، برای طبقه‌بند KNN، روش پیشنهادی

جدول ۶. مقایسه کارایی روش پیشنهادی بر اساس معیار AUC برای طبقه‌بند KNN

مجموعه داده	بدون داده‌افزایی	LoRAS	H-SMOTE	HDAM
D1	۰/۹۰۸۴	۰/۸۷۴۸	**۰/۹۳۲۶**	۰/۹۳۲۲
D2	۰/۷۹۸۵	۰/۸۶۴۳	۰/۸۶۲۹	**۰/۸۷۴۱**
D3	۰/۸۲۴۷	۰/۸۶۶۳	۰/۸۷۱۱	**۰/۸۹۰۸**
D4	۰/۷۸۳۵	۰/۸۰۷۹	**۰/۸۷۷۷**	۰/۸۶۶۷
D5	۰/۵۲۹۴	۰/۷۶۰۴	**۰/۷۶۷۱**	۰/۷۶۰۸
D6	۰/۵۸۱۰	۰/۸۳۹۱	**۰/۸۵۸۳**	۰/۸۵۷۹
D7	۰/۵۰۰۰	۰/۶۲۹۳	۰/۷۶۱۵	**۰/۷۷۰۰**
D8	۰/۶۳۹۴	۰/۶۲۱۷	۰/۶۶۱۱	**۰/۷۱۳۱**
D9	۰/۴۹۹۵	**۰/۵۴۹۳**	۰/۵۰۰۷	۰/۵۵۳۹
D10	۰/۵۰۰۰	**۰.۶۱۳۵**	۰.۵۸۶۸	۰/۵۵۶۲
میانگین	۰/۶۵۶۴	۰.۷۴۷۴	۰.۷۶۷۹	**۰/۷۷۷۵**

در چهار مجموعه داده به AUC بیشتری دست یافته است. در مجموعه داده‌هـــای D5، D4، D1 و D6 روش H-SMOTE و در مجموعه داده‌های D9 و D10 روش LoRAS بهتر عمل کرده اســت. به‌طور میانگین، روش پیشـــنهادی HDAM با مقدار ۰/۷۷۷۵ برای AUC عملکرد بهتری نســبت به سایر روش‌ها دارد.

جدول (۷) نیز نشـــان می‌دهد که بـــرای طبقه‌بند MLP، روش پیشـــنهادی HDAM با میانگین امتیاز ۰/۷۹۳۶ برای AUC به نتایج بهتری در مقایســـه با ســـایر روش‌ها دست یافته است. برای این طبقه‌بند، روش پیشنهادی در ۹ مورد از ۱۰ مجموعـــه داده مـــورد بررسی، عملکـــرد بهتری در مقایسه با دیگر روش‌ها داشته است.

جدول (۸) نشان می‌دهد هر روش در چند مجموعه داده بیشترین کارایی برحسب معیارهای دقت/ فراخوانی را در مقایسه با سایر روش‌ها به‌دست آورده است. مطابق جدول (۸)، روش HDAM در ۲۵ مـــورد از ۳۰ آزمایـــش صورت گرفتـــه دارای دقت بالاتری اســـت و روش H-SMOTE با ٤ مـــورد در جایگاه دوم قـــرار دارد. همچنین، روش HDAM در ۲۲ مورد از مجموع ۳۰ آزمایش انجام شـــده فراخوانی بالاتری در مقایسه با دو روش دیگر به‌دست آورده است. بـــه دلیل این که برخی روش‌ها بر روی بعضی از مجموعه داده‌ها عملکرد یکســـانی دارند، در برخی سطرهای جدول

جدول ۷. مقایسه کارایی روش‌های پیشنهادی بر اساس معیار AUC برای طبقه‌بند MLP

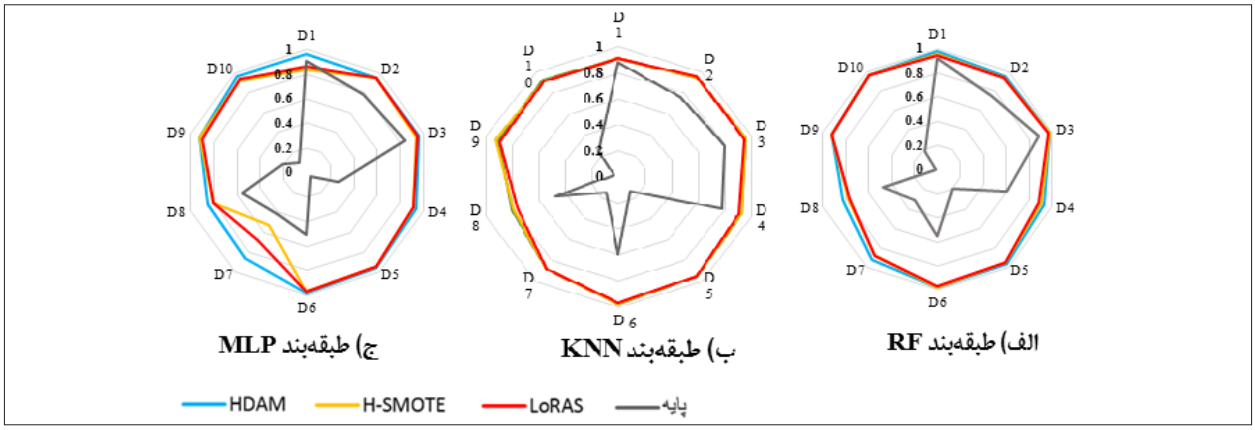
مجموعه داده	بدون داده‌افزایی	LoRAS	H-SMOTE	HDAM
D1	۰/۸۵۹۲	۰/۸۰۱۶	۰/۸۳۶۹	**۰/۹۶۲۸**
D2	۰/۸۱۲۶	۰/۸۴۶۸	۰/۸۸۵۲	**۰/۸۹۷۷**
D3	۰/۸۵۶۷	۰/۸۳۶۸	۰/۸۶۷۷	**۰/۸۸۷۸**
D4	۰/۷۸۹۰	۰/۸۳۳۲	۰/۸۶۰۱	**۰/۸۹۷۰**
D5	۰/۵۲۸۲	۰/۶۷۵۵	۰/۶۸۶۰	**۰/۷۵۸۰**
D6	۰/۵۰۰۰	۰/۷۳۵۹	۰/۷۶۲۰	**۰/۷۹۳۳**
D7	۰/۵۰۰۰	۰/۵۱۱۱	۰/۵۹۳۷	**۰/۷۳۱۶**
D8	۰/۷۳۷۰	۰/۶۹۹۳	۰/۷۱۴۳	**۰/۸۰۱۴**
D9	۰/۵۰۰۰	**۰/۵۵۷۴**	۰/۵۰۳۹	۰/۵۳۶۹
D10	۰/۵۰۰۰	۰/۵۹۴۲	۰/۵۸۵۷	**۰/۶۶۸۱**
میانگین	۰/۶۵۸۲	۰/۷۰۹۱	۰/۷۲۹۶	**۰/۷۹۳۶**

جدول ۸. کارایی روش پیشنهادی بر اساس معیارهای دقت و فراخوانی

دقت			
طبقه‌بند	HDAM	H-SMOTE	LoRAS
RF	۱۰	۰	۰
KNN	۵	۴	۳
MLP	۱۰	۰	۰
مجموع	۲۵	۴	۳
فراخوانی			
طبقه‌بند	HDAM	H-SMOTE	LoRAS
RF	۱۰	۰	۰
KNN	۳	۵	۲
MLP	۹	۱	۰
مجموع	۲۲	۶	۲

(۸) حاصل‌جمع مقادیر از ۱۰ بیشتر است. به‌عنوان مثال از مقایســـه دقت طبقه‌بند KNN این نتیجه حاصل شد که برای مجموعه داده D5، دو روش LoRAS و HDAM مقدار یکسان و برتر ۰/۹۶۰۶ را کسب کرده‌اند.

شـــکل (۲) معیـــار F1 حاصـــل از اعمـــال روش‌هـــای متوازن‌ســـازی مورد بررسی را روی ۱۰ مجموعه داده با استفاده از ســـه طبقه‌بند KNN، RF و MLP مقایسه می‌کند. همچنیـــن جدول (۹)، میانگین معیـــار F1 بر روی مجموعه داده‌ها را برای هر طبقه‌بند نشان می‌دهد. مطابق شکل (۲-الف) نتایج حاصل از طبقه‌بند RF نشـــان‌دهنده آن است که



شکل ۲. مقادیر F1 حاصل از سه طبقه‌بند (الف) جنگل تصادفی، (ب) Kنزدیک‌ترین همسایه و (ج) پرسپترون چندلایه

جدول ۹. میانگین مقادیر F1 حاصل از ۱۰ مجموعه داده برای سه طبقه‌بند KNN ،RF و MLP

طبقه‌بند	بدون داده‌افزایی	LoRAS	H-SMOTE	HDAM
RF	۰/۴۹۰۴	۰/۹۱۸۰	۰/۹۲۶۰	۰/۹۴۲۴
KNN	۰/۴۸۴۲	۰/۹۰۶۴	۰/۹۱۵۶	۰/۹۱۶۱
MLP	۰/۴۶۳۹	۰/۸۹۲۲	۰/۸۷۵۳	۰/۹۳۶۲

روش پیشنهادی HDAM در کلیه ۱۰ مجموعه داده، از F1 بالاتری نسبت به روش‌های پیشین و حالت پایه برخوردار هستند. جدول (۹) نیز نشان می‌دهد، در مقایسه با حالت پایه (بدون داده‌افزایی)، روش HDAM به‌طور میانگین به بهبود ٤٥/٢ درصد در F1 دست یافته است.

بر اساس جدول (۹) برای طبقه‌بند KNN، روش پیشنهادی HDAM بیشترین میزان بهبود (٤٣/١٩ درصد) را نسبت به حالت پایه دارد. همچنین بر طبق شکل (۲-ب)، با به‌کارگیری طبقه‌بند HDAM، KNN در ٤ مجموعه داده D4، D5، D8 و D10 مقادیر F1 بالاتری را در مقایسه با روش‌های پیشین کسب نموده است.

در نهایت مطابق شکل (۲- ج)، روش HDAM در ۹ مجموعه داده (بجز D9) بالاترین میزان F1 را در مقایسه با روش‌های پیشین کسب کرده است. جدول (۹) نیز نشان می‌دهد در صورت استفاده از طبقه‌بند MLP بهبود ٤٧/٢٣ درصدی F1 روش پیشنهادی در مقایسه با حالت پایه بدون متوازن‌سازی حاصل گردیده است.

در ارزیابی‌های انجام شده، روش پیشنهادی بر روی مجموعه‌داده‌های مختلف و با استفاده از چندین طبقه‌بند مورد آزمایش قرار گرفت. نتایج نشان داد که روش پیشنهادی در تمام موارد با طبقه‌بندهای RF و MLP موجب بهبود قابل توجه معیارهای AUC، دقت، فراخوانی و F1 نسبت به روش‌های موجود شده است. با این حال، در طبقه‌بند نزدیک‌ترین همسایه (KNN) با وجود این که به‌طور میانگین بهبود اندکی داشته است، در برخی مجموعه‌داده‌ها عملکرد اندکی ضعیف‌تر از سایر روش‌ها بود. دلیل این امر آن است که KNN به‌شدت به توزیع داده‌ها در فضای ویژگی حساس است و افزایش نمونه‌های مصنوعی در مناطق مرزی گاهی می‌تواند منجر به طبقه‌بندی اشتباه شود. بنابراین، هر چند روش پیشنهادی در اکثر طبقه‌بندها به‌ویژه مدل‌های مبتنی بر شبکه عصبی و درخت تصمیم، نتایج بهتری را ارائه داده است، اما باید محدودیت‌های آن در مواجهه با طبقه‌بندهای مبتنی بر فاصله نیز مورد توجه قرار گیرد.

۵- نتیجه‌گیری

در این مقاله یک روش ترکیبی برای داده‌افزایی معرفی شد که در هر گام آن، به‌صورت تصادفی یکی از دو الگوریتم H-SMOTE یا LoRAS انتخاب شده و یک نمونه از کلاس اقلیت با روش انتخاب شده تولید می‌شود. این فرآیند تا برآورده شدن شرط برابری تعداد نمونه‌های کلاس اقلیت و اکثریت و متوازن شدن مجموعه داده ادامه می‌یابد. نتایج ارزیابی‌های متعدد انجام شده بر روی ۱۰ مجموعه داده از

مخازن UCI و KEEL نشان می‌دهد که طبقه‌بندهای یادگیری ماشین هنگامی که بر روی مجموعه داده‌های متوازن شده با روش پیشنهادی آموزش دیده‌اند، در مقایسه با دو روش HSMOTE و LoRAS عملکرد بهتری بر اساس معیارهای رایج طبقه‌بندی یعنی دقت، فراخوانی، F1 و AUC نشان داده‌اند. این نتایج نشان می‌دهند که ترکیب دو یا چند روش داده‌افزایی با خصوصیات متفاوت و مکمل می‌تواند از مزایای همه روش‌ها بهره‌مند شود و مشکلات استفاده تکی از هر روش را کاهش دهد. به‌عنوان کار آتی، برآنیم تا با ترکیب روش‌های داده‌افزایی مناسب برای طبقه‌بندی چندکلاسه، کارایی روش پیشنهادی را برای طبقه‌بندی چندکلاسه مورد بررسی قرار دهیم.

۶. مراجع

[1] Khan, M. H., Iqbal, M. S., Muhammad Shah, F., Al Mahmud, J., Popel, M. H., Showrov, M. I. H., Ahmed, S. & Rahman, O. 2020. “A Survey of Methods for Managing the Classification and Solution of Data Imbalance Problem،” Journal of Computer Science, Vol. 16, No. 11, pp. 1546-1557.

[2] Tsai, C. F., Lin, W. C., Hu, Y. H. & Yao, G. T. 2019. “Under-Sampling Class Imbalanced Datasets by Combining Clustering Analysis and Instance Selection،” Information Sciences، vol. 477, pp. 47-54, 2019.

[3] Huang, M. W., Tsai, Ch. F. & Lin, W. Ch. 2021. “Instance selection in medical datasets: Adivide-and-conquer framework,” Vol. 90, pp. 106957.

[4] Moran, M., Cohen, T., Ben-Zion, Y. & Gordon, G. 2022. “Curious instance selection,” Information Sciences, Vol. 608, pp. 794-808.

[5] Chawla, N. V., Bowyer, W., Hall, L. O. & Kegelmeyer, W. P. 2002. “SMOTE: Synthetic Minority Over-sampling Technique,” Journal of Artificial Intelligence Research, Vol. 16, pp. 321-357.

[6] Douzas, G. & Bacao, F. 2019. “Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE,” Information Sciences, Vol. 501, pp. 118-135.

[7] Sáez, J. A., Luengo, J., Stefanowski, J. & Herrera, F. 2015. “SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering,” Information Sciences, Vol. 291, pp. 184-203, 2015.

[8] Bunkhumpornpat, C., Sinapiromsaran, K. & Lursinsap, C. 2009. “Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem,” in Advances in Knowledge Discovery and Data Mining, New York.

[9] Chao, X. & Zhang, L. 2023. Few-shot imbalanced classification based on data augmentation. Multimedia Systems, Vo. 29, No. 5, pp.2843-2851.

[10] Bej, S., Davtyan, N., Wolfien, M., Nassar, M. & Wolkenhauer, O. 2021. “LoRas: an oversampling approach for imbalanced datasets,” Machine Learning, Vol. 110, No. 2, pp. 279-301.

[11] Temraz, M. & Keane, M. T. 2022. “Solving the class imbalance problem using a counterfactual method for data augmentation,” Machine Learning with Applications, vol. 9, pp.100375.

[12] Melícias, F. S., Ribeiro, T. F., Rabadão, C., Santos, L., & Costa, R. L. D. C. 2024. “GPT and interpolation-based data augmentation for multiclass intrusion detection in IIoT,” IEEe Access, Vol. 12, pp. 17945-17965.

[13] Dai, H., Liu, Z., Liao, W., Huang, X., Cao, Y. Wu, Z. & Li, X. 2025. “Auggpt: Leveraging chatgpt for text data augmentation,” IEEE Transactions on Big Data.

[14] Blake, C. & Merz, C. 1978. “UCI Machine Learning Repository,” Department of Information and Computer Science, [Online] Available: https://archive.ics.uci.edu/datasets. [Accessed 2025].

[15] Derrac, J., Garcia, S., Sanchez, L. & Herrera, F. 2015. “Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework,” J. Mult. Valued Logic Soft Comput, 17, pp. 255-287.

[16] Chih-Wei, H. & Chih-Jen, L. 2002. “A comparison of methods for multiclass support vector machines,” IEEE Transactions on Neural Networks, Vol. 13, No. 2, pp. 415-425.

[17] Bishop, C. M. & Nasrabadi, N. M. 2006. “Pattern recognition and machine learning,” New York: springer, Vol. 4, No. 4, pp. 738.